



Universität Ulm | 89069 Ulm | Germany

**Fakultät für Ingenieurwissenschaften,
Informatik und Psychologie**

Institut für Neuroinformatik

Direktor: Prof. Dr. Dr. Daniel Alexander Braun

Deep Learning for Robust and Explainable Models in Computer Vision

Dissertation zur Erlangung des Doktorgrades
Doktor der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Ingenieurwissenschaft, Informatik und Psychologie
der Universität Ulm

vorgelegt von
Mohammadreza Amirian
aus Shiraz

Ulm 2023

Amtierende Dekanin: Prof. Dr. Anke Huckauf

Gutachter: Prof. Dr. Friedhelm Schwenker

Gutachter: Prof. Dr. Thilo Stadelmann

Gutachter: Prof. Dr. Martin Jaggi

Tag der Promotion: 20th October 2023

Acknowledgements

I want to extend my heartfelt gratitude to the individuals who have played essential roles in my academic journey while completing my Ph.D. dissertation. Their support, mentorship, and presence have been invaluable, and I am deeply appreciative. I must begin by acknowledging the support of my family: Edalat, Parivash, and Milad. Their belief in my aspirations and constant encouragement has strengthened me throughout this demanding journey.

I am profoundly grateful to my distinguished Ph.D. committee, particularly Professor Friedhelm Schwenker, to whom I owe special thanks for his mentorship. The wisdom and expertise of Professors Thilo Stadelmann, Martin Jaggi, Hans Kestler, Daniel Alexander Braun, Günther Palm, and Matthias Tichy have been instrumental in shaping my research and academic growth. My gratitude extends to project managers Stefan Scheib and Frank-Peter Schilling, whose guidance and collaboration significantly enriched my research endeavors.

The diverse and talented group of coauthors, who have made invaluable contributions to my track record, have enriched the academic discourse and expanded my horizons within the realm of research. Their collaborative spirit, combined with our shared dedication to the subject matter, has played a pivotal role in advancing my research. I want to extend my gratitude to each coauthor for their instrumental contributions to my academic journey, including Lukas Tuggener, Patrick Thiam, Viktor Kessler, Markus Kächele, Stefan Scheib, Javier Montoya, Alexander Züst, Ivo Herzig, Peter Eggenberger Hotz, Rudolf Marcel Füchslin, Pascal Paysan, Igor Peterlik, Samuel Wehrli, Corinna Hertweck, Stefan Glüge, Lukas Lichtensteiger, Marco Morf, Fernando Benites, Pius von Däniken, Peter Bellmann, Georg Layher, Yan Zhang, Maria Velana, Sascha Gruss, Steffen Walter, Harald Christhelm Traue, Daniel Schork, Jonghwa Kim, Elisabeth André, Heiko Neumann, Taye Girma Debelee, Abraham Gebreselasie, Dereje Yohannes, Kamran Kazemi, Mohammad Javad Dehghani, Jonathan Gruss, Yves D. Stebler, Ahmet Selman Bozkir, Marco Calandri, Ricardo Chavarriaga, Yvan Putra Satyawana, and Dandolo Flumini. Each has left an indelible mark on my academic journey, and I am grateful for their collaborative efforts.

I also want to acknowledge the colleagues I've worked with at Ulm University, Zurich University of Applied Sciences (ZHAW), and the Swiss Federal Institute of Technology in Lausanne (EPFL). Special thanks to Patrick, Viktor, and Heinke from Ulm. Your inspiration and shared commitment to academic excellence have impacted my Ph.D. journey. Special thanks to my colleagues and collaborators at ZHAW: Jonas, Katrin, Yasmin, Sean, Adhiraj, Yvan, Sebastiano, Susanne, Norman, Raphael, Pascal, Peng, Katsiaryna, Daniel, Philipp, Jonathan, Catherine (for her invaluable proofreading), Claude, Gabriel, and Ahmad. And to my colleagues at EPFL: Mary-Anne, Anastasia, Lie, Matteo, Amirkeivan, El Mahdi,

Atli, Thijs, Jean-Baptiste, Tao, Prakhar, Praneeth, and Seyed-Mohsen. Your shared experiences, diverse backgrounds, and academic commitment have enriched my academic journey.

I expand my appreciation to the administrative staff at Ulm University, ZHAW, and EPFL, including Traude, Birgit, Annette, Cornelia, Regula, Pamela, and Jennifer, whose efficient and supportive work has facilitated my academic journey.

Special mention to my flatmates, Patrick, Draženka, Patrick, Valentina, and Sandro, who have contributed to my social life and created a supportive and nurturing environment. Close friends and community in Ulm, Winterthur, and Zurich area, including Abbas, Ramanjeet, Alper, Meissam, Gholamhossein, Mahsa, Maryam, Fatemeh, Sajad, Mahdi, Elham, Timo, Ronak, Reza, Romina, Farhad, Najmeh, Pooya, Helmut, Javier, Giani, and Roshan, have been a source of joy and balance, significantly enhancing my social life. I would also like to express my gratitude to those who have enriched my life in Switzerland beyond academia, including the FC Kreuzlingen football club, the Keller family, and the Bailamos Salsa Club.

The support, mentorship, and friendship of countless individuals have profoundly shaped my Ph.D. dissertation journey. I am deeply thankful for the lessons, experiences, and relationships that contributed to my academic and personal growth. Your contributions have been invaluable in making this academic journey possible, and I am sincerely grateful to each of you.

Abstract

Recent breakthroughs in machine and deep learning (ML and DL) research have provided excellent tools for leveraging enormous amounts of data and optimizing huge models with millions of parameters to obtain accurate networks for image processing. These developments open up tremendous opportunities for using artificial intelligence (AI) in the automation and human assisted AI industry. However, as more and more models are deployed and used in practice, many challenges have emerged. This thesis presents various approaches that address robustness and explainability challenges for using ML and DL in practice.

Robustness and reliability are the critical components of any model before certification and deployment in practice. Deep convolutional neural networks (CNNs) exhibit vulnerability to transformations of their inputs, such as rotation and scaling, or intentional manipulations as described in the adversarial attack literature. In addition, building trust in AI-based models requires a better understanding of current models and developing methods that are more explainable and interpretable a priori.

This thesis presents developments in computer vision models' robustness and explainability. Furthermore, this thesis offers an example of using vision models' feature response visualization (models' interpretations) to improve robustness despite interpretability and robustness being seemingly unrelated in the related research. Besides methodological developments for robust and explainable vision models, a key message of this thesis is introducing model interpretation techniques as a tool for understanding vision models and improving their design and robustness. In addition to the theoretical developments, this thesis demonstrates several applications of ML and DL in different contexts, such as medical imaging and affective computing.

Contents

Abstract	iii
1 Introduction	1
1.1 Motivation	2
1.2 Problem Statement	4
1.3 Contributions	5
1.4 Publications	8
1.4.1 Journal Papers	8
1.4.2 Conference Papers	9
1.4.3 Book Chapter	11
1.5 Organization of Thesis	12
2 Theoretical Foundations	13
2.1 Convolutional Neural Networks	14
2.1.1 Convolution Operator	15
2.1.2 Feature Maps	17
2.1.3 Pooling Layers	17
2.1.4 FeedForward Neural Networks	17
2.1.5 Convolutional Neural Networks	18
2.1.6 Advanced Blocks	19
2.1.7 Architecture Search	22
2.1.8 3D Convolutional Neural Networks	23
2.2 Vision Transformers	24
2.2.1 Preliminaries	24
2.2.2 Attention	25
2.2.3 Self-Attention	25
2.2.4 Positional Encoding	25
2.2.5 Relative Positional Encoding	26
2.2.6 Vision Transformers for Classification	26
2.2.7 Vision Transformers for Segmentation	27
2.3 Optimizing Neural Networks	28
2.3.1 Optimizing Trainable Parameters	28
2.3.2 Optimization to Generalization	30
2.4 Related Work	32

3	RBF Classifiers for Explainable Computer Vision Using CNNs	35
3.1	Introduction	36
3.2	Related Work	38
3.3	Radial Basis Function Networks	40
3.4	Adapting RBFs for CNNs	43
3.4.1	Introducing Unsupervised Learning Loss	43
3.4.2	Quadratic Kernel	44
3.5	Experimental Results	44
3.6	Visualization of the RBF Classifiers	47
3.6.1	Visualization of the Training Process	47
3.6.2	Similarity Metric Learning and Interpretability	49
3.7	Discussions and Conclusions	52
4	Using Interpretability to Detect Adversarial Attacks for Robust CNNs	53
4.1	Introduction	54
4.2	Related Work	55
4.3	Background	56
4.3.1	Adversarial Attacks	56
4.3.2	Feature Response Estimation	59
4.4	Explainable Adversarial Attacks Detection	60
4.4.1	Tracing Adversarial Attacks	60
4.4.2	Detecting Adversarial Attacks	60
4.5	Experimental Results	62
4.6	Discussion and Conclusion	64
5	Motion Compensation in Computed Tomography Using CNNs	65
5.1	Introduction	66
5.2	Related Work	68
5.3	Materials and Methods	69
5.3.1	CBCT Reconstruction	69
5.3.2	Motion Simulation	70
5.3.3	Datasets	71
5.4	Supervised Learning for Motion Artifact Reduction	72
5.4.1	DL-Enabled CBCT Reconstruction	72
5.4.2	Evaluation Metrics	75
5.4.3	Experimental Details	75
5.5	Experimental Results	76
5.5.1	Quantitative Results	76
5.5.2	Clinical Evaluation	79
5.6	Discussions and Conclusions	82

6	Applications in Affective Computing, Medical Imaging and Beyond	85
6.1	Affective Computing	86
6.1.1	Facial Expression Estimation	86
6.1.2	Emotion Recognition	87
6.2	Automated Data Analysis	92
6.2.1	Automated Machine Learning	92
6.2.2	Automated Deep Learning	93
6.3	Medical Diagnosis and Imaging	96
6.3.1	Pain Estimation	96
6.3.2	Data Homogenization	97
6.4	Face Recognition	100
6.4.1	Algorithmic Bias in FR Systems	100
6.4.2	Measuring Bias and Awareness	100
6.5	Rotation-Invariant Vision Transformers	102
6.5.1	Introduction and Problem Statement	102
6.5.2	Method and Experimental Results	105
7	Conclusions	109
7.1	Summary of Thesis	109
7.2	Future Research Work	110
7.3	Practical Discussions	112
	Bibliography	115

List of Figures

2.1	The cross-correlation function is often implemented in deep learning libraries for convolutional neural networks. For an input image, the output (kernel response) is the dot product of the vectorized kernel with a field which is the same size as the input image. The kernel slides over the entire image area with a given step size (figure adopted from [146]).	16
2.2	Max-pooling (MP) and average-pooling (AP) layers with kernel size and stride of 2 for CNNs (figure adopted from [291]).	17
2.3	A convolutional neural network for representation learning from an input image, followed by a feedforward network for object classification.	18
2.4	The residual connection between a layer’s input and output improves the gradient flow (figure is adopted from [96]).	19
2.5	This figure, adopted from [260], depicts the idea of the inception blocks and their practical implementation.	20
2.6	Convolutional block attention module (CBAM) with its two main components for refining channel and spatial features (figures are adopted from [260]).	21
2.7	Two- and three-dimensional convolutions. 2D convolutions target images, while 3D convolutions are suitable for volumetric data (figures are adopted from [108]).	23
2.8	Vision transformers (ViTs) for image classification (figure is adopted from [65]).	27
2.9	Vision transformers (ViTs) for image semantic segmentation (figure is adopted from [253]).	28

3.1	<p>Figures on the top and bottom rows visualize the position of a test image in the clusters optimized using the unsupervised loss function. The output of CNN backbones is connected to RBFs' input through a fully connected layer, and the input features of the RBFs are referred to as embeddings in this chapter. The model compares the embeddings of each image with cluster centers using a trainable similarity distance metric. The same distance metric can be used to find similar and dissimilar images to a test sample amongst training images (visualized in the table in the middle row). The RBFs apply an activation function to the distance of the training images from the cluster centers to compute activation values. The output layer of the RBF is optimized for classification based on these activation values. The entire CNN-RBF architecture is optimized end-to-end with a specific initialization (figure adopted from [14]).</p>	36
3.2	<p>Activation functions for RBF networks. Here is the list of the parameters for depicting the kernels: $\sigma = 1$, $\alpha = 1/2$, and $\beta = 1/2$. The proposed quadratic activation kernel is linear based on the r^2. Consequently, the CNN goes through a completely linear forward path, and thus, gradients are computed and backpropagated efficiently (figure adopted from [14]).</p>	41
3.3	<p>Hyperparameter search results from CIFAR-10 (top) and CIFAR-100 (bottom). The top five performing sets of hyperparameters for each dataset are highlighted in yellow (figures adapted from [14]).</p>	46
3.4	<p>This figure presents the location of data samples compared to the cluster centers during the training process. The centers of the clusters are in the middle of the figures. The training samples are located at a random angle based on their distance from the center of the clusters. The vertical and horizontal axes show the normalized distances (figure adopted from [14]).</p>	48
3.5	<p>Two-dimensional representation of the training process: the figure presents the embeddings of the convolutional backbone (top row), and the activations of the RBFs (bottom row) mapped to a two-dimensional space using t-SNE [277]. The vertical and horizontal axes depict the normalized values; however, all sub-figures use the same normalization factors (figure adopted from [14]).</p>	48
3.6	<p>This figure depicts similar and dissimilar training images for given test images based on the similarity metric computed in Equation 3.1. The figure depicts the top 7 most similar and dissimilar training images for a given test image in every two rows. The images shown in every two consecutive rows belong to one of the datasets in Table 3.1 in the same order (figure adopted from [14]).</p>	50

3.7	The presented figure visualizes the top 14 images selected using different distance metrics in the embedding space for a given test image (figure adopted from [14]).	51
3.8	The Figure illustrates the clusters contributing to a CNN-RBF network’s correct class (top row) and the wrong class (bottom row). The larger image with red borders in each cluster representation is the test sample. Red circles show the distance of the samples to the cluster center, and the background is proportional to the activation values of the cluster. The brighter the activation value, the larger it is, and the maximum activation at the cluster center is equal to one (figure adopted from [14]).	52
4.1	Examples of different state-of-the-art adversarial attacks on a VGG19 model: original images and labels (left), perturbations (middle), and mislabeled adversarial examples (right). In the middle column, zero difference is encoded white, and the maximum difference is black because of visual enhancement (figure adopted from [15]).	57
4.2	a) Distribution of average local spatial entropy in clean images (green) versus adversarial examples (red) as computed on the ImageNet validation set [223]. b) Receiver operating characteristic (ROC) curve of the performance of the detection algorithm on different attacks (figure adopted from [15]).	62
4.3	Successful adversarial examples created by DeepFool [188] for binary and ternary classification tasks are only possible with noticeably visible perturbations (figure adopted from [15]).	64
5.1	Motion Artifacts. Left: CBCT scans with motion artifacts from the test dataset. Right: Scan with artificially produced motion artifacts from the motion simulation. The scans are presented in HU with W/L=1000/0 (figure adopted from [12]).	71
5.2	The architecture of the proposed dual-domain model for end-to-end optimization consists of the following components: (i) a projection enhancement network (PE-Net), (ii) a projection-to-volume reconstruction layer, and (iii) a volume enhancement network (VE-Net) (figure adopted from [12]).	73
5.3	Example results for FDK reconstruction (volume domain optimization). Presented is the uncorrected volume using default reconstruction (left), the ground truth volume, both as difference and absolute image (“average volume”, top right), as well as the corrected volume (bottom right). Images are presented in HU with W/L=1000/0 (figure adopted from [12]).	78

5.4	Example results for FDK reconstruction (volume domain optimization). The uncorrected volume using default reconstruction (left), the ground truth volume, both as difference and absolute image (“average volume”, top right), as well as the corrected volume (bottom right) are depicted in the table. Images are presented in HU with W/L=1000/0 (figure adopted from [12]). . . .	80
5.5	The table shows example results for iCBCT reconstruction for real-world test dataset, using the two options for the choice of ground truth. The uncorrected volumes using default reconstruction (left), the residual corrections (middle), as well as the corrected volumes (right) are presented (figure adopted from [12]). .	81
6.1	Several action units (AUs) used for facial expression estimation (figure is adopted from [8]).	86
6.2	Several dictionary-learned facial templates used for facial feature extraction (figure is adopted from [8]).	87
6.3	Presented is the proposed sequence of blocks for the automatic fusion of audiovisual and biophysiological information to predict arousal levels (figure is adopted from [10]).	89
6.4	Presented are the echo state networks (ESNs) based architectures for modeling temporal information dependencies and fusing multimodal information. One ESN is trained for each modality, and the predictions of all modalities are combined using precomputed weights according to the importance of modalities per task. . . .	90
6.5	The figure depicts the temporal mismatch between audio features and gold standard labels. The average performance of predicting the arousal level of participants from audio increases considerably when features are aligned with gold-standard labels (figure is adopted from [10]).	91
6.6	Performance of four different vision datasets in terms of ALC of MobileNetV2 as a function of weight decay and learning rate (top two rows) and averaged performance over all datasets (bottom row). The green dot indicates the best performance (figures are adopted from [271]).	95
6.7	The proposed architecture for <i>PrepNet</i> model with three modules: (i) an auto-encoder aims at CT dataset homogenizer; (ii) a multiclass classifier to recognized CT-datasets; and (iii) a binary classifier for diagnosis (COVID-19). The loss functions of the dataset classifier and auto-encoder were trained adversarially against each other. The binary classifier for diagnosis (COVID-19) was trained independently using the preprocessed scans by auto-encoder (figure adopted from [11]).	98

6.8	Original images from the datasets with different preprocessing methods applied (figure adopted from [11]).	99
6.9	Probability density distribution of pairwise (Euclidean and Cosine) distances between test images' embeddings of different races. The embeddings are computed using the VGG model fine-tuned for face recognition (VGGFace2 [35]) with different embedding dimensionalities (128, 256 and 2048). The figure shows that the faces from the Caucasian race, which have the largest share of data samples, have a larger average distance than those of Africans, Asians, and Indians (figure adopted from [87]).	101
6.10	Classification and segmentation performance of vision transformers under different degrees of rotation. Augmentation improves the robustness of vision transformers against rotation; however, rotation invariance encoded in the method as inductive bias can improve the sample efficiency of the models.	103
6.11	The proposed patch embedding methods for vision transformers: a) Isotropic patch embedding for the entire image. Every patch is sampled based on the circles around the center of the patch. b) Radial patch embedding in the patch level technique samples every patch based on the pixels on a circle radius positioned at the patch's center. c) Radial patch embedding for the entire image.	106
6.12	Robust training against rotation using rotation invariant patch embedding techniques.	107

List of Tables

3.1	An overview of computer vision benchmark datasets used to evaluate the performance of CNN-RBFs (table adopted from [14]). . .	45
3.2	List of the final hyperparameters used for each computer vision benchmark dataset to achieve the performance of CNN-RBF architectures (table adapted from [14]).	46
3.3	Comparing the performance of various CNN-RBF architectures with pretraining and augmentation on benchmark computer vision datasets. The best results column is the top performance of the current state-of-the-art architecture on the benchmark dataset (table adapted from [14]).	47
4.1	Effect of adversarial attacks on feature responses: (left) original images, and their feature responses, (right) perturbed versions, and their feature responses (figure adopted from [15]).	59
4.2	Input, feature response maps, and local spatial entropy for clean and perturbed images, respectively (table adopted from [15]). . .	61
4.3	The table describes the numerical evaluation of detection performance on different adversarial attacks. Column two gives the number of tested images and approximate elapsed run time. The success of an adversarial attack is defined if a perturbation changes the prediction. Columns four and five show average confidence values of the true (ground truth) and wrong (target) classes after a successful attack. Finally, the last columns show detection rates for different false positive rates (table adopted from [15]).	63
4.4	This table describes the performance of similar state-of-the-art adversarial attack detection methods. The Area Under Curve (AUC) is the average value of all attacks in the third and last row (table adopted from [15]).	63

5.1	Presented are the quantitative results of DL-based motion correction for CBCT data with simulated motion. The table presents the performance of the proposed motion reduction framework based on the RMSE, PSNR, and SSIM metrics and reports the mean and standard deviation of the body-masked difference (correction) volumes. The metrics are calculated between the reconstructed and ground truth volumes, converted to HU with slope and intercept of 48200 and -1106 , respectively. All numerical values are averaged over the test set. The table shows the average metric together with the average gain (or loss) and the latter's standard deviation to clarify the contribution of the motion correction. For example, in the last row, the average PSNR is reported as 33.00 dB, corresponding to an average improvement of 4.62 dB, with a standard deviation of 0.82 dB. The models noted by † are used for clinical evaluation (Section 5.5.2) (figure adopted from [12]).	77
5.2	Results of the clinical evaluation. This table shows the preferences for CNN-based or default iCBCT reconstruction when using CNN models trained using either average volume or average amplitude ground truth concerning motion artifact reduction and potential applications such as plan adaptation and dose calculation, patient positioning and segmentation (table adopted from [12]).	82
6.1	Performance of three automated machine learning algorithms with different paradigms on AutoML challenge datasets and their convergence time [94] (table adopted from [272]).	93
6.2	Test and cross-dataset performance of different methods. Using an adversarial loss to train a PrepNet improves the cross-dataset average performance (table adopted from [11]).	98
6.3	The performance of rotation invariant vision transformers on several vision benchmark vision datasets. Rotation invariant patch embedding increases the robustness of ViTs at the expense of a decrease in performance.	107

1 Introduction

As a result of the widespread interest in applying artificial intelligence (AI) in practice, several intriguing challenges and research topics have recently emerged due to the trustworthiness of models in various circumstances being brought into question [109]. Researchers have cited explainability, robustness, and fairness among other hindrances in developing trustworthy AI [123]. Therefore, understanding the reasons for failure and creating ability to explain the inner workings of neural networks has attracted researchers' attention [64, 268]. Furthermore, developing interpretable and explainable models has become a research focus in its own right [85].

The three terms *robustness*, *explainability*, and *interpretability* are the fundamental concepts behind this thesis. The term *robustness* refers to a clearer concept compared with *explainability* and *interpretability*. Model *robustness* is proportional to the consistency of the model's performance against naturally-induced or manually-computed corruption and alterations affecting the data to deviate from the training distribution [66, 170]. However, the other two terms, *explainability* and *interpretability*, and their boundaries and overlaps are still subjects of research at the taxonomy level [92]. In this thesis, the terms *explainability* and *interpretability* are used analogously to their usage in [222]. Accordingly, *explainability* refers to the explanation of models' decisions (even though these models can be intrinsically black boxes), and *interpretability* refers to the design patterns that are inherently interpretable and understandable by humans.

The scope of this work is narrowed down from AI in general to focus on computer vision models, including convolutional neural networks (CNNs) and vision transformers (ViTs). This thesis is motivated by practical applications and presents relevant research concerning neural networks' robustness, fairness, interpretability, and explainability. Moreover, the thesis provides not only theoretical and fundamental advances but also offers several applications in which computer vision models have been successfully used. The remainder of this chapter explains the motivations for this research and describes the scientific problem we address. Finally, this chapter provides a list of papers and publications related to this research, followed by the thesis organization.

1.1 Motivation

This thesis aims at using ML and DL in practical applications, and presents several success stories in Chapter 5 and Chapter 6. Despite the numerous successful applications of DL, there are deterrents for putting it into practice in sensitive applications where humans are involved. This thesis presents relevant research tackling such challenges as follows: 1) adversarial robustness in Chapter 4, 2) explainability via interpretable classifiers in Chapter 3. The remainder of this section describes the motivation of the researchers, elaborates on related efforts, and describes the niches to which this thesis contributes.

Researchers attempted to expose the complications of using DL in practice by studying robustness [22], fairness [179], explainability [64], interpretability [36], accountability [130], reliability [226], safety [2], and privacy [107] etc. Although these themes were independently the subject of research and scientific concern, they have only recently been grouped under the overarching topic referred to as *trustworthy AI* in the literature [123]. Trustworthy AI literature summarizes the research effort as developing models which are effective in practice and aligned with positive societal effects. The following paragraphs explain the key components of trustworthy AI which are considered in this thesis, *explainability*, *interpretability*, *robustness* and *fairness*, and describe the goals of the related research.

After researchers found flaws in the preciseness and biasedness in vision models [262, 48] as well as natural language processing methods [28, 301], the European Union introduced the “right to explain” in the general data protection regulation (GDPR) as an attempt to protect human rights when decisions are automated [64]. This human right relates to the human’s ability to understand the AI-based agent logic in human-machine interaction [220]. Some of the research work related to *explainability* targets explaining the models’ decisions even if researchers treat the models as black boxes [47]. The concept of *interpretability* includes research attempting to open the black box of neural networks with revealing patterns about the inner mechanism of models [315]. Saliency map visualizations [26] and feature response visualization methods [247] are examples of researchers’ endeavors in targeting interpretability.

The topic of *robustness* is directly related to the performance of vision models. Vision models have shown a drop in performance as the consequence of changes in data distribution based on naturally-induced variations [66] or manually-computed perturbations [170]. Robustness is a significant hurdle to overcome in life-long deployments of vision models, which has inspired many recent research works. An example of the output of such research is equivariant CNNs, which aim to improve robustness against rotation and translation of input images [45, 218, 111]. Moreover, adversarial training research attempts to neutralize the targeted attack’s

effect in fooling vision models [270]. Still, there are several gaps in dataset collection and robust model development for naturally-induced variations and different illumination conditions [149].

Social activists and computer vision researchers recently raised concerns regarding *fairness* in automated decisions [99, 23]. Automated decisions are considered unfair if they rely on sensitive variables such as gender, ethnicity, sexual orientation, or disability [281]. The researchers identified sources of bias leading to unfair decisions, which can be divided into two general categories of algorithmic biases, and biases in the training datasets [179]. This topic gained much attention and media coverage after the deployment of face recognition (FR) systems in public surveillance [99, 169, 161, 48]. Since then, many researchers have attempted to tackle the problem of biasedness at the algorithmic level [217, 238] and collect diverse datasets to achieve fair modeling for all genders and races [289, 244].

This thesis explores the obstacles to using AI, particularly machine learning (ML) and deep learning (DL), in practical applications of computer vision in which robustness and explainability are of high importance. The ultimate goal of this work is the successful application of ML and DL algorithms, although this is not a trivial task and raises many additional questions that require further research. A short phrase that summarizes the long-term goal alongside the focus of this work is the development of *trustworthy AI*. Although the term *trustworthy AI* has only recently come into vogue, this thesis addresses its various components, including fairness, robustness, interpretability, and explainability. Trustworthiness includes other elements outside the scope of this research, such as security, privacy, and accountability. This thesis presents several applications besides fundamental research that support robust and explainable AI (XAI).

Understanding the behavior of computer vision models (explainability) has always been a subject of curiosity for scientific endeavors. The first group of researchers who analyzed current models considered them to be black boxes and predicted their performance by changing the input and observing the behavior of the models' output. The second group of researchers proposed intrinsically more interpretable and explainable models. This thesis presents a chapter on using radial basis function networks (RBFs) as classifiers on top of CNNs to improve the interpretability of decision-making in computer vision models which contributes to XAI research.

Early in the development of CNNs, researchers found that computer vision models were only robust in a limited range of rotation and scaling in the input images¹. In addition, lighting conditions and other environmental disturbances caused errors in recognizing image patterns. Last but not least, the researchers found that optimizing images made it possible to fool the computer vision models into interpreting two images, which humans perceive to be identical, differently, leading

¹<http://yann.lecun.com/exdb/lenet>

to the computation of so-called adversarial perturbations. Since then, improving the robustness of computer vision models has become a popular research topic. Researchers have made enormous efforts to understand the models, identify the reasons behind failures, and improve the computer vision models. Interpreting the behavior of computer vision models can serve as a tool to monitor the reasons for failures. Therefore, interpretability and robustness are closely related in the literature. For example, researchers have found that computer vision models can focus on the wrong features or background information when classifying an object. This thesis includes a chapter on using feature response maps—where a computer vision focuses its attention in response to the visual input—for identifying adversarial examples.

In addition to the above theoretical developments, this thesis presents many applications inspired by its original goal. It targets the vulnerabilities (motion artifacts) found in classical computed tomography (CT) reconstruction methods as the main practical contribution. Moreover, it presents many other side contributions to affective computing, pain estimation, AutoML, AutoDL, medical data homogenization, and fairness in face recognition systems.

1.2 Problem Statement

This thesis is motivated by solving real-world problems using computer vision methodology. The applications presented in Chapter 5 and Chapter 6 are derived from several real-world problems where ML and DL are useful. However, there are still gaps in the current methodologies that must be addressed in order to achieve trustworthy models for general applications. Chapter 3 and Chapter 4 present the fundamental research targeting the explainability and robustness gaps required to apply computer vision models in practice. The remainder of this section details the scientific problems addressed in each chapter individually.

Current architectures for computer vision models based on CNNs and ViTs use a stack of convolutional or self-attention layers to develop a representation of the inputs. Despite the different architectures in the image encoder of most computer vision models, all of these models use a stack of multiple fully connected (FC) layers or multilayer perceptrons (MLP), on top of learned latent representations [24]. Researchers commonly used FC layers as the optimal classifiers for deep models because of their efficiency in gradient backpropagation [148]. MLPs divide the embedding space in their last layer into multiple classes using hyperplanes. The distance of the input image representations from the decision boundary drawn by the hyperplanes in the last layer of such classifiers determines the decision confidence of the models. Researchers found that such classification is not optimal for outliers, since models developed using MLP classifiers demonstrate low reliability for random (garbage) classes. This is due to the outliers being far from the

classifier’s decision boundary, which contributes to the models’ flawed high confidence in these samples. In addition, computer vision models trained using linear classifiers are vulnerable to optimized perturbations (adversarial attacks). Ian Goodfellow has attempted to thoroughly analyze various classifiers to evaluate their robustness to adversarial attackers and garbage classes [89]. His preliminary speculations hint that RBFs might be more robust than MLPs. However, the study was inconclusive due to the difficulties in optimizing RBF networks, even for simple tasks such as classifying handwritten digits. This thesis proposes modifications to RBF networks that improve the optimization of RBFs and shows how RBF classifiers are beneficial for interpreting the decision-making of computer vision models.

After the advent of CNNs, researchers were very skeptical and curious about their functionality. They work as highly accurate models but seem to appear as black boxes. The result of this curiosity and scientific venture is a vast amount of literature analyzing the behavior of CNNs by computing the models’ feature response maps through the inversion of the forward path. As extensive as the techniques for visualizing models are, their applications are rare. This thesis presents an example of detecting adversarial attacks using feature response maps. The intention is that this idea inspires researchers to use their knowledge of interpreting computer vision models for architecture development and debugging.

Neural networks have outperformed their competitors in approximating arbitrary functions and learning patterns from enormous amounts of data. However, AI projects still face high risks due to not achieving the intended goal, unforeseen delays, extensions, and application failures. This thesis presents several successful applications that allow the reader to understand where using ML and DL-based methods are beneficial. For example, we address motion artifacts in cone beam computed tomography (CBCT) scans. Volumetric (3D) data from CBCT scans are reconstructed from hundreds of 2D X-ray images from different angles. The analytical reconstruction algorithms are robust when the target volumes are constant and free of motion. However, this assumption does not hold due to respiratory or cardiac motions present in the human body. In this work, we demonstrate how CNNs can be used to compensate for motion artifacts in CBCT scans. Along with this application, this thesis offers several other applications to show some possible and successful venues in which ML- and DL-based models are superior to classical computer vision methods in practice.

1.3 Contributions

The main contributions of this thesis to ML and DL research are as follows:

- Chapter 3: Modern vision architectures use multilayer perceptrons (MLPs)

in the form of fully connected layers as classifiers, as researchers have largely abandoned radial basis function networks (RBFs) due to optimization problems. This thesis provides the following developments in training RBFs as classifiers for convolutional neural networks (CNN) backbones: 1) Presentation of the first successful attempt to use RBFs as the classifier of modern computer vision models for object classification. 2) Introduction of a novel quadratic activation function to build a linear computational graph with RBFs. 3) Simultaneous optimization of supervised loss for classification and unsupervised loss for clustering [14].

- Chapter 3: Solving the technical problems of optimizing RBFs as classifiers for computer vision models opens several possibilities for training computer vision models: 1) Combining supervised and unsupervised learning by simultaneously optimizing two target losses. 2) Learning a similarity distance metric to find similar images by optimizing the covariance matrix in the embedding space. 3) Improving the interpretability of the computer vision models by visualizing the data using prototypes and learning more about the models' decision-making [14].
- Chapter 4: This thesis presents findings on a well-known vulnerability in the robustness of computer vision models referred to as adversarial attacks in related literature. First, the research presented in Chapter 4 shows how adversarial perturbations leave a detectable trace on the feature response map of CNNs, even though the input image remains identical. Then, feature response maps of CNNs are used with a simple and effective algorithm to detect adversarial attacks with a very competitive accuracy compared to state-of-the-art methods [15].
- Chapter 5: Motion artifacts in medical images are a common problem, especially for lengthy acquisition times. This work provides a data-driven solution based on supervised learning to reduce motion artifacts where no analytical solution exists. The proposed solution addresses motion reduction in two reconstruction methods (analytical and iterative) and reduces artifacts in raw data (acquired projections) and reconstructed scans (volume domain). The target domain of this method is cone-beam computed tomography (CBCT) scans, which are used for automatic segmentation and dose calculation in cancer therapy. In this thesis, we present techniques for training models on simulated data that achieve an improvement of over 6 *dB* in terms of signal-to-noise ratio (PSNR). Moreover, the proposed models generalize to real-world data, and clinical experts have verified their performance in the first attempt at motion compensation for CBCT scans.
- Chapter 6: Optimizing ML and DL models and finding the best models and architectures for small datasets is an intriguing area of research. This

thesis presents the most relevant findings from research in automated machine and deep learning (AutoML and AutoDL). The experiments in the context of automated machine learning show that optimizing the parameters of Gaussian processes as surrogate models for hyperparameter spaces (HPs) is the most successful method for HP tuning and meta-learning in AutoML. Moreover, the experimental results in the context of automated deep learning show that regularization and augmentation are the keys for fitting computer vision models to small datasets, that pre-trained models consistently outperform randomly initialized ones, and that large classifiers train faster than smaller ones [272, 271].

- Chapter 6: Domain adaptation and merging datasets from multiple data sources in medical imaging is a current research challenge. This thesis proposes an autoencoder-based architecture trained using an adversarial loss to preprocess 2D computed tomography scans for merging multiple datasets with minimal changes in the original scans. The proposed method extends classical training, validation, and testing performance to evaluate cross-dataset generalization and improves the cross-dataset performance for COVID detection from lung CT scans by over 10% [11].
- Chapter 6: This thesis presents relevant findings on the measurement of different sorts of biases in face recognition (FR) systems and the relationship between algorithmic bias and awareness. First, after analyzing the results of different models and network embeddings, this work concludes that awareness is not a good proxy for measuring racial bias in FR systems. Second, this thesis presents evidence that models which are designed to be unaware of race are not necessarily unbiased and suggest that further measures are critical for achieving fairness in FR systems [87, 295].

1.4 Publications

This section presents the list of peer-reviewed and published research papers connected to this thesis, divided based on the publication venue into two groups of journal and conference contributions.

1.4.1 Journal Papers

The following is a list of peer-reviewed and published research papers in scientific journals contributing to this thesis:

- **Mohammadreza Amirian**, Javier Montoya, Thilo Stadelmann, Frank-Peter Schilling, Rudolf Marcel Fuchslin, Ivo Herzig, Peter Eggenberger Hotz, Lukas Lichtensteiger, Marco Morf, Alexander Züst, Pascal Paysan, Igor Peterlik, and Stefan Scheib. “Mitigation of motion-induced artifacts in Cone Beam Computed Tomography using Deep Convolutional Neural Networks.” *Journal of Medical Physics*, pp. 6228-6242 (2023) [12].
- Ivo Herzig, Pascal Paysan, Stefan Scheib, Alexander Züst, Frank-Peter Schilling, Javier Montoya, **Mohammadreza Amirian**, Thilo Stadelmann, Peter Eggenberger Hotz, Rudolf Marcel Fuchslin, Lukas Lichtensteiger. “Deep learning-based simultaneous multi-phase deformable image registration of sparse 4D-CBCT”. *Medical Physics*, pp. e325-e326 (2022) [98].
- Samuel Wehrli, Corinna Hertweck, **Mohammadreza Amirian**, Stefan Glüge, and Thilo Stadelmann. “Bias, awareness, and ignorance in deep-learning-based face recognition.” *AI and Ethics*, pp. 1-14 (2022) [295].
- Lukas Tuggener, **Mohammadreza Amirian**, Fernando Benites, Pius von Däniken, Prakhar Gupta, Frank-Peter Schilling, and Thilo Stadelmann. “Design patterns for resource-constrained automated deep-learning methods.” *AI*, pp. 510-538 (2020) [271].
- **Mohammadreza Amirian**, and Friedhelm Schwenker. “Radial basis function networks for convolutional neural networks to learn similarity distance metric and improve interpretability.” *IEEE Access*, pp. 123087-123097 (2020) [14].
- Patrick Thiam, Viktor Kessler, **Mohammadreza Amirian**, Peter Bellmann, Georg Layher, Yan Zhang, Maria Velana, Sascha Gruss, Steffen Walter, Harald Christhelm Traue, Daniel Schork, Jonghwa Kim, Elisabeth André, Heiko Neumann, and Friedhelm Schwenker. “Multi-modal pain intensity recognition based on the senseemotion database.” *IEEE Transactions on Affective Computing*, pp. 743-760 (2021) [265].

- Taye Girma Debelee, Abraham Gebreselasie, Friedhelm Schwenker, **Mohammadreza Amirian**, Dereje Yohannes. “Classification of mammograms using texture and cnn based extracted features.” *Journal of Biomimetics, Biomaterials and Biomedical Engineering*, pp. 79-97 (2019) [55].
- Markus Kächele, **Mohammadreza Amirian**, Patrick Thiam, Philipp Werner, Steffen Walter, Günther Palm, and Friedhelm Schwenker. “Adaptive confidence learning for the personalization of pain intensity estimation systems.” *Evolving Systems*, pp. 71-83 (2017) [116].
- Markus Kächele, Patrick Thiam, **Mohammadreza Amirian**, Friedhelm Schwenker, Günther Palm. “Methods for person-centered continuous pain intensity assessment from bio-physiological channels.” *IEEE Journal of Selected Topics in Signal Processing*, pp. 854-864 (2016) [117].
- Kamran Kazemi, **Mohammadreza Amirian**, Mohammad Javad Dehghani. “The S-transform using a new window to improve frequency and time resolutions.” *Signal, Image and Video Processing*, pp. 533-541 (2014) [124].

1.4.2 Conference Papers

Here is the list of the peer-reviewed and presented research papers in scientific conferences contributing to this thesis:

- **Mohammadreza Amirian**, Javier A. Montoya-Zegarra, Jonathan Gruss, Yves D. Stebler, Ahmet Selman Bozkir, Marco Calandri, Friedhelm Schwenker, and Thilo Stadelmann. “PrepNet: A Convolutional Auto-Encoder to Homogenize CT Scans for Cross-Dataset Medical Image Analysis.” In 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 1-7 (2021) [11].
- **Mohammadreza Amirian**, Lukas Tuggener, Ricardo Chavarriaga, Yvan Putra Satyawana, Frank-Peter Schilling, Friedhelm Schwenker, and Thilo Stadelmann. “Two to trust: Automl for safe modelling and interpretable deep learning for robustness.” In International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning, pp. 268-275 (2021) [16].
- Stefan Glüge, **Mohammadreza Amirian**, Dandolo Flumini, and Thilo Stadelmann. “How (not) to measure bias in face recognition networks.” In Proceedings of the IAPR Workshop on Artificial Neural Networks in Pattern Recognition, pp. 125-137 (2020) [87].

- **Mohammadreza Amirian**, Katharina Rombach, Lukas Tuggener, Frank-Peter Schilling, Thilo Stadelmann. “Efficient deep CNNs for cross-modal automated computer vision under time and space constraints.” In Proceedings of the ECML-PKDD 2019, pp. 16-19 (2019) [13].
- Lukas Tuggener, **Mohammadreza Amirian**, Katharina Rombach, Stefan Lörwald, Anastasia Varlet, Christian Westermann, and Thilo Stadelmann. “Automated machine learning in practice: state of the art and recent results.” In Proceedings of the 6th Swiss Conference on Data Science (SDS), pp. 31-36 (2019) [272].
- Thilo Stadelmann, **Mohammadreza Amirian**, Ismail Arabaci, Marek Arnold, Gilbert François Duivesteijn, Ismail Elezi, Melanie Geiger, Stefan Lörwald, Benjamin Bruno Meier, Katharina Rombach, and Lukas Tuggener. “Deep learning in the wild.” In IAPR Workshop on Artificial Neural Networks in Pattern Recognition, pp. 17-38 (2018) [250].
- Benjamin Bruno Meier, Ismail Elezi, **Mohammadreza Amirian**, Oliver Dürr, and Thilo Stadelmann. “Learning neural models for end-to-end clustering.” In Proceedings of the IAPR Workshop on Artificial Neural Networks in Pattern Recognition, pp. 126-138 (2018) [180].
- **Mohammadreza Amirian**, Friedhelm Schwenker, and Thilo Stadelmann. “Trace and detect adversarial attacks on CNNs using feature response maps.” In Proceedings of the IAPR Workshop on Artificial Neural Networks in Pattern Recognition, pp. 346-358 (2018) [15].
- Viktor Kessler, Patrick Thiam, **Mohammadreza Amirian**, Friedhelm Schwenker. “Pain recognition with camera photoplethysmography.” In Proceedings of the Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1-5 (2017) [128].
- Viktor Kessler, Patrick Thiam, **Mohammadreza Amirian**, Friedhelm Schwenker. “Multimodal fusion including camera photoplethysmography for pain recognition.” In Proceedings of the International Conference on Companion Technology (ICCT), pp. 1-4. (2017) [127].
- Taye Girma Debelee, **Mohammadreza Amirian**, Achim Ibenthal, Günther Palm, Friedhelm Schwenker. “Classification of mammograms using convolutional neural network based feature extraction.” International Conference on Information and Communication Technology for Development for Africa, pp. 89-98 (2017) [54].
- **Mohammadreza Amirian**, Markus Kächele, Günther Palm, and Friedhelm Schwenker. “Support vector regression of sparse dictionary-based features for view-independent action unit intensity estimation.” In Proceedings

of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 854-859 (2017) [8].

- **Mohammadreza Amirian**, Markus Kächele, Patrick Thiam, Viktor Kessler, Friedhelm Schwenker. “Continuous multimodal human affect estimation using echo state networks.” In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. pp. 67–74 (2016) [10].
- **Mohammadreza Amiria**, Markus Kächele, Friedhelm Schwenker. “Using radial basis function neural networks for continuous and discrete pain estimation from bio-physiological signals.” In Proceedings of the IAPR Workshop on Artificial Neural Networks in Pattern Recognition. pp. 269-284 (2016) [9].
- Markus Kächele, Patrick Thiam, **Mohammadreza Amirian**, Philipp Werner, Steffen Walter, Friedhelm Schwenker, Günther Palm. “Multimodal data fusion for person-independent, continuous estimation of pain intensity.” In Proceedings of the International Conference on Engineering Applications of Neural Networks, pp. 275-285 (2015) [118].

1.4.3 Book Chapter

Here is the contribution published as a book chapter in conjunction with this thesis:

- Lukas Hollenstein, Lukas Lichtensteiger, Thilo Stadelmann, **Mohammadreza Amirian**, Lukas Budde, Jürg Meierhofer, Rudolf M Fuchsli, Thomas Friedli. “Unsupervised learning and simulation for complexity management in business operations.” Applied Data Science. pp. 313-331 (2019) [102].

1.5 Organization of Thesis

The remainder of this thesis is organized as follows:

- Chapter 2 provides an overview of the necessary prerequisites and theoretical background for understanding this thesis, summarizes the related work, and relates the following chapters to the current literature. This chapter begins with preliminary content, such as the fundamentals of convolution operation and self-attention. The chapter continues with best practices in architecture search and hyperparameter tuning methods.
- Chapter 3 introduces the main theoretical contribution of this thesis, namely the use of RBF networks as classifiers of CNNs for interpretable decisions. This chapter also proposes changing the training process and introduces a novel quadratic activation function to adapt RBFs for optimization with conventional CNNs.
- Chapter 4 demonstrates how understanding neural networks using feature response map visualizations can improve their robustness by detecting adversarial attacks. In addition, this chapter explains guided backpropagation, a well-known technique for inverting CNN architectures and visualizing the regions of input images which are relevant to the model's classification, and shows the application of feature responses in detecting adversarial attacks.
- Chapter 5 presents the main practical contribution of this thesis, in which neural networks reduce motion artifacts from CBCT scans for various reconstruction techniques. This chapter describes the first attempt to reduce motion artifacts in CBCT scans. It explains the architecture and underlying idea of how supervised learning with simulated data can address a solution for a real-world problem in which there are no ground-truth labels. This chapter includes a clinical evaluation of this method using real-world data and shows how improvement in numerical measures translates to the preferences of clinical experts.
- Chapter 6 provides an overview of several applications in conjunction with this thesis. This chapter aims to draw the readers' attention to several practical problems with ongoing research, present related solutions, and suggest promising areas for future research in these applications.
- Chapter 7 concludes the thesis and discusses a roadmap for future research opportunities in the niches to which this research work contributes.

2 Theoretical Foundations

This chapter summarizes the general prerequisites and theoretical background necessary to understand the remainder of the thesis. Specific niche techniques used in each part of the scientific and applied contributions are explained in each chapter individually. Therefore, reading this chapter is recommended only for those interested in refreshing their fundamental knowledge of computer vision techniques. Furthermore, the following chapters contain a more detailed theoretical overview of the related concepts, a knowledgeable reader can read them independently of this chapter.

This thesis’s main fundamental and methodological contributions are related to computer vision techniques for object recognition. First, this chapter explains the basics of convolutional neural networks (CNNs) in Section 2.1, which have revolutionized computer vision research by outperforming the classical methods. Second, this chapter briefly reviews the history of CNNs, including their major architectures and exciting recent developments. CNN-based models are the most recurring theoretical theme in this thesis, and understanding these models is important for being able to follow Chapter 3, Chapter 4, and parts of Chapter 6. The brief introduction to 3D-CNNs at the end of Section 2.1 is also necessary for understanding Chapter 5.

Moreover, this chapter briefly summarizes computer vision techniques using vision transformers (ViTs) in Section 2.2.6. Researchers investigating machine and deep learning (ML and DL) methods have long sought efficient methods for modeling attention-inspired mechanisms to focus on the most relevant information in time series, images, and to fuse information from several data modalities. Transformers and self-attention provide an excellent solution for attention in natural language processing (NLP). Transformers have recently been applied to computer vision problems by adapting self-attention for object recognition and segmentation. ViTs belong to more recent research compared to CNNs. Although their underlying theory only supports Section 6.5 in this thesis. ViTs are more likely to gain more attention in future computer vision research because of their ability to train on very large datasets compared to CNNs.

Deep learning has opened a great opportunity to distill information from massive datasets and optimize millions of parameters. However, these methods depend on optimization techniques that converge rapidly to an optimum which generalizes well. Therefore, understanding optimization techniques is necessary to bring the computer vision models to their optimal performance. The computer vision models presented in this thesis can overcome challenging problems that occur when using enormous datasets. These models are prone to overfitting, but they can be optimized to make correct predictions for the training data. However, the models cannot generalize to the unseen data, as is expected and observed in human vision. The last two sections of this chapter discuss the optimization methods, how to avoid overfitting, and how to improve the performance of computer vision architectures in generalization tasks with unseen data. The optimization and generalization of vision models are not the direct subjects of any chapter in this thesis; they are running themes throughout all chapters, especially in Chapter 3, Chapter 5 and parts of Chapter 6.

2.1 Convolutional Neural Networks

This section reviews the basics as well as recent advances in developing CNNs for computer vision. It begins with explanations of the building blocks used in CNNs. The computer vision community initially focused on manually improving these models' internal building blocks and introduced novel and suitable building blocks. The focus changed to automated model developments and architecture search when compute resources became widely available. After explaining the basics, this section presents some of the architectural breakthroughs that have improved the accuracy of computer vision models.

Automated neural architecture search has replaced manual architecture development attempts in the next generation of image processing models. Therefore, this section also describes some of the efforts in the automated search for optimal computer vision neural architectures. Finally, this section concludes with an explanation of the basics of 3D-CNNs and UNet architectures, which are necessary for understanding the content in the final chapters of this thesis.

CNN backbones have replaced hand-crafted feature extraction techniques such as scale-invariant features (SIFT) [163] because they can automatically learn representations of images during the optimization process. Based on this analogy, a model can be divided into two parts: 1) an encoder that converts the visual information (e.g., images) into a set of latent (intermediate) representations (model embeddings), and 2) a classifier that identifies the existing objects or segment patterns in the images. The main advantage of CNNs over manual feature extraction is the ability to optimize and fine-tune millions of parameters for encoding visual information into discriminative representations using large datasets. Much com-

puter vision research focuses on optimizing models' architecture to compute more generic representations (embeddings) of images. The ultimate goal of this area of research is not only to develop models that can learn image representations but also to train models that generalize well to unseen images from the same data distribution as the training data. A computer vision model's ultimate goal is learning representations that generalize to new categories of images outside the data sets used for optimization. Although the encoder part of neural networks has been the subject of much recent research, feed-forward neural networks have often been chosen for classifiers in the literature because of their efficiency in optimization.

2.1.1 Convolution Operator

The convolution operator of two functions shows how two input functions change their shape when shifted against each other for all possible shift values. For two one-dimensional real-valued time series (x and w), their convolution (s) can be defined as follows:

$$s(t) = \int x(a)w(t-a)da = (x * w)(t) \quad (2.1)$$

where t is the time, and $*$ denotes the convolution operator. For a given time shift (t), the convolution of two time series is equal to the dot product of one multiplied by the mirrored and shifted version of the other. The convolution operator is commutative due to the time inversion in the definition of the function ($f * g = g * f$). Similarly, the convolution operator can be defined for two 1D discrete functions (X and W) with time stamps i and j within the validity range determined by m as follows:

$$S(i) = \sum_{j=1}^m X(j)W(i-j) = (X * W)(i) \quad (2.2)$$

Based on this interpretation of the 1D convolution operator, we can define the 2D convolution (S) for the two-dimensional image as follows:

$$S(i, j) = (I * W)(i, j) = \sum_m \sum_n I(m, n)W(i-m, j-n) \quad (2.3)$$

where I and W represent two images, i and j define the spatial coordinates of these images. The valid range for images is indicated by m and n , respectively. The convolution function shows how a given kernel (W) changes an input image (I) after the kernel is applied. The commutativity properties also apply to

two-dimensional convolutions because of the mirroring the images. Since commutativity is not an essential property of neural networks, most libraries use *cross-correlation* instead of convolutions for implementation:

$$\hat{S}(i, j) = (I * W)(i, j) = \sum_m \sum_n I(m, n)W(i + m, j + n) \quad (2.4)$$

The *cross-correlation* function computes the dot product of an image patch and the kernel (W) by shifting the kernel vertically and horizontally over the input image in the range of the images' definition. The step size at which the kernel shifts after each convolutional step is called stride and is a parameter of a convolutional layer in neural networks.

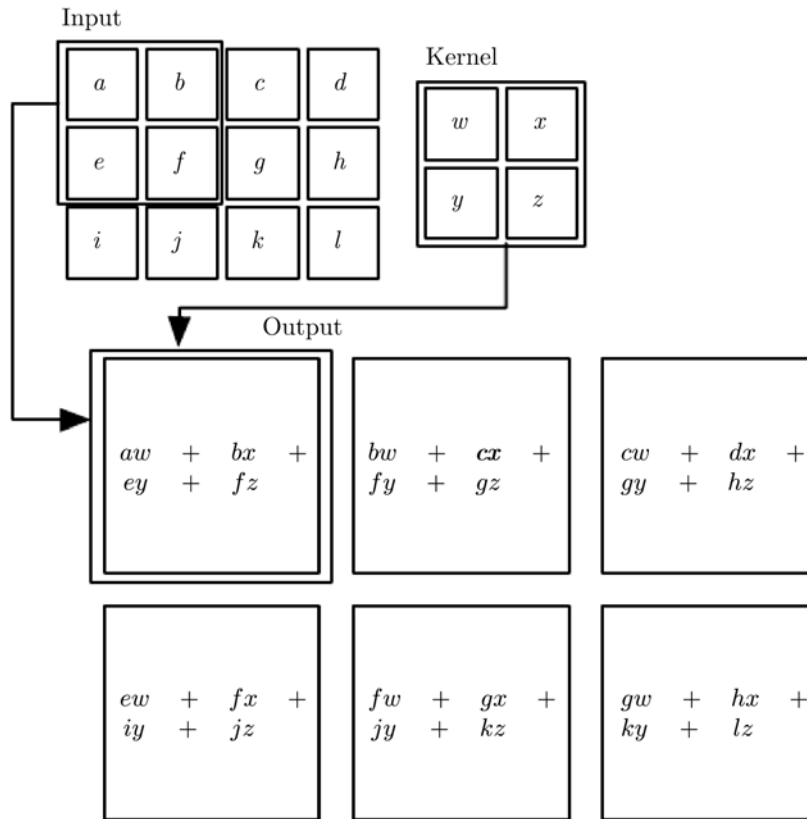


Figure 2.1: The cross-correlation function is often implemented in deep learning libraries for convolutional neural networks. For an input image, the output (kernel response) is the dot product of the vectorized kernel with a field which is the same size as the input image. The kernel slides over the entire image area with a given step size (figure adopted from [146]).

2.1.2 Feature Maps

Applying a kernel with the cross-correlation function in equation 2.4 to an image leads to computing a so-called feature map. Depending on the type of kernels, the feature maps contain different information (see figure 2.1). Feature maps are the first representations of the images computed in the CNNs, and visually inspecting them, along with the first layer inputs, is crucial for understanding the behavior of the entire network. The filters have the same depth as the input images (three for RGB and one for grayscale), and it is possible to visualize them along with the feature maps without complications.

2.1.3 Pooling Layers

The pooling layers aim to summarize the previous layer's output by merging the information in a given neighborhood. Two standard techniques for pooling local information are using the maximum or average value around the center of the kernel. The output of a pooling operator, as shown in Figure 2.2, is independent of the order of values in that specific region. The pooling layer and the cross-correlation function are the key components of the CNN for translation invariance as *inductive bias*¹ in CNNs.

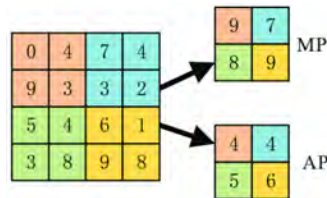


Figure 2.2: Max-pooling (MP) and average-pooling (AP) layers with kernel size and stride of 2 for CNNs (figure adopted from [291]).

2.1.4 FeedForward Neural Networks

Deep feedforward neural networks, also called multilayer perceptrons (MLPs), are often used in deep architectures to approximate functions. The goal of these networks is to approximate a function (f) that maps a set of input features (\mathbf{x}) to ground truth labels ($y \approx f(\mathbf{x})$). Feedforward networks, as shown on the right side of the figure 2.3, consist of an input, an output, and several hidden layers. Each hidden layer of the feedforward network is a *fully connected* layer that contains an intermediate representation of features by computing the weighted combination of

¹Inductive biases are a set of assumptions encoded in a learning algorithm to counter hypothetical input data and cases.

all features in the previous layer. Each fully connected layer in the architecture of an MLP can have a trainable bias term, denoted by x_0 in Figure 2.3 and trainable weights. Feedforward networks are optimized using *backpropagation* as described in the next sections.

2.1.5 Convolutional Neural Networks

The simplest form of convolutional nets consists of the two basic layers (convolution and pooling) explained in the previous sections. Basic networks can be constructed using successive convolutional layers to compute input representations and a pooling layer to summarize the information. However, modern convolutional architectures for vision use much more than these two layers. Figure 2.3 illustrates a simple convolutional network. The original image's pixel values reveal only the mapped object's information for the given pixel size. *Feature maps*, which show the representations of the first convolutional layer, combine the local information for a given filter size (usually 3×3). The pooling layer combines more local information from multiple filter activations (typically 2×2) and increases the size of the input region that contributes to a single activation value. The region's size in the original input image contributing to a single value at each network's layer determines the so-called *receptive field* at a given layer.

Finally, multiple convolutional and pooling layers are connected to a feedforward network for classification. A convolutional net, also called convolutional *backbone*, aims to compute discriminative (latent) representations of the images for each image class. These representations are finally transformed into the form of a feature vector in the last layer by flattening or global pooling. Flattening rearranges all the activations of a convolutional layer into a single vector, whereas global pooling applies the maximum and average functions to the spatial dimensions of the representations. The one-dimensional vector computed for each image is often referred to in the literature as *embeddings*. The *embeddings* of a convolutional neural network are passed to a feedforward network for object classification (Figure 2.3).

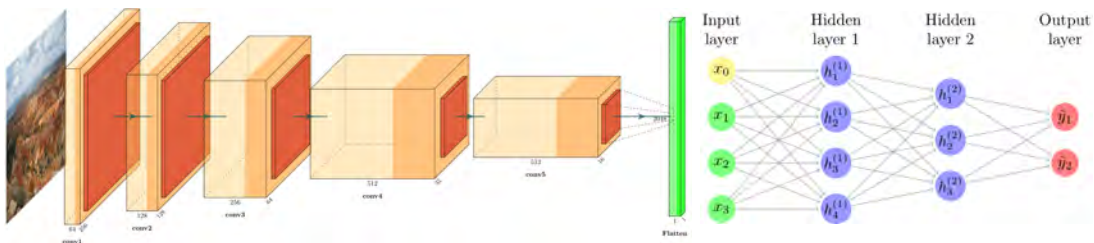


Figure 2.3: A convolutional neural network for representation learning from an input image, followed by a feedforward network for object classification.

2.1.6 Advanced Blocks

The recent history of convolutional neural networks has had many exciting breakthroughs. However, one of the first modern convolutional neural network prototypes (LeNet-5) was only able to classify handwritten digits [148]. Training AlexNet [140], a relatively small model compared to currently available networks, was only made possible by splitting the model between two graphics processing units (GPUs). Even two years after introducing AlexNet, it was impossible to train very deep VGG models end-to-end without pretraining the model layer by layer [243]. Given the limitations of resources and algorithms prior to the feasibility of automatic neural architecture search, computer vision researchers mainly tried to incorporate *inductive biases* to develop better convolutional architectures. These techniques are inspired by image processing tasks and failure cases in the classification task or improvement of the optimization process and gradient flow. The following sections review three interesting architectural advances in computer vision.

2.1.6.1 Residual Connections

Residual connections in CNNs establish a bridge between the input and output of a layer [96]. Although using a stack of multiple convolutional layers and forming deep models showed better generalization properties than shallow networks, the researchers have traditionally proposed residual connections to improve gradient flow on the backward path. Researchers introduced the idea of using residual connections at the same time that gradient vanishing was the focus of research in computer science for long short-term memory (LSTM) models [100]. The improvement of the gradient flow also led to the breakthrough of highway networks at the same time [249]. However, in the following years, residual networks were more commonly used in the research community. Figure 2.4 from the original paper shows one of the most straightforward and practical ideas in the history of deep learning.

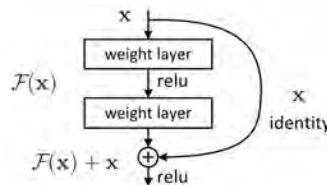


Figure 2.4: The residual connection between a layer’s input and output improves the gradient flow (figure is adopted from [96]).

2.1.6.2 Inception blocks

The original idea of inception blocks is to summarize the sparse latent representations of image patches into a dense form and cluster the relevant samples using convolutional filters with different patch sizes. Inspired by Arora et al. [19], the naive inception block finds the correlations between image patches or representations and clusters them into groups and units of highly correlated samples. Szegedy et al. [260] suggested using a layer of 1×1 convolutions to cover a small region with many clusters, which is practical for regions where clusters are densely distributed. Furthermore, larger convolutions of size 3×3 and 5×5 are used for the more spatially spread clusters. Inception blocks also include a pooling operator to maintain the translation invariance property (see Figure 2.5a).

The concept of a naive inception block is immensely appealing; however, it suffers from practical feasibility since the computational cost of such blocks blows up within the first few layers. Thus, to reduce the computational complexity of naive inception blocks, they are implemented with 1×1 filters to downsample the input representations in practice, while the outputs of the layers are computed by concatenating the representations of the input computed using all four sets of filterbanks depicted in Figure 2.5b. As a result, the inception models improved state-of-the-art performance in image recognition tasks after their advent.

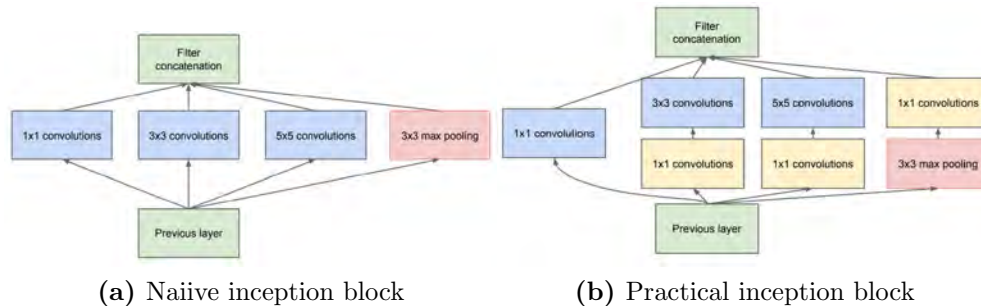


Figure 2.5: This figure, adopted from [260], depicts the idea of the inception blocks and their practical implementation.

2.1.6.3 Convolutional Block Attention Module

The main goal of the attention module in convolutional layers is to provide the ability to focus on a specific channel as well as spatial information [302]. Therefore, this module uses a channel attention module similar to squeeze and excitation techniques [105] in addition to a very similar spatial attention module. The high-level idea, shown in Figure 2.6a, is to compute a channel and a spatial attention map for the input of a given layer and multiply the activation values by these

maps to focus on specific channel-spatial information. The whole convolutional block attention module (CBAM) can profit from residual connections to improve gradient flow and allow the model to skip the attention modules.

The computation of the attention maps is relatively straightforward, as shown in Figure 2.6. To compute the channel attention maps (see Figure 2.6b), we first employ a mean and a max pooling over the input feature maps to obtain a vector of the average and maximum of the activation values for each channel. Then, these two vectors are passed through a trainable MLP with shared weights for both pooling outputs. Finally, the activations of the MLPs are averaged and passed through a sigmoid activation to form the final channel attention maps. A similar system is used for spatial attention mechanisms by computing the average and max-pooling over the spatial information instead of the channels and by replacing the MLP with a convolutional layer (see Figure 2.6c).

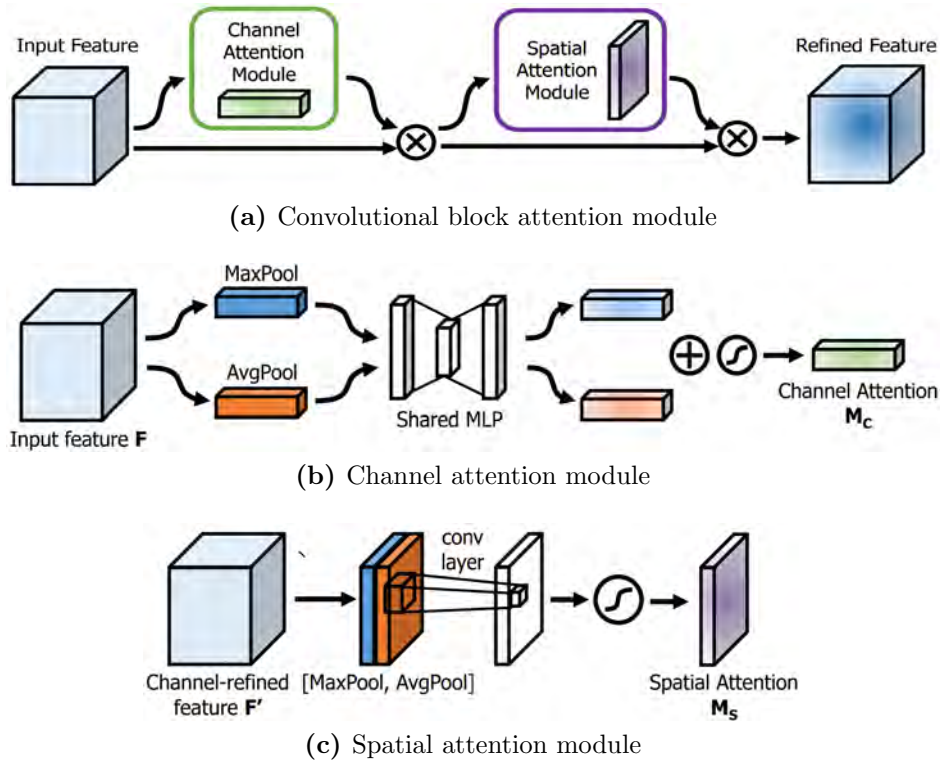


Figure 2.6: Convolutional block attention module (CBAM) with its two main components for refining channel and spatial features (figures are adopted from [260]).

2.1.7 Architecture Search

With the rapid increase in computer resources, computer vision researchers began to find ways to expand their search space, from optimizing hyperparameters to exploring new model architectures. Scientists who intuitively invented novel blocks for imaging models began to use their intuition to find the best search space for computer vision architectures and to optimize search techniques. The remainder of this section presents two breakthroughs in neural architecture search.

2.1.7.1 NASNets

Zoph et al. [327] introduced NASNets in the first famous attempt to search for the optimal architecture for image recognition. They performed the architecture search on a dataset with images of size 32×32 pixels from 10 object classes (CIFAR10 [139]). However, the heuristics and inductive biases allowed successful scaling of the sought after architectures to a large dataset with 299×299 images of 1000 classes (ImageNet [58]). Zoph et al. designed a controller using recurrent neural networks (RNN) to find the optimal architecture of two motifs named *normal cell* and *reduction cell*. The convolutional architecture is constructed using a stack of such searched architectures. They considered the number of initial convolutions and motif repetitions as free parameters to solve the problem of scaling from a small dataset (CIFAR10) to a larger dataset (ImageNet). Although this research improved state-of-the-art image recognition performance by 1.2% and reduced the number of best model parameters by 28%, it was only the beginning of more exciting research in this area.

2.1.7.2 EfficientNets

Tan and Le made the next breakthrough in the search for a neural architecture with a model that achieved the same performance as state of the art in image recognition, importantly it was $8.4\times$ smaller and $6.1\times$ faster [264] than competitor models. Their research focused on two directions: 1) improving architectural search and 2) introducing a compound scaling technique. As in their previous research on developing mobile neural architectures for searched networks (MnasNet [263]), Tan and Le used a reinforcement learning (RL) based method to optimize their objective function. Their objective function is to find a Pareto-optimal mobile network called EfficientNet. It includes two components: 1) maximizing the accuracy of the network, similar to MnasNets (mobile NasNets), and 2) minimizing the number of FLOPs (required floating point computations) instead of latency, which is considered in MnasNets. Inspired by the architecture of mobile networks (MobileNet and MobileNetV2 [104, 225]), the authors used the mobile inverted bottleneck (MBCon [225]) as the building block of Efficient-

Net. This work’s second breakthrough was introducing a compound approach to scaling neural architectures, while maintaining a balance between their height, width, and depth. Their scaling method demonstrates improvements in scaling EfficientNets, MobileNets, and ResNets.

2.1.8 3D Convolutional Neural Networks

Researchers have extended the idea of two-dimensional convolutions to three-dimensional spaces where data samples span multiple images (slices) per input instance, such as in videos and volumetric medical images [185]. Although the goal of video processing and 3D medical image processing is different, both can utilize 3D convolutional neural networks with the same architectures as in Figure 2.7. 3D convolutional neural networks (3D-CNNs) aim to find temporal dependencies in video processing [110] and 3D spatial dependencies in medical imaging and point clouds [317]. As an extension of 2D filters, 3D filters have one dimension higher - a size of $3 \times 3 \times 3$ voxels² is a common choice - to find spatial or temporal information in (3D) volumetric data. The feature responses of 3D filters are computed similarly by finding the correlation between the filter and a particular spatial position of the data volume. Applying a single 3D filter to a data volume results in a 3D feature map calculation. Stacking the feature maps of multiple 3D filters results in a four-dimensional feature map at each layer of a 3D model. The pooling operation is extended to find a volume’s average or maximum value with the typical size of $3 \times 3 \times 3$. Similar to techniques used with 2D-CNNs, such as flattening and global spatial pooling, the feature maps of the last layer can be converted into embeddings for classification. The high dimensionality of the feature maps of 3D-CNNs makes their implementation very memory intensive, and processing the 3D inputs increases the computational complexity of the 3D-CNNs. However, researchers have recently explored 3D-CNNs for medical applications as the memory limits of modern GPUs have increased significantly. It seems that 3D CNNs will receive more attention in the future as computational resources continue to develop.

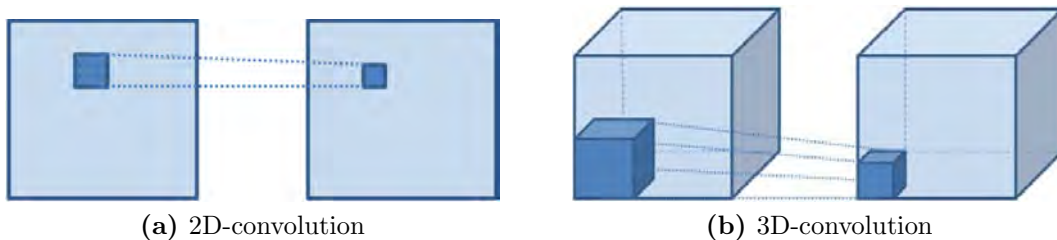


Figure 2.7: Two- and three-dimensional convolutions. 2D convolutions target images, while 3D convolutions are suitable for volumetric data (figures are adopted from [108]).

²Voxel is a single value in a data volume analogous to pixels in images.

2.2 Vision Transformers

Vision transformers (ViTs) for computer vision have emerged through the adaptation of the self-attention mechanism developed in the field of natural language processing (NLP) [279]. Researchers have searched for efficient attention mechanisms to optimally focus on the most relevant information to recognize patterns from different information sources. However, researchers in the field of NLP have only recently discovered a practical and efficient implementation of attention. The use of attention in NLP was so successful that the models developed in several NLP applications quickly outperformed the state-of-the-art [60, 158]. The core of these recent breakthroughs in NLP promptly found its way to image processing applications. The remainder of this section aims to review ViTs and explain the main components of these models used in computer vision. This section lays the theoretical foundation for the ViTs used in the following chapters of this thesis.

2.2.1 Preliminaries

The theoretical background of ViTs and attention mechanisms is grounded in machine translation. A brief explanation of the basic concepts is necessary to understand the rest of this section. Let us use a simple example from daily life to explain these concepts. Imagine that we make a text *query* to find a relevant research paper in a search engine. The search engine evaluates the query based on several *keys* that summarize the titles of the available papers and return the most relevant papers (*values*)³. Upon receiving a query, the search engine may use a *tokenizer* to segment the query sentence or break it into multiple *tokens* (words or punctuations). The model maps the tokens to their token IDs based on a particular tokenizer and pads it with zeros up to a certain length to form the *embedding* vectors of queries, keys, and values.

Dosovitskiy et al. introduced information conversion into tokens in image processing for the first time [65]. Based on this definition, commonly used in recent studies, they divided the input images into smaller patches of size 16×16 with three color channels. A random projection of the vectorized shape of these image patches is computed, and then the tokens that form the input of the vision transformer are generated. After tokenizing the information from any source, including images or text, the model focuses on the most relevant information in a query and relates them with a key to find the most appropriate values.

³The terms *key*, *query*, and *values* are used frequently in this section with very similar meanings to those used in the information retrieval literature.

2.2.2 Attention

The concept of attention, as first described in [21], is nothing more complicated than a weighted average of values (h) defined as follows:

$$\mathbf{c} = \sum_j \alpha_j \mathbf{h}_j \quad (2.5)$$

where $\sum_j \alpha_j = 1$. $\alpha_j = 1$ corresponds to the importance of each element in the vector \mathbf{h} .

Attention has been the key component in training outstanding models in NLP, such as BERT and RoBERTa [60, 158], through the use of keys, queries, and values from different sources in a supervised scenario. In addition, the attention mechanism is also used in self-supervised training of language models to predict missing information in training models such as GPT [32]. Attention can estimate dependencies between two sequences and can be extended to self-attention (SA) for modeling dependencies within a text sequence. Self-attention techniques are commonly used in image processing to find local correlations between tokens computed from the same image.

2.2.3 Self-Attention

The following steps describe how to compute the self-attention (SA) layer's output for an image ($(X) \in \mathbb{R}^{N \times T}$) converted into N projected patches (tokens): 1) Calculate the projections of all tokens based on three different matrices to compute the keys, queries, and values based on all tokenized image patches. 2) Compute the attention matrix by multiplying keys and queries and normalizing the results using the softmax operation, which is defined for a vector x as follows: $\text{softmax}(\mathbf{x}) = \frac{\exp(\mathbf{x}_i)}{\sum_j \exp(\mathbf{x}_j)}$. 3) Multiply the attention matrix by the values to calculate the SA matrix. In practice, ViTs consist of a stack of several such SA layers, which provide the opportunity to compute the dependencies of each pixel to every other one and combine the correlations based on their importance using the attention matrix.

2.2.4 Positional Encoding

SA layers, as described, are an effective tool for finding correlations between individual pixels. However, after converting an image to patches and computing the tokens, the SA layer is invariant to the order of the input tokens. In other words, the SA layer is independent of the order of the patches in the input images

and ignores the order of the tokens in the input data. To address this deficiency of the transformers, researchers added an (absolute) positional encoding that considers the order of the input tokens in NLP models and the image patches in the ViTs. The vector of positional encoding contains additive information proportional to the absolute position of words in a phrase or patches in an image.

2.2.5 Relative Postional Encoding

Absolute positional encoding in transformers retains the spatial information of a single patch, but fails to account for the relative distances between various patches. Shaw et al. used relative positional coding to address this shortcoming of self-attention in ViTs [239]. First, relative positional coding computes a distance function between image patches. Then, it applies a function based on these distances to the attention matrix, instead of absolute positional encoding, which adds the positional encoding to the input tokens.

2.2.6 Vision Transformers for Classification

Dosovitskiy et al. trained the first vision of ViTs on ImageNet [65], three years after Vaswani et al. introduced the attention mechanism [279]. The architecture of their vision transformer, as shown in Figure 2.8, consists of a transformer encoder (backbone) with a multilayer perceptron (MLP) head for classification. The input images used for image recognition are divided into patches of size 16×16 . Then, each patch is converted (flattened) into a vector, and a linear projection is applied to compute the input tokens for the transformer architecture. The backbone of the transformer contains multiple layers of multi-head attention⁴. Positional embeddings corresponding to the position of the patches in the original image are added to the computed tokens. A stack of multiple layers of multi-head attention computes a deep representation of the input images. These latent representations of the input images are passed to the classifier to classify them into distinct categories, such as dog, cat, and car. The possibility of computing the correlation between each pixel in the input images via an attention matrix makes ViTs more powerful than CNNs for a given dataset; nevertheless, ViTs are prone to overfitting. However, the higher learning capacity of ViTs provides the opportunity to use more data for training. The larger version of ImageNet with more than 21,000 classes (ImageNet-21k) is useful for pre-training ViTs (usually, optimal performance of pre-trained CNN models is achieved with ImageNet-1k. Additional data was not as helpful as in the case of ViTs).

⁴Multi-head attention is an extension of the attention mechanism that computes multiple attention matrices with different weights from keys and queries and combines the results of these many self-attention layers [279].

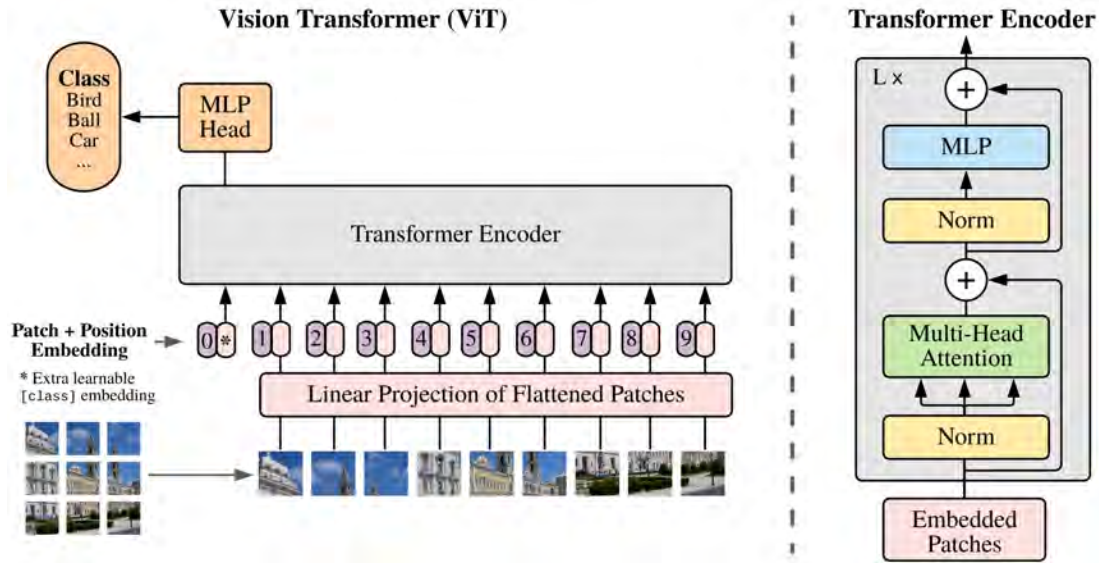


Figure 2.8: Vision transformers (ViTs) for image classification (figure is adopted from [65]).

2.2.7 Vision Transformers for Segmentation

ViTs can be extended from classification to semantic segmentation using similar encoding backbones. Figure 2.9 shows how pre-trained ViTs for classification can serve as the first building block of ViTs for semantic segmentation. The main difference between classification and segmentation ViTs starts after encoding the images into patch embeddings ($\mathbf{z}_L \in \mathbb{R}^{N \times D}$ with N patches and tokens of D dimensions). Classifiers then predict a vector with elements that sum to one, with the values being proportional to the probability of the predicted class. The segmenter ViTs, on the other hand, approximate a segmentation map $s \in \mathbb{R}^{H \times W \times K}$ that represents the segmentation predictions for each pixel of K classes in an image of a given height (H) and width (W). Two additional components of randomly initialized class embeddings and a mask transformer support the adaptation of ViT architecture to compute the segmentation map. After calculating the patch embeddings of the input images using pre-trained classification ViTs, the patch embeddings are concatenated with class embeddings ($([cls_1, \dots, cls_k] \in \mathbb{R}^{K \times D})$). The mask transformer includes several multi-headed self-attention layers where each class embedding attends each patch's pixel. At the end of the mask transformer, the normalized patch embeddings $\mathbf{z}'_L \in \mathbb{R}^{N \times D}$ and the class embeddings are separated, normalized based on their ℓ_2 , and a scalar dot product of each class embedding and patch embedding is computed to create a mask for each class. To predict the final image masks, the segmentation model includes an *argmax* function to find the most likely class per pixel and reduce the predictions to the same size as the input image.

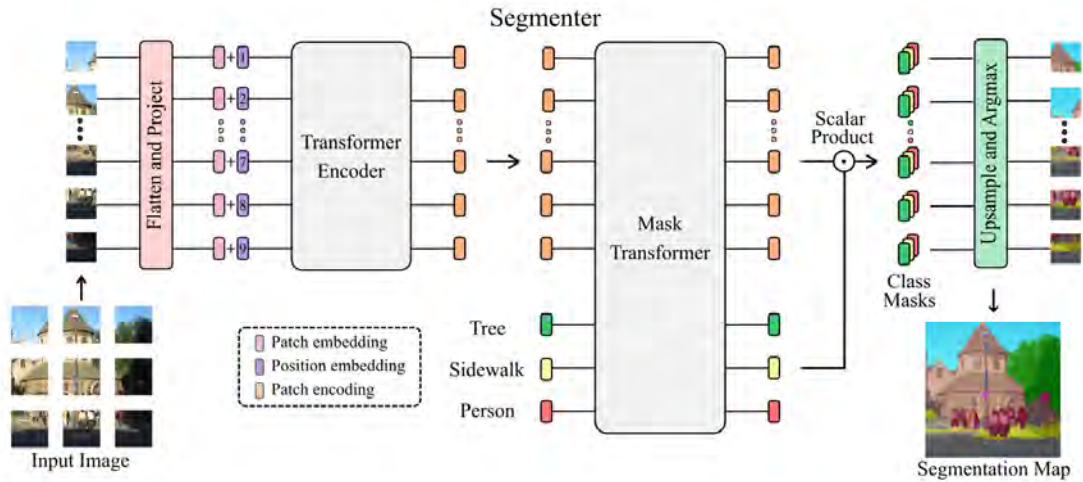


Figure 2.9: Vision transformers (ViTs) for image semantic segmentation (figure is adopted from [253]).

2.3 Optimizing Neural Networks

So far, this chapter has explained several aspects of different models for image processing. This section describes how these models are optimized to fit a given dataset and find patterns in the images within a dataset. The goal of the optimization process is tuning the trainable parameters of the models (θ) for an objective function (L) such that the model can generalize to unseen data.

2.3.1 Optimizing Trainable Parameters

The optimization's objective function, the so-called loss function, reflects the dataset's target task. For example, a classifier has a loss function that provides the highest probability for the presence of the correct class in an input image. Likewise, the segmenter computes the highest probability for the object surrounding a single pixel. The goal of the optimization algorithms is to minimize the expected value (\mathbb{E}) of a loss function over the entire training dataset (\hat{p}_{data}) as follows [88]:

$$J(\theta) = \mathbb{E}_{\mathbf{x}, y \sim \hat{p}_{data}} L(f(\mathbf{x}; \theta), y) \quad (2.6)$$

where \mathbf{x} and y denote the pair of data samples and ground truth labels. Although the training process of ML and DL models uses the training data (\hat{p}_{data}) for optimization, the main goal is to find a model that fits the data distribution (p_{data}):

$$J_{\boldsymbol{\theta}}^* = \mathbb{E}_{\mathbf{x}, y \sim p_{data}} L(f(\mathbf{x}; \boldsymbol{\theta}), y) \quad (2.7)$$

where $J(\boldsymbol{\theta})^*$ is the expected value of the error over the data distribution (not just the training set). The main difference between ML and DL optimization and classical problems is that the loss function depends on the training data. The search for optimal parameters for an ML ($\boldsymbol{\theta}_{ML}$) for a maximum likelihood problem can be described as follows:

$$\boldsymbol{\theta}_{ML} = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^N p_{model}(\mathbf{x}^{(i)}, y^{(i)}; \boldsymbol{\theta}) \quad (2.8)$$

Maximizing the likelihood of predictions with ground truth labels is equivalent to minimizing the prediction error. For discrete pairs of data samples and labels (\mathbf{x} and y), generalization is expressed as follows:

$$J_{\boldsymbol{\theta}}^* = \sum_{\mathbf{x}} \sum_y p_{data} L(f(\mathbf{x}; \boldsymbol{\theta}), y) \quad (2.9)$$

The gradient of the loss function (\mathbf{g}) is calculated for all parameters using training data in practice to find the optimal model's parameters:

$$\mathbf{g} = \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}} = \mathbb{E}_{\mathbf{x}, y \sim \hat{p}_{data}}(\mathbf{x}, y) \nabla_{\boldsymbol{\theta}} L(f(\mathbf{x}; \boldsymbol{\theta}), y) \quad (2.10)$$

Theoretically, to use a gradient descent algorithm to optimize the neural networks based on the gradients of the parameters with respect to the loss function, we need to compute the average of gradients over the entire training dataset before updating the parameters. However, this method is computationally very expensive, and the optimization algorithm converges faster when multiple updates are made from subsets of the training dataset (*mini-batches*). Therefore, the gradient of the parameters with respect to the loss function is calculated for a mini-batch with a size of m samples:

$$\hat{\mathbf{g}} = \frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_i L(f(\mathbf{x}^{(i)}; \boldsymbol{\theta}), y^{(i)}) \quad (2.11)$$

Larger mini-batch sizes result in cleaner gradients toward the minima of the objective function, and gradients computed with smaller batch sizes are noisier.

However, this gradient noise can regularize the training process and improve the models' generalization. The use of mini-batches became a practical optimization approach due to their advantages in generalization and convergence speed. Neural networks and deep vision models consist of many layers with a depth of more than a hundred. The algorithm for computing the gradient of parameters for all layers is based on the chain rule, which is called *backpropagation*. After calculating the loss function at the end of the models' computational graph, its partial derivatives are calculated with respect to the trainable variables of the models' last layer. These gradients are backpropagated toward input images to compute the gradients of all parameters minimizing the loss function. Then, the gradients are multiplied by a learning rate, and the parameters are updated based on these multiplications. The iterative training process continues until a stopping criterion is met.

2.3.2 Optimization to Generalization

Training neural networks and computer vision models, as described in Section 2.3.1, focuses on minimizing the prediction error on the training set. However, the main goal of training is to optimize models that generalize well to the entire data distribution outside the training set. The art of bringing computer vision models to optimal performance involves many techniques and a lot of empirical trial and error. The main goal of these techniques is to limit the networks' capacity or artificially increase the amount of data by presenting different variations of the original dataset, which forces the model to learn more generic patterns instead of memorizing individual images from the training dataset. The remainder of this section describes some well-known methods for improving generalization after optimization.

Dropout

There are several reasons for poor generalization (overfitting) in deep neural networks described in the literature. These reasons include neurons' coadaptation to a particular image with poor generalization and learning dense representations of the input images. Srivastava et al. [248] proposed dropout as an effective technique to improve the neural network's generalization to address the above reasons for overfitting. Dropout is equivalent to randomly setting the activations of a layer for a given input image to zero with a certain probability during training. The random suppression of activations via dropout prevents the model from coadapting neurons. Furthermore, dropout leads to learning sparse representations of input images and consequently improving the generalization. Although the usage of dropout in the classifiers (fully connected layers or MLPs) is more common, it is possible to use dropout in convolutional layers as well. Srivastava

et al. [248] originally introduced dropout to reduce overfitting during training. Nonetheless, Gal and Ghahramani proposed a framework for estimating neural network uncertainty using dropout in test time as an additional application of dropout [81]. In their theoretical framework, dropout in neural networks has been successfully used for Bayesian inference in Gaussian processes to estimate uncertainty.

Regularization

The other technique to limit the neural networks' capacity is adding a penalty to the neural network loss function that increases with the absolute value of the trainable weights. Researchers used different functions of the trainable weights as additive penalties to the original (classification or segmentation) loss function. Tibshirani and Zheng used the ℓ_1 - norm of the weights to calculate such a penalty [267, 323]. The ℓ_1 - norm regularization, also called Lasso regression regularization, keeps the sum of the absolute values of the trainable parameters small. Lasso regularization leads to sparse weight vectors by setting some weights to zero. An alternative to Lasso regularization is to compute the penalty term using the square of the weights (ℓ_2 - norm), which leads to Ridge regression regularization [192]. The computed regularization penalty (loss) is multiplied by a regularization factor and then added to the loss value of the training. Loshchilov and Hutter have shown that decoupling the regularization penalty from the classification loss by defining an independent weight decay from the learning rate for adaptive gradient algorithms improves the generalization performance [162].

Augmentation

Besides dropping neurons from the neural network architecture and regularizing the weights, neural network generalization improves by presenting the models with different variations of the input data. Since the early days of deep learning research, studies have shown that CNNs have limited robustness to rotation and scaling [148]⁵. However, researchers quickly found that computer vision models can recover such weaknesses by presenting variations of input images to the network during training time. Thus, computer vision models trained with rotated versions of the input images are robust against rotations. *Augmentation* is the term proposed in the literature for training computer vision models with transformed images. Since then, researchers have used various strategies to augment their input images depending on the application. Techniques for augmentation such as shearing, translation, rotation, rectification, and changing contrast, color, brightness, and sharpness are broadly used in the computer vision research com-

⁵<http://yann.lecun.com/exdb/lenet/>

munity [42, 228, 242, 283, 140].

Similar to the architectural design literature explained in section 2.1.7, augmentation methods have also evolved towards automation. An important work of research on this subject resulted in *AutoAugment* [50], which presents optimal strategies that are automatically checked against one of the largest computer vision datasets (ImageNet [58]). Further improvement in the speed of such a resource-exhaustive search led to the development of *Fast AutoAugment*, an algorithm that is lighter in terms of computational complexity and more suitable for exploring optimal augmentation strategies on private datasets [156].

2.4 Related Work

Two critical components of the rapid increase in computational resources and datasets' size have revived machine and deep learning (ML and DL) techniques in practical applications for image pattern classification [147]. Initially, raw pixel values for simple tasks such as classifying handwritten digits were sufficient to train neural networks and support vector machines for pattern classification [46]. However, the urge to extract robust features for more complex computer vision problems led researchers to develop advanced methods for representation learning [164].

ML pipelines became increasingly complicated in the first decade of the twentieth century as problem-oriented feature extraction techniques grew rapidly [142]. Classical computer vision researchers, who moved away from Fourier transforms and brought image-based prior knowledge to multiscale wavelet transform, began training sparse dictionaries from data [221], and robust hand-crafted features [163] for pattern recognition received enormous attention. Moreover, the growth of datasets quickly saturated the performance of classical ML models, and task diversity made the search space for the best priors exhaustive. DL appeared as the next breakthrough to increase the capacity of models for massive datasets and automate feature extraction and representation learning in a wide range of different tasks [146].

The first DL milestone in computer vision was reviving the convolutional neural networks (CNNs) for pattern recognition [147]. CNNs introduced in the late 1980s finally found their way into practice by overcoming their high computational complexity. LeNet5 [148] is one of the first CNN models applied to handwritten digit classification algorithms, and AlexNet [80] is among the first successes of deep CNNs in image classification on large datasets. The second half of the 2010s is the most prosperous time in the history of DL and CNNs in computer vision with a lot of exciting research and developments regarding various architectures [243, 261, 225, 263] and optimization techniques [216, 129, 135, 162].

Researchers trained CNNs using neural architectures, search became fashionable, optimization methods evolved based on large datasets such as ImageNet [223], and this distilled knowledge was successfully applied to smaller datasets using transfer learning [197].

After CNNs matured in image classification [24] and segmentation [145], the next generation of research focused on automatic neural architecture search [69] and finding optimal search spaces for efficient networks with minimal delay [263] and computational power consumption for mobile applications [264]. However, the parallel increase in computational resources and the presence of massive datasets, motivated by data-driven artificial intelligence (AI) research, created the opportunity for the next breakthrough in computer vision. The next breakthrough occurred in the early 2020s by introducing vision transformers (ViTs) [65], and adapting self-attention, originally discovered in natural language processing (NLP) literature [279], for computer vision tasks. ViTs are widely used for image classification and segmentation, and these models have improved their performance with a larger version of the ImageNet dataset called ImageNet-21k with over 21,000 classes [213].

Researchers have expressed doubts and concerns about the robustness of computer vision models using CNNs [89] and their explainability [1] from their inception. CNNs lack some basic properties of classical methods, such as rotation equivariance⁶, despite their capability to learn translational equivariant features [148]. Due to the lack of rotation equivariance, the performance of CNNs decreases when the input images rotate [120]. Similar instabilities and inaccuracies have been reported due to changes in lighting conditions, contrast, image acquisition techniques, and overall data distribution drift [62]. Studies even show that CNNs focus more on the texture than the shapes when classifying objects [84]. Researchers also discovered that they could compute minimal perturbations, called adversarial attacks, for an input image to fool CNN models with images that are indistinguishable from one another to the human eye [89]. They even optimized a so-called universal adversarial attack that generalizes to many images [187]. Among all of the challenges mentioned in computer vision research related to robustness, this work presents a solution for rotational invariance in ViTs and adversarial attack detection in CNNs [15].

CNNs were known as powerful black-box models following a similar trend to many other ML and DL-based techniques in information processing [33]. These models can be used in many applications without additional reasoning; however, understanding these high-precision decisions is critical for applications that affect human safety and health, such as autonomous driving systems [20] or healthcare [6]. The literature that has developed around explainable AI (XAI) [268] and inter-

⁶A rotationally equivariant representation of an image rotates with the same angle as its input rotates. Edge detection filters, for example, are rotationally equivariant

pretation of computer vision models [318] is the result of researchers' concerns about the usage of black-box models in critical applications. Researchers have two main approaches to the interpretability and explainability of neural networks. The first group analyzes the trained models to understand the predictions using post-processing and post-hoc techniques [47]. Another group disagrees with the idea of interpretability solely as an add-on for neural networks. Instead, these researchers point to changing the design of the neural architectures so that the decisions are transparent and explainable [175]. Concerning the issue of explainability, this thesis proposes using radial basis neural networks as classifiers on top of the CNNs to provide more understandable information to humans about the decision-making of the models [14].

Despite all the challenges mentioned above, computer vision breakthroughs have found their way into numerous applications [5]. CNNs have outperformed all other methods in the majority of applications, such as object detection, recognition, and segmentation [24, 145]. Furthermore, CNNs perform well in other classical image processing tasks such as image denoising [63], super-resolution [308], and motion deblurring [256]. Moreover, the applications of CNNs extend not only to medical imaging for diagnosis [227] and automatic segmentation [290], but also to image quality enhancement and motion artifact reduction [157]. In addition to all the above theoretical contributions, this paper presents practical applications of ML and DL, especially in medical imaging, and demonstrates a variety of empirical findings [251].

3 RBF Classifiers for Explainable Computer Vision Using CNNs

Radial basis function neural networks (RBFs) are prime candidates for pattern classification and regression and have been used extensively in classical machine learning applications. However, RBFs have not been integrated into contemporary deep learning research, and computer vision has continued using conventional convolutional neural networks (CNNs) because of technical difficulties. This chapter presents the techniques to adapt RBF networks as a classifier on top of CNNs by modifying the training process and introducing a new activation function to train modern vision architectures end-to-end for image classification. The specific architecture of RBFs enables them to learn a similarity distance metric to compare and categorize similar and dissimilar images. Furthermore, this chapter demonstrates that using an RBF classifier on top of any CNN architecture provides new human-interpretable insights about the decision-making process of the vision models. Finally, RBFs are successfully applied to a range of CNN architectures, and their performance on benchmark computer vision datasets is presented in this chapter. This chapter is adopted from the research published in [14], licensed under CC BY 4.0 ¹.

¹<https://creativecommons.org/licenses/by/4.0>

3.1 Introduction

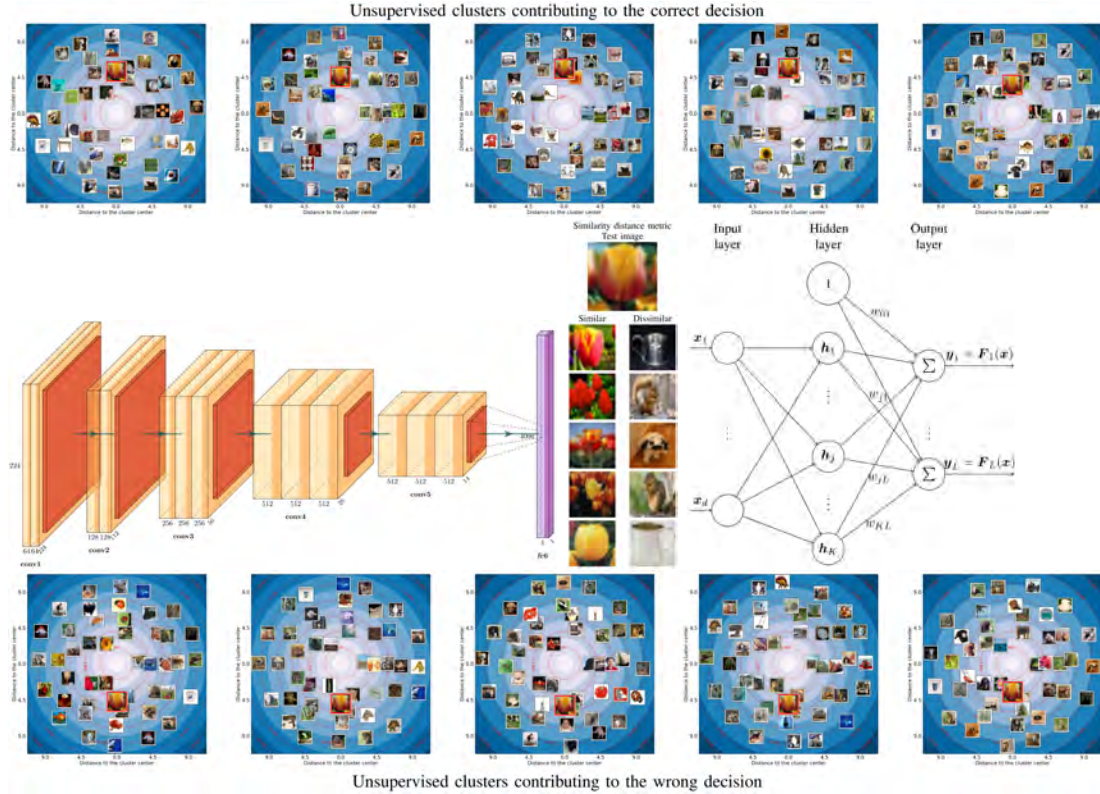


Figure 3.1: Figures on the top and bottom rows visualize the position of a test image in the clusters optimized using the unsupervised loss function. The output of CNN backbones is connected to RBFs' input through a fully connected layer, and the input features of the RBFs are referred to as embeddings in this chapter. The model compares the embeddings of each image with cluster centers using a trainable similarity distance metric. The same distance metric can be used to find similar and dissimilar images to a test sample amongst training images (visualized in the table in the middle row). The RBFs apply an activation function to the distance of the training images from the cluster centers to compute activation values. The output layer of the RBF is optimized for classification based on these activation values. The entire CNN-RBF architecture is optimized end-to-end with a specific initialization (figure adopted from [14]).

Inspired by the locally tuned response of biological neurons, Broomhead and Lowe introduced radial basis function neural networks (RBFs) in 1988 [31]. The modeling concept behind RBFs is a combination of unsupervised and supervised learning for pattern classification and regression. However, due to structural deficiencies, RBFs have not been integrated into contemporary research in computer vision using Convolutional Neural Networks (CNNs). This chapter presents developments in a new area of research and lays the foundation for using RBFs

in deep learning and computer vision by modifying their architecture and learning process. The results demonstrate that integrating RBFs into CNN models for computer vision provides a similarity distance metric and an interpretable decision-making process.

This chapter is motivated by RBF architectures' unique opportunities when used with CNN models because of their explainability and robustness compared to linear classifiers. The new training process introduced for RBFs in this chapter provides the opportunity to use labeled and unlabeled data by optimizing two loss functions combining supervised and unsupervised learning. Moreover, the training process of RBF architectures includes optimizing a distance metric that serves as a similarity distance metric to find similar and dissimilar images. Additionally, this chapter proposes visualization techniques to illustrate the clusters and activations with training and test images to gain more insights about the reason behind the decisions made by the networks, thus improving interpretability. The contributions of this chapter to computer vision literature can be summarized as follows:

- Combining supervised and unsupervised learning.
- Learning a similarity distance metric to find similar images.
- Improving the interpretability of decision-making.

Despite the advantages of combining RBFs with modern CNN architectures, two factors in the architecture and training process of RBFs hinder their integration into CNNs. First, the nonlinear activations and computational graphs of RBFs used in the literature prevent efficient gradient flow. Secondly, RBFs assume that the training features are fixed, so the cluster centers are initialized accordingly. Nonetheless, CNN architecture dynamically learns the embeddings used as input features of RBFs. This chapter tackles the limitations of the original RBFs and presents the following contributions to RBF literature:

- Introducing a quadratic activation function and a linear computational graph for end-to-end learning.
- Adding an unsupervised loss term to update the cluster centers in the training process with the learned embeddings.
- Applying the RBFs to computer vision in a first attempt at using deep CNN architectures.

The remainder of the chapter covers the related work in Section 3.2 followed by the theoretical background of RBFs in Section 3.3. Then, Section 3.4 presents

the original research and contributions with the proposed modifications to RBFs, followed by a visual explanation of the new proposed training and decision-making process in Section 3.6.1. The experimental results of applying the proposed RBF-CNN architectures using a range of CNN backbones on benchmark datasets are presented in Section 3.5. The potential contributions of the proposed similarity distance metric on computer vision to enhance the transparency of the decision-making process is demonstrated in Section 3.6.2. This chapter concludes with discussions and conclusions in Section 3.7.

3.2 Related Work

The research followed two approaches to optimize RBF architectures. The first approach concentrates on the training process and initialization of the networks, while the second aims to find superior activation functions. This chapter presents improvements in both research directions to integrate the RBFs into contemporary computer vision models using CNNs.

RBFs were originally introduced as supervised models for classification and regression tasks. Broomhead and Lowe initially proposed drawing the cluster centers either from a uniform distribution or randomly from the training samples and then optimizing the output weights using a pseudo-inverse analytic solution [31]. Initializing the cluster centers randomly and only training the output weights is called a one-phase training process for RBFs. Two-phase training for RBFs uses various methods to initialize the cluster centers before optimizing the output weights. Research since 1988 has used supervised and unsupervised methods to initialize the cluster centers. Moody and Darken proposed an unsupervised algorithm to initialize these cluster centers [186], while Schwenker et al. proposed supervised vector quantization [236]. Decision trees were used to find centers independently by [141] and [234] before training the output weights. Finally, Schwenker et al. proposed the third phase to optimize the entire RBF network end-to-end, including output weights, the cluster center, and trainable parameters of activation functions using gradient descent [235].

These methods for cluster center initialization assume a fixed feature space for the input layer. However, CNNs learn the embeddings automatically and develop the feature space of the images during the training process. Therefore, this research suggests optimizing an unsupervised learning loss during the training to cope with this change in the feature space. This work differs from previous research as it combines supervised and unsupervised learning by optimizing two separate losses simultaneously using gradient descent.

The technical requirements of new applications and implementations have motivated the use of several activation functions presented in the literature of RBFs [67].

The Gaussian function is the kernel developed by modeling the data through a multivariate Gaussian distribution [31]. Other functions adapted in the RBF architecture include linear kernels, thin-plate splines, logistic functions, and multi-quadratic functions [79, 208, 155, 40]. Hardy’s multiquadratic functions motivated an activation function for RBFs used by Karimi et al., and Zhao et al. [121, 322]. Du et al. proposed a kernel for digital signal processing (DSP) units 3.9. This chapter presents a novel quadratic kernel to build a linear computational graph for efficient gradient flow and RBF integration for end-to-end training with CNN architectures.

Besides the mature fundamental research, RBFs have been applied to many applications for pattern classification and regression in recent years. For example, Nicodemou et al. used RBF networks for 3D hand pose estimation [191], Dehghan and Mohammadi estimated a numerical solution for Fokker-Planck differential equations with RBFs [57], Li et al. used sparse multiscale RBFs for seizure detection in EEG signals [152]. Furthermore, Zhao et al. predicted interfacial interactions by training RBFs [322], and Geng et al. introduced deep RBF networks and applied the method to food safety inspection data. Finally, RBFs are used to train models for classification and regression in discrete and continuous pain quantification [9].

RBFs can be applied to computer vision tasks and image classification as well. Schwenker et al. used raw images as feature vectors to classify hand-written digits [235]. Er et al. extracted the features from facial images using principal component analysis (PCA) and processed these features using Fisher’s linear discriminant (FLD) technique before classifying the faces using RBFs [182]. However, the successful rise of modern CNNs, such as LeNet-5 [148] and AlexNet [224], led to a paradigm shift from using hand-crafted features to automated deep CNN-based feature and representation learning. In recent years, most computer vision tasks, like facial recognition [177], are dominated by modern CNN architectures as they present superior performance compared to classical methods for image processing. To the best of our knowledge, this chapter presents the first attempt to integrate RBFs into modern CNN architectures for computer vision.

This chapter relates to literature focusing on deep metric learning since RBFs automatically optimize a similarity distance metric during training based on their architecture. Euclidean distance, Mahalanobis distance, and cosine similarity have been used to evaluate the similarity between the embeddings (the features extracted from CNNs) of two images in the literature [101, 304, 300]. Researchers have applied different strategies and loss functions to optimize these similarity metrics for same-class images while also maximizing the distance of different-class images. The research in this area concentrates on the training process and the design of a loss function which brings similar images closer in the embedding space based on a similarity measure. Hu et al. proposed minimizing the inter-class

scores and maximizing the intra-class scores based on Euclidian distances [106]. Hoffer and Ailon suggested optimizing a similarity-based loss function defined for selected triplets of images [101]. Song et al. used the pairwise distances between images of an entire batch and proposed a structured loss function for metric learning [246]. Similar research work has aimed at optimizing angular distance, cosine distance, and large-margin Euclidean distance of similar and dissimilar images [287, 300, 41].

This chapter presents a method to retrieve a ranked list of similar and dissimilar images, leading to visually appealing similarity metric learning results. However, the proposed similarity metric learned by the RBFs does not require any complicated triplet sample selection or loss design. Instead, these results have been obtained using a typical supervised loss function for classification (softmax cross-entropy). Furthermore, RBFs can not only optimize for Euclidean and Mahalanobis distances but also for the entire covariance matrix.

3.3 Radial Basis Function Networks

This section briefly reviews and explains the theoretical foundation of radial basis function networks. RBFs are presented in the literature as a global approximation method for learning a mapping \mathbf{F} from a given feature space with the dimensionality of d to a label space with K dimensions ($\mathbf{F} : \mathbb{R}^d \rightarrow \mathbb{R}^K$) [31]. In this chapter, the function \mathbf{F} of features \mathbf{x} approximates the one-hot encoded labels \mathbf{y} . The features used to train the RBFs in this chapter are the embeddings of deep CNNs, which are used to predict the class labels using end-to-end optimization. A fully connected layer connects the CNN architectures and RBFs to provide compatibility between the two architectures. The architecture of the RBF consists of input layers, a single trainable hidden layer with C cluster centers (\mathbf{c}_j) 3.1, and an output layer.

During the evaluation, also known as inference in deep learning, the RBF computes a distance between embeddings of CNNs and the cluster centers and applies an activation function to this distance. The network outputs are then computed by multiplying the output layer weights with the activation values. This forward path of RBFs is formally defined as:

$$r^2 = (\mathbf{x} - \mathbf{c}_j)^T \mathbf{R}_j (\mathbf{x} - \mathbf{c}_j) \quad (3.1)$$

$$\mathbf{y}_k = \mathbf{F}_k(\mathbf{x}) = \sum_{j=1}^C w_{jk} h(\|\mathbf{x} - \mathbf{c}_j\|_{\mathbf{R}_j}^2) + w_{0k} \quad (3.2)$$

where r represents the distance, \mathbf{R}_j is the positive definite covariance matrix (trainable distance metric), T denotes the matrix transposition, w_{jk} shows the

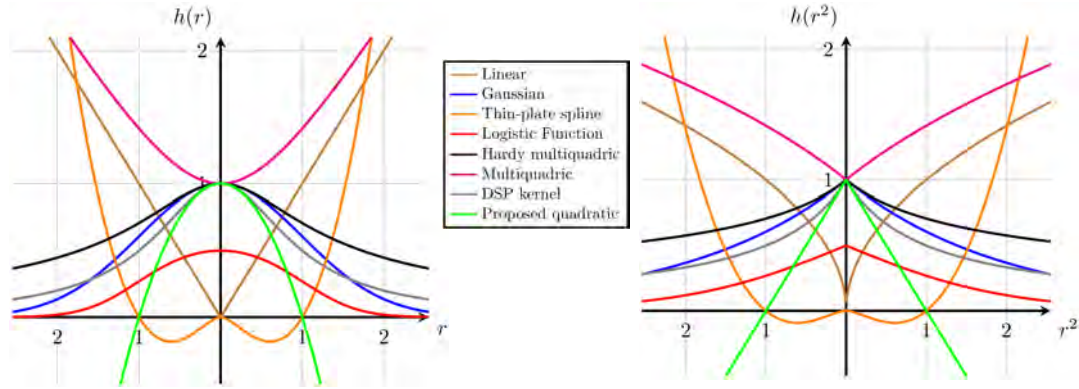


Figure 3.2: Activation functions for RBF networks. Here is the list of the parameters for depicting the kernels: $\sigma = 1$, $\alpha = 1/2$, and $\beta = 1/2$. The proposed quadratic activation kernel is linear based on the r^2 . Consequently, the CNN goes through a completely linear forward path, and thus, gradients are computed and backpropagated efficiently (figure adopted from [14]).

weights of the output layers, h is the activation function, and w_{0k} are the biases. In these Equations, μ , j , and k enumerate the number of samples, cluster centers, and classes. Trainable parameters in Equation 3.1 and 3.2 are the output weights, cluster centers, and covariance matrix.

Optimizing the RBF networks with an identity covariance matrix is equivalent to training in Euclidean space. It is possible to optimize a Mahalanobis distance [53] by training the main diagonal on the covariance matrix. Any arbitrary distance metric can be trained by optimizing the entire covariance matrix, while projecting the matrix to the space of positive definite matrices. The distance, r , computed in Equation 3.1 is not only a measure of the proximity of an image to a cluster center; it can also be used to compare images and find similar and dissimilar images in the embedding space.

The linear and nonlinear activation functions used in RBFs are as follows [208, 155, 40]:

$$\text{Linear :} \quad h(r) = r \quad (3.3)$$

$$\text{Gaussian :} \quad h(r) = e^{-r^2/2\sigma^2} \quad (3.4)$$

$$\text{Thin-plate spline :} \quad h(r) = r^2 \ln r \quad (3.5)$$

$$\text{Logistic function :} \quad h(r) = \frac{1}{1 + e^{(r^2 - r_0^2)/\sigma^2}} \quad (3.6)$$

$$h(r) = \frac{1}{(r^2 + \sigma^2)^\alpha}, \quad \alpha > 0 \quad (3.7)$$

$$h(r) = (r^2 + \sigma^2)^\beta, \quad 0 < \beta < 1 \quad (3.8)$$

$$h(r) = \frac{1}{1 + r^2/\sigma^2} \quad (3.9)$$

In addition to the standard machine learning activation kernels in Equations 3.3 to 3.6, the kernel presented in Equation 3.7 is derived from the generalized Hardy's multiquadratic function [79]. Du et al. [67] proposed the kernel in Equation 3.9 because of its convenience for implementation on DSP units. Various activation functions for RBFs are depicted in Figure 3.2.

The complete process of training RBFs was introduced by Schwenker et al. [235] as a three-phase process:

Unsupervised learning: This step aims to find cluster centers that are representative of the given dataset. The k-means [17] clustering algorithm is widely used for this purpose. K-means iteratively finds a set of cluster centers and minimizes the overall distance between cluster centers and members over the entire dataset. The target of the k-means algorithm can be written in the following form:

$$\text{Loss}_{\text{unsupervised}} = \sum_{j=1}^K \sum_{\mathbf{x}^\mu \in \vartheta_j} \|\mathbf{x}^\mu - \mathbf{c}_j\|^2 \quad (3.10)$$

where $\mathbf{x}^\mu \in \vartheta_j$ denotes the members of the j^{th} cluster shown by ϑ_j .

Computing weights: The output weights of an RBF network can be computed using a closed-form solution. The matrix of activation of the samples is defined from the training set (H) as follows:

$$\mathbf{H} = h(\|\mathbf{x}^\mu - \mathbf{c}_j\|_{\mathbf{R}_j}^2)_{\mu=1, \dots, M, j=1, \dots, C} \quad (3.11)$$

Based on Equation 3.2, the matrix of output weights (W), which estimates the matrix of labels (Y), is computed using the following equation:

$$\mathbf{Y} \approx \mathbf{H}\mathbf{W} \Rightarrow \mathbf{W} \approx \mathbf{H}^\dagger \mathbf{Y} \quad (3.12)$$

where \mathbf{H}^\dagger is the Moore–Penrose pseudo-inverse matrix [206] of \mathbf{H} and is computed as:

$$\mathbf{H}^\dagger = \lim_{\alpha \rightarrow 0^+} (\mathbf{H}^T \mathbf{H} + \alpha \mathbf{I})^{-1} \mathbf{H}^T \quad (3.13)$$

End-to-end optimization: After initializing the RBF weights and cluster centers with clustering algorithms such as k-means, it is possible to optimize the network end-to-end via backpropagation and gradient descent. Schwenker et al. computed the gradients of the loss function for a Gaussian activation function in [235].

3.4 Adapting RBFs for CNNs

This section presents the adaptation steps for using RBF classifiers for CNNs as depicted in Figure 3.1. The deep embeddings of the CNNs, computed using standard convolutional layers and inception blocks, are flattened and fed to RBFs after a fully connected layer in the architecture. The network ends in an output layer with softmax activation and is optimized end-to-end. Integrating the RBFs into deep structures and using them in conjunction with CNNs presents three challenges:

Initialization: Training the RBFs from scratch with randomly initialized weights using gradient descent is quite inefficient due to inconvenient initial cluster centers. The large initial distances in high dimensional spaces lead to small activation values, and the gradients attenuate considerably after the RBF hidden layer during backpropagation. Therefore, the k-means algorithm initializes the cluster centers before starting the training. Furthermore, computing the weights from Equation 3.12 is not feasible due to the scale of computer vision datasets such as ImageNet [58], which has over 14 million images and 1000 classes. Hence, using gradient descent and optimizing randomly initialized output layer weight is the optimal way to proceed.

Dynamic input features: The input features of classical RBFs are fixed, but this assumption is not valid with respect to CNNs. As the embeddings of CNNs develop during the training process, the cluster centers initialized by the k-means algorithm are no longer optimal after a few epochs of training. This research work proposes the optimization of the k-means algorithm's target with unsupervised loss during the training process as defined in Equation 3.10.

Activation: The nonlinear computational graph drawn by computing the distance in Equation 3.1 and applying the activations in equations 3.3-3.9 leads to inefficient gradient flow. Therefore, this research attempts to build a linear computational graph in RBFs through the introduction of a new activation function.

This section presents two modifications to classical RBFs to make them suitable for deep CNNs. First, it introduces an additional loss term to the RBFs' hidden layer. This term is based on the target function of the k-means algorithm defined in Equation 3.10 and continues in the unsupervised learning process during the development of the embeddings. The second contribution of this section is the introduction of a new quadratic kernel to build a linear computational graph for efficient optimization using backpropagation.

3.4.1 Introducing Unsupervised Learning Loss

The embeddings of CNNs change during the training process, which necessitates updating the cluster centers with an unsupervised loss. Therefore, introducing

an additional term to the RBFs' supervised loss function to optimize the cluster centers during training using the k-means unsupervised loss in Equation 3.10 is crucial for optimizing CNNs with an RBFs classifier (CNN-RBFs) end-to-end. The final loss of a CNN-RBF network is computed as follows:

$$\text{Loss}_{\text{rbf}} = \text{Loss}_{\text{supervised}} + \lambda \text{Loss}_{\text{unsupervised}} \quad (3.14)$$

where the classification loss $\text{Loss}_{\text{classification}}$ is any arbitrary loss function, for instance categorical cross-entropy.

It is conventional to use clustering algorithms such as the k-means or expectation-maximization (EM) algorithms to initialize the cluster centers. The loss function in Equation 3.14 is optimized using gradient descent by minimizing the distance of the embeddings for each sample from its nearest cluster center regardless of the class labels. The distance from the nearest cluster center is computed using the distance metric \mathbf{R}_j defined in Equation 3.1.

3.4.2 Quadratic Kernel

The kernels used for classical RBFs are nonlinear and increase the model's complexity. The architectures proposed in Figure 3.1 profit from using the state-of-the-art models for representation learning, i.e., CNNs, as a backbone. Therefore, CNN-RBF architectures can be trained with simpler linear models to improve the gradient flow during backpropagation. The proposed quadratic activation function is linear in the space of r^2 and is defined as follows:

$$h(r) = 1 - r^2/\sigma^2 \quad (3.15)$$

where σ is the trainable parameter that determines the width of the kernel. The proposed kernel is depicted in Figure 3.2 alongside the conventional activation functions. The proposed quadratic kernel reduces the nonlinearity of the CNN-RBF computational graph for backpropagation. The squares of the distances between cluster centers and samples are computed by linear matrix multiplication in Equation 3.1 and the application of the proposed linear activation for r^2 . Thus, the gradients of deep embeddings propagate backward through a distance computation with matrix multiplication and linear activations.

3.5 Experimental Results

This section presents the experimental results that reinforce the applicability of RBFs to CNNs on several standard computer vision benchmark datasets and

investigates the effect of tweaking various hyperparameters of the CNN-RBF architectures in the training phase and generalization to test data. Three convolutional backbones EfficientNet-B0 [264], InceptionV2 [261], and ResNet50 [96] compute the embedding of CNN-RBFs in this section. A list of the benchmark computer vision datasets used in this section is presented in Table 3.1.

Dataset	Train Size	Test Size	# Classes
CIFAR-10 [139]	50 000	10 000	10
CIFAR-100 [139]	50 000	10 000	100
Oxford-IIIT Pets [200]	3 680	3 369	37
Oxford Flowers [193]	1 020	6 140	102
FGVC Aircraft [173]	6 667	3 333	100
Caltech Birds [297]	5 996	5 794	200

Table 3.1: An overview of computer vision benchmark datasets used to evaluate the performance of CNN-RBFs (table adopted from [14]).

Figure 3.3 shows the hyperparameter search results for object classification on two benchmark computer vision datasets: CIFAR-10 and CIFAR-100. The backbone CNN model in these experiments is EfficientNet-B0, with a layer of RBFs for classification. The image preprocessing pipeline, called AutoAugment [50], consists of a set of optimal and automatically discovered augmentation policies for the ImageNet [58] dataset. The CNN-RBF architecture demonstrated in Figure 3.1 has two further hyperparameters: the number of cluster centers and the input dimensions of the RBF network. The models are optimized using an AdamW [162] optimizer with different learning rates and weight decay serving as tunable hyperparameters. The other hyperparameters are the loss constant (λ) from Equation 3.14, dropout rate, and batch size.

The hyperparameter searches in Figure 3.3 are conducted using the hyperband [150] algorithm with 4 agents running in parallel on two *Tesla V100* graphic processing units (GPUs) for approximately 10 days. It should be noted that dropout is only applied after the CNN backbone and before the fully connected layer in Figure 3.1. The output of the fully connected layer, without any activation function, is used as input features of the RBFs. The results in Figure 3.3 show that training CNN-RBF architectures leads to high performances, even with a wide range of hyperparameters. However, achieving good test performance with a high dropout rate and a large input dimension is challenging. CNN-RBF architectures show a better performance without dropout and rectified linear unit (ReLU) activations in the input layer of RBFs. Thus, dropout is neglected for further hyperparameter searches conducted on the other datasets in Table 3.1. The list of optimal hyperparameters for all datasets is presented in Table 3.2.

Several CNN-based backbone architectures are used within the CNN-RBFs to train models for all computer vision datasets. The experimental results of applying the CNN-RBFs to the computer vision benchmark datasets with the standard

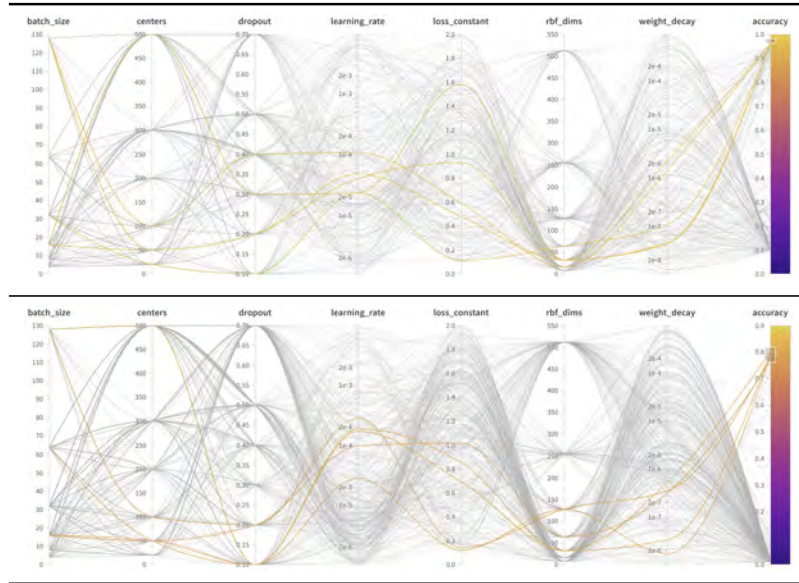


Figure 3.3: Hyperparameter search results from CIFAR-10 (top) and CIFAR-100 (bottom). The top five performing sets of hyperparameters for each dataset are highlighted in yellow (figures adapted from [14]).

Dataset	Loss constant	Learning rate	Embeddings dimensions	Batch size	Number of centers	Weight decay
CIFAR-10	0.1141	2.355e-5	64	32	20	1.090e-7
CIFAR-100	0.8557	1.873e-4	32	64	50	5.369e-7
Oxford-IIIT Pets	1.067	7.487e-5	64	16	50	1.150e-7
Oxford Flowers	1.562	1.076e-4	16	64	100	3.843e-6
FGVC Aircraft	0.5471	1.103e-4	8	8	50	1.222e-6
Caltech Birds	0.5156	2.603e-4	32	32	50	1.416e-8

Table 3.2: List of the final hyperparameters used for each computer vision benchmark dataset to achieve the performance of CNN-RBF architectures (table adapted from [14]).

train and test splits are presented in Table 3.3. CNN-RBFs show the capacity to learn the entire training dataset in all of the cases. There is, however, a small gap between the best-reported performances in computer vision literature and CNN-RBF architectures. Using dropout with CNN-RBFs for regularization does not lead to desirable results. Reducing the number of parameters of the RBFs while limiting their input size is the best empirically proven regularization strategy for RBFs besides data augmentation. Developing regularization methods for RBFs to improve generalization is an open research topic for reducing the gap between current performances and state-of-the-art computer vision models.

Dataset	Backbone	CNN-RBFs			Best result
		EfficientNet-B0	InceptionV2	ResNet50	
CIFAR-10	No-Augment	0.966	0.963	0.969	0.993
	Auto-Augment	0.975	0.977	0.942	
CIFAR-100	No-Augment	0.797	0.752	0.693	0.936
	Auto-Augment	0.822	0.805	0.778	
Oxford-IIIT Pets	No-Augment	0.840	0.804	0.622	0.967
	Auto-Augment	0.887	0.820	0.829	
Oxford Flowers	No-Augment	0.609	0.659	0.595	0.997
	Auto-Augment	0.828	0.757	0.667	
FGVC Aircraft	No-Augment	0.723	0.717	0.665	0.945
	Auto-Augment	0.842	0.843	0.828	
Caltech Birds	No-Augment	0.613	0.428	0.281	0.904
	Auto-Augment	0.618	0.587	0.503	

Table 3.3: Comparing the performance of various CNN-RBF architectures with pre-training and augmentation on benchmark computer vision datasets. The best results column is the top performance of the current state-of-the-art architecture on the benchmark dataset (table adapted from [14]).

3.6 Visualization of the RBF Classifiers

This section attempts to visually explain the training and test processes of vision models using RBF classifiers. First, it demonstrates the training process for the simple task of classifying hand-written digits. Then, the distribution and training samples of active clusters for a test sample are depicted for more complicated object detection tasks described in the previous section.

3.6.1 Visualization of the Training Process

This section visualizes the performance of the RBFs classifiers for CNNs on a simple dataset. The experiments are conducted on the modified national institute of standards and technology (MNIST) dataset [148], a dataset of hand-written digits including ten classes. Learning the dataset is considered a simple task in computer vision. The simplicity of the dataset and learning task allows for the visualization of the training process at a fine level of detail. CNN-RBF architecture has the same number of cluster centers as classes (ten) in the dataset to depict the training process in this experiment. The network architecture in this section consists of a four-layer CNN, and the output of these layers is connected to the RBF after a global average pooling layer and a fully connected layer.

Figure 3.4 demonstrates the evolution of the representations around the cluster center during the training process. The data samples in Figure 3.4 are placed according to their distance from the center and at a random angle. The samples

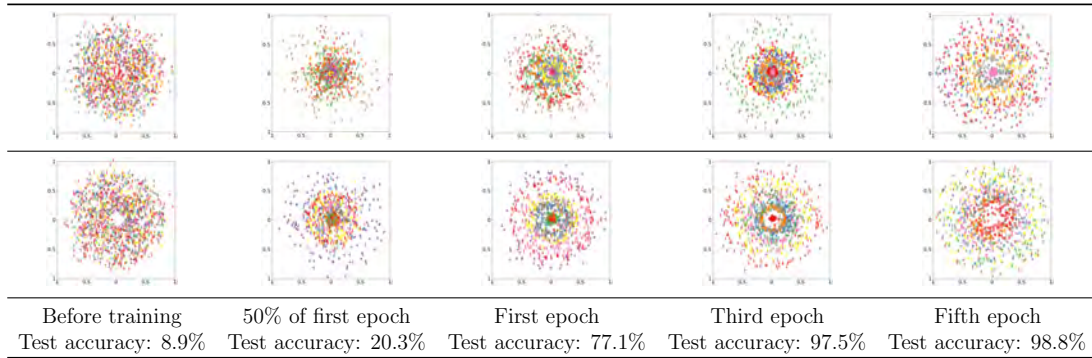


Figure 3.4: This figure presents the location of data samples compared to the cluster centers during the training process. The centers of the clusters are in the middle of the figures. The training samples are located at a random angle based on their distance from the center of the clusters. The vertical and horizontal axes show the normalized distances (figure adopted from [14]).

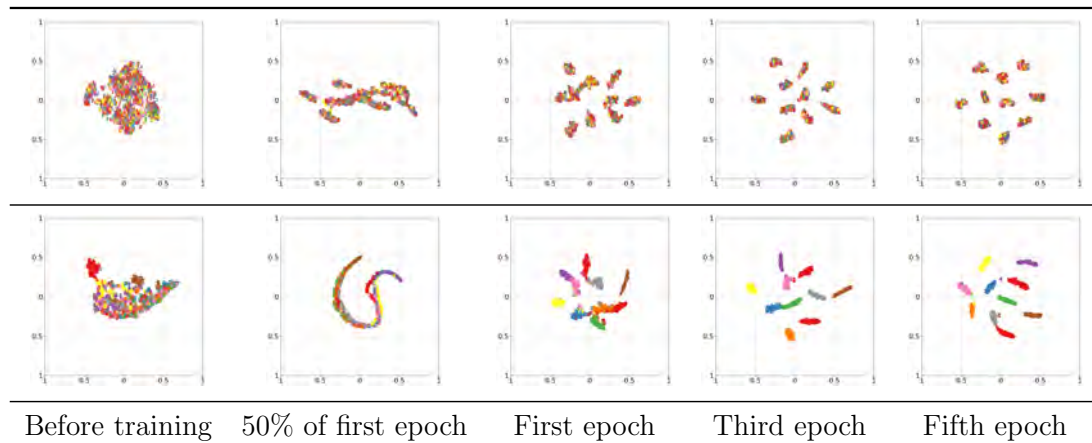


Figure 3.5: Two-dimensional representation of the training process: the figure presents the embeddings of the convolutional backbone (top row), and the activations of the RBFs (bottom row) mapped to a two-dimensional space using t-SNE [277]. The vertical and horizontal axes depict the normalized values; however, all sub-figures use the same normalization factors (figure adopted from [14]).

are shown with a number corresponding to their class, and the color is similar for samples of the same class in Figure 3.4. To reduce the overlap between close samples, a random uniform noise of amplitude 0.1 is added to the distance of the samples from the cluster centers.

Minimizing the unsupervised loss in Equation 3.10 reduces the distance of the data samples from the cluster centers. Furthermore, the supervised loss enforces the samples of the same class to maintain the same distance from cluster centers, as the activations are the only information for the network’s final decision. The

circles with samples of the same class around the cluster centers demonstrate the effect of supervised loss in training. Notably, the clusters presented in Figure 3.4 are selected to illustrate the concepts underlying training CNN-RBFs optimally.

Figure 3.5 illustrates the two-dimensional mapping of the CNN embeddings (top row) and RBF activations (bottom row) using t-SNE [277]. The effect of both supervised and unsupervised loss from Equation 3.14 is evident in this figure. The data samples split into clusters regardless of their class labels in the embedding space of CNN due to the unsupervised loss (top row in Figure 3.5). The activation values divide into clusters corresponding to the class labels, a process encouraged by the supervised loss.

3.6.2 Similarity Metric Learning and Interpretability

Using a different approximation strategy compared with fully connected layers provides CNN-RBFs with the chance to probe the decision-making process based on the following visual clues:

- Similar images as measured by the similarity distance metric of RBFs trained on CNN embeddings
- Clusters visualization with a higher contribution to the network's decision and distance of the samples from the centers of these clusters

The embeddings of the CNN are compared with the position of the cluster centers using the learned distance metric from the RBFs. The learned distance metric of the RBFs can measure the similarity between a test image and similar images from the training data. Figure 3.6 shows the similar images found in the training dataset for a given test sample as determined by the similarity distance metric in Equation 3.1. The most similar and dissimilar images are computed using the following criteria:

$$\mathbf{x}_{similar} = \arg \min_{\mathbf{x}_{train}^{\mu}} \|\mathbf{x}_{train}^{\mu} - \mathbf{x}_{test}\|_{\mathbf{R}}^2 \quad (3.16)$$

$$\mathbf{x}_{dissimilar} = \arg \max_{\mathbf{x}_{train}^{\mu}} \|\mathbf{x}_{train}^{\mu} - \mathbf{x}_{test}\|_{\mathbf{R}}^2 \quad (3.17)$$

where x_{test} presents the input of RBFs for a given test image, x_{train}^{μ} shows the input vector for training samples, and μ enumerates the training samples from 1 to N . $\mathbf{x}_{similar}$ and $\mathbf{x}_{dissimilar}$ represent the most similar and dissimilar images to the given test image (x_{test}) respectively, and \mathbf{R} denotes the positive definite covariance matrix similar to Equation 3.1. The same similarity metrics in Equation 3.16 and Equation 3.17 allow computing a ranked list of similar and dissimilar images for a given test sample.

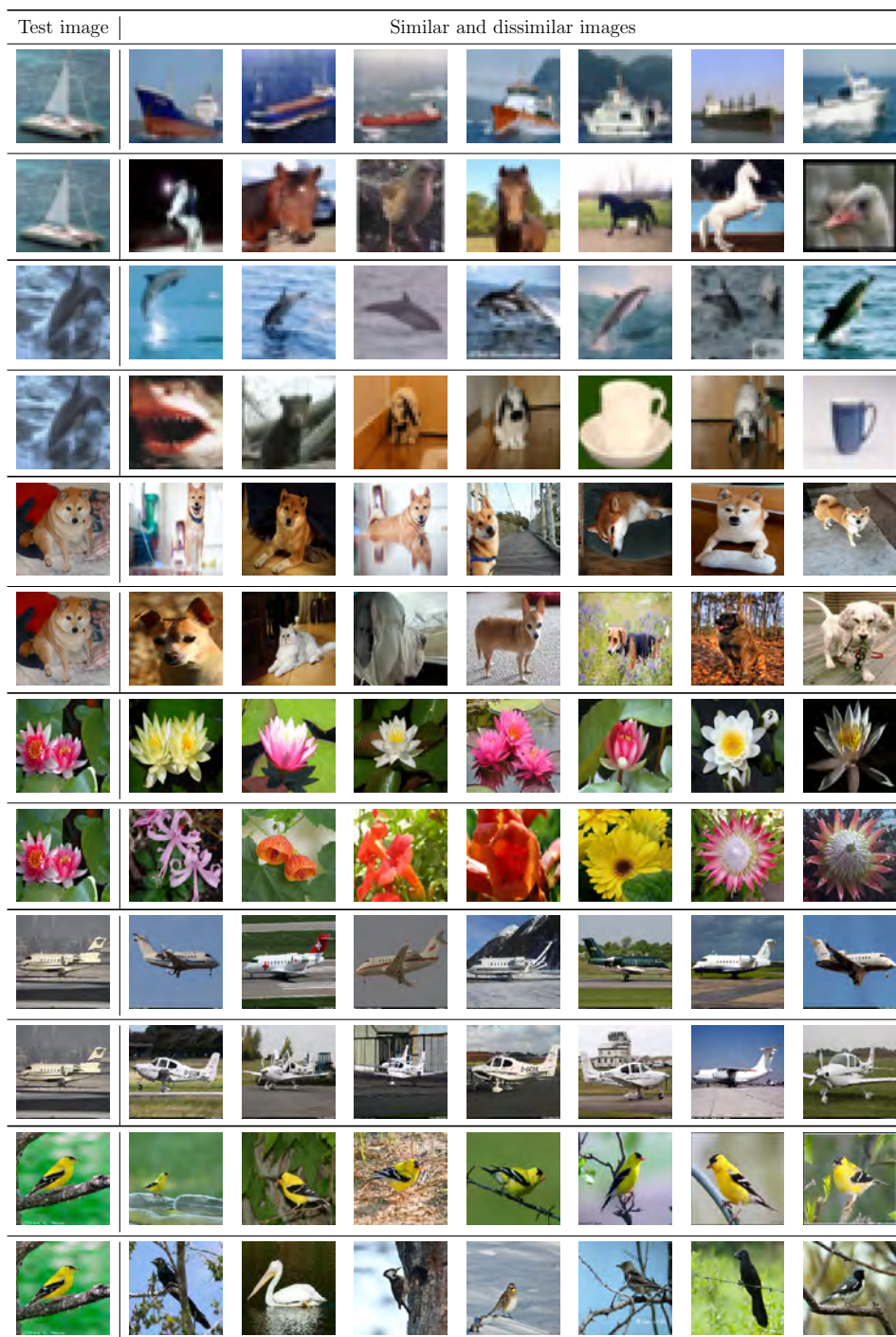


Figure 3.6: This figure depicts similar and dissimilar training images for given test images based on the similarity metric computed in Equation 3.1. The figure depicts the top 7 most similar and dissimilar training images for a given test image in every two rows. The images shown in every two consecutive rows belong to one of the datasets in Table 3.1 in the same order (figure adopted from [14]).

Figure 3.7 compares the performance of the similar sample selection for given test images. The figure suggests that the learned metric and Euclidean distances outperform the cosine distance for similar sample selection. Furthermore, the learned metric slightly outperforms the Euclidean distance in these specific cases.

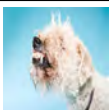
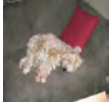
Test image	Similar images from training set in the embedding space													
														
														
														
														
														
														

Figure 3.7: The presented figure visualizes the top 14 images selected using different distance metrics in the embedding space for a given test image (figure adopted from [14]).

The active clusters for every sample provide the reasoning behind the final decision of CNN-RBFs. The clusters can be depicted using the distance of images from their centers. Figure 3.8 shows training samples and their distances from the cluster centers against a test sample. The product of activations and output weights determines the final decision of the RBFs. Thus, the importance of a cluster for a decision can be determined by sorting the product of activations and class weights. Figure 3.8 depicts the clusters with the highest contributions to the correct class (ground truth) and the wrong class based on this product. The wrong class here refers to the class with the second-highest confidence level.

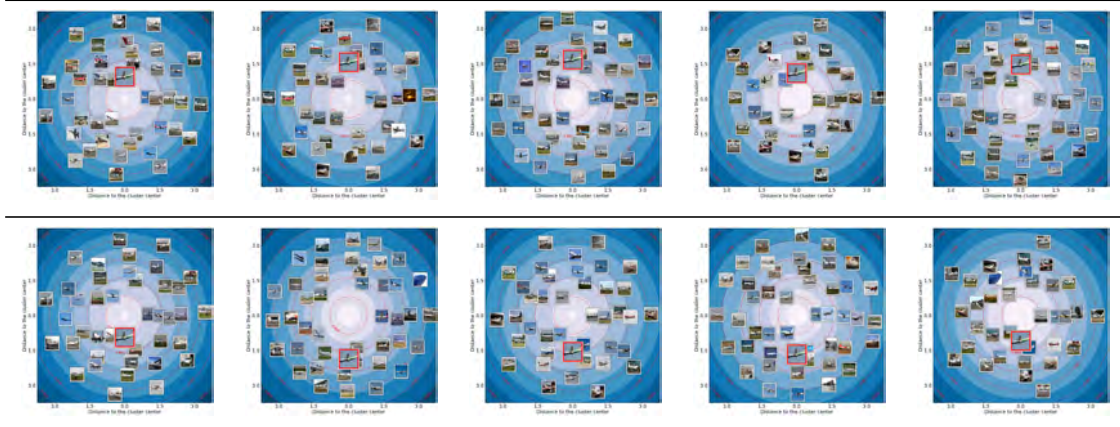


Figure 3.8: The Figure illustrates the clusters contributing to a CNN-RBF network’s correct class (top row) and the wrong class (bottom row). The larger image with red borders in each cluster representation is the test sample. Red circles show the distance of the samples to the cluster center, and the background is proportional to the activation values of the cluster. The brighter the activation value, the larger it is, and the maximum activation at the cluster center is equal to one (figure adopted from [14]).

3.7 Discussions and Conclusions

This chapter presents fundamental architectural modifications to RBFs for integrating them into CNN architectures for computer vision. The experimental results indicate that integrating RBFs classifiers into CNN architectures achieves competitive performances in benchmark computer vision datasets by combining supervised and unsupervised learning. The proposed activation and training process is compatible with any arbitrary state-of-the-art CNN architecture, including inception blocks and residual connections. The small gap between the CNN-RBFs performance and the best CNN models is a subject for future research to find optimal regularization methods for RBF networks. Using RBF architectures with CNNs introduces two unique and network-specific opportunities for learning a similarity distance metric and interpreting the decision-making process in more detail. The classification of similar and dissimilar images found using a similarity distance metric trained by RBFs is interpretable by humans. The cluster representations are currently only used to trace the decision-making process. In the current research, the distribution of images around clusters is not visually conclusive because they are optimized in an unsupervised manner.

4 Using Interpretability to Detect Adversarial Attacks for Robust CNNs

The existence of adversarial attacks on convolutional neural networks (CNN) questions the fitness of such models for sensitive and serious practical applications. The adversarial attacks are minimal changes computed for a given input image, provoking a misclassification even though both images appear identical to a human observer; they are, therefore, difficult to detect. In a different context, backpropagated activations of CNNs' hidden layers for a given input image, so-called feature responses, have been helpful for humans to visualize and understand what the CNN looks at while computing its output class. This chapter presents a novel method to detect adversarial examples and identify manipulated images by tracking adversarial perturbations in feature responses. This method is fully human-explainable and allows the automatic detection of adversarial attacks using the average local spatial entropy of feature maps for a given input image without altering the original network architecture. Experiments confirm the validity and functionality of our approach for detecting state-of-the-art attacks on large-scale models trained on ImageNet.

Alongside the contribution of this chapter to the robustness of computer vision models through the detection of adversarial attacks, this chapter presents one of the few applications of explainable artificial intelligence (XAI) for debugging models. Although feature response (maps) were initially intended to open the black box of vision models, they can also be used in practice for debugging, as presented in this chapter. This chapter demonstrates a successful application of vision model interpretability and XAI, which will hopefully inspire future work in the direction of debugging and designing models based on human-explainable methods. This chapter is adopted from the research published in [15], licensed under CC BY 4.0 ¹.

¹© 2018 Springer Nature Switzerland AG: <https://creativecommons.org/licenses/by/4.0>

4.1 Introduction

The success of deep neural nets for pattern recognition [230] has been one of the primary drivers behind the recent surge of interest in AI. A substantial part of this success is due to the application of convolutional neural networks (CNNs) [148, 43] and their descendants on image recognition tasks. Moreover, the respective methods have reached a level of sophistication where they are now being used in business and industry [251] and lead to a wide variety of deployed models for critical applications like automated driving [27] or biometrics [325].

However, concerns regarding the reliability of deep neural networks have been raised after the discovery of so-called adversarial examples or attacks [262]. These examples are specifically generated to fool a CNN into misclassifying visually very similar images or images that appear identical to the human eye with high confidence through the addition of barely visible perturbations [188] (see Figure 4.1). The perturbations are computed using an optimization process on the input: the network weights are fixed, and the input pixels are optimized for the dual target of (a) classifying the input differently than the ground truth class and (b) minimizing the changes to the input. A growing body of literature confirms the importance of this discovery on practical applications of neural networks [3]. The existence of adversarial attacks provokes questions on how CNNs achieve their performance and in what respect their decision-making differs from humans. In addition, adversarial attacks can threaten serious deployments of CNNs in the applications with the possibility of tailor-made attacks.

For instance, Su et al. [255] report on successfully attacking neural networks by modifying a single pixel. This attack works without having access to the internal structure or the gradients in the network under attack. Moosavi-Dezfooli et al. [187] further show the existence of universal adversarial perturbations that can be added to any image to fool a specific model. Furthermore, the impact of similar attacks extends beyond classification [184], attacks are transferable to other modalities [44], and also work on models distinct from neural networks [198]. Finally, adversarial attacks have been shown to work reliably even after perturbed images have been printed and captured again via a mobile phone camera [144]. Apparently, this research area touches on a weak spot concerning the robustness of CNNs in critical applications involving human privacy or security.

On the other hand, there is a recent interest in the explainability of AI agents, particularly using machine and deep learning models [280, 195]. It goes hand in hand with societal developments, like the new European legislation on data protection that affects any organization using algorithms on personal data [90]. While neural networks are publicly perceived as “black boxes” concerning how they arrive at their conclusions [1], several methods have been developed recently to deliver insight into the representation and decision surface of trained models,

improving interpretability. Prime candidates amongst these methods are feature response visualization approaches that provide information regarding operations in hidden layers of a CNN visible [316, 247, 194]. They can thus serve a human engineer as a diagnostic tool in support of reasoning over the success and failure of a model on the task at hand.

This chapter presents a method for using a specific form of CNN feature visualization, namely feature response maps, using guided backpropagation technique [247], to not only *trace* the effect of adversarial attacks but also to *detect* them. This method traces the attacks on algorithmic decisions throughout CNNs. Moreover, it uses feature response maps as input to a novel automated detection approach based on a statistical analysis of feature maps' average local spatial entropy. The goal is to decide if a model is currently under attack by the given input. The proposed approach has the advantage over existing methods because it does not change the network architecture, i.e., it does not affect the classification accuracy but is explainable to humans. Experiments on the validation set of ImageNet [223] with VGG19 networks [243] show the validity of our approach for detecting various state-of-the-art attacks.

The remainder of this chapter is organized as follows: Section 4.2 reviews related work in contrast to our approach. Section 4.3 presents the background on adversarial attacks and feature response estimation before introducing the proposed approach in detail in Section 4.4. Section 4.5 reports on the experimental evaluations, and Section 4.6 concludes the chapter with an outlook on future work.

4.2 Related Work

Work on adversarial examples for neural networks is a very active research field. Potential attacks and defenses are published at a high rate and have been surveyed by Akhtar and Mian [3]. Amongst potential defenses directly comparable to our approach are those that focus solely on detecting a possible attack and not on additionally recovering correct classification.

On the one hand, several detection approaches exist that exploit specific abnormal behavioral traces that adversarial examples leave while passing through a neural network. Liang et al. [153] consider the artificial perturbations as noise in the *input* and attempt to detect it by quantizing and smoothing image filters. This method used a similar concept, which is the basis of SqueezeNet introduced by Xu et al. [307], which compares the network's *output* on the raw and filtered input, and raises a flag if it detects a large difference between both. Feinman et al. [74] observe the network's output confidence as estimated by dropout in the forward pass [81], and Lu et al.'s SafetyNet [165] looks for abnormal patterns

in the ReLU activations of *higher layers*. In contrast, the method presented in this chapter performs detection based on statistics of activation patterns in the complete *representation learning* part of the network as observed in feature response maps, whereas Li [151] directly observes convolutional filter statistics there.

On the other hand, the second class of detection approaches trains sophisticated classifiers for directly sorting out adversarially optimized inputs: Meng and Chen’s MagNet [181] learns the manifold of friendly images, rejects far away ones as hostile and modifies close outliers to be attracted to the friendly images’ manifold before feeding them back to the network under attack. Grosse et al. [93] enhance the output of an attacked classifier by an additional class and retrain the model to classify adversarial examples as such directly. Metzen et al. [183] have a similar goal but target it via an additional subnetwork. In contrast, this chapter presents a method that uses a simple threshold-based detector and pushes all decision power to the human-explainable feature extraction via the feature response maps.

Finally, as shown in [3], different and mutually exclusive explanations for the existence of adversarial examples and the nature of neural network decision boundaries exist in the literature. Because our method enables a human investigator to trace attacks visually, it will be instrumental in taking this debate further.

4.3 Background

This section briefly presents adversarial attacks and the theory of feature response estimation before assembling both parts into the proposed detection approach in the next section.

4.3.1 Adversarial Attacks

The principal idea behind adversarial attacks is to find a small perturbation for a given image that changes the prediction of a CNN. Pioneering work demonstrated that negligible and visually insignificant perturbations could lead to considerable deviations in the networks’ output [262]. The optimization problem of finding a perturbation $\boldsymbol{\eta}$ for a normalized clean image $\mathbf{I} \in \mathbb{R}^m$, where m is the size (width×height) of the image, is stated as follows [262]:

$$\min_{\boldsymbol{\eta}} \|\boldsymbol{\eta}\|_2 \quad \text{s.t.} \quad \mathcal{C}(\mathbf{I} + \boldsymbol{\eta}) \neq \ell ; \quad \mathbf{I} + \boldsymbol{\eta} \in [0, 1]^m \quad (4.1)$$

where $\mathcal{C}(\cdot)$ presents the classifier, and ℓ is the ground truth label. Szegedy et al. [262] proposed a solution for the optimization problem of finding the perturba-

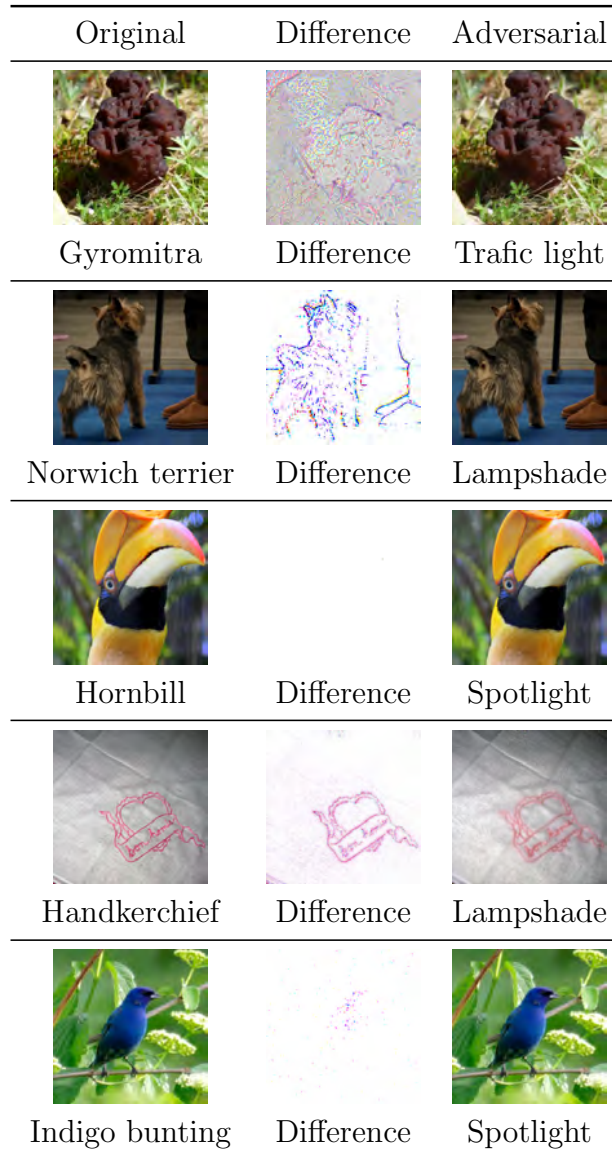


Figure 4.1: Examples of different state-of-the-art adversarial attacks on a VGG19 model: original images and labels (left), perturbations (middle), and mislabeled adversarial examples (right). In the middle column, zero difference is encoded white, and the maximum difference is black because of visual enhancement (figure adopted from [15]).

tions in Equation 4.1 for arbitrary labels ℓ' that differ from the ground truth. However, they used box-constrained limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) [78] to find perturbations satisfying Equation 4.1. Optimization based on the L-BFGS algorithm for finding adversarial attacks is computationally inefficient compared to gradient-based methods. Therefore, in this chapter, a few different gradient-based attacks, a one-pixel attack, and a boundary attack are

used to compute adversarial examples, as explained in the following paragraphs (see Figure 4.1).

Fast gradient sign method (FGSM) [89] is a method suggested for computing adversarial perturbations based on the gradient $\nabla_{\mathbf{I}}J(\boldsymbol{\theta}, \mathbf{I}, \ell)$ of the cost function with respect to the original image pixel values:

$$\boldsymbol{\eta} = \epsilon \operatorname{sign}(\nabla_{\mathbf{I}}J(\boldsymbol{\theta}, \mathbf{I}, \ell)) \quad (4.2)$$

where $\boldsymbol{\theta}$ represents the network parameters and ϵ is a constant factor that constrains the max-norm (l_{∞}) of the additive perturbation ($\boldsymbol{\eta}$). The ground truth label is presented by ℓ in Equation 4.2. The sign function is Equation 4.2 which computes the elementwise sign of the gradient of the loss function with respect to the input image. Optimizing the perturbation in Equation 4.2 in a single step is called the fast gradient sign method (FGSM) in the literature. This method is a white box attack, i.e. the algorithm for finding the adversarial example requires information on weights, gradients, and the network's architecture.

Gradient attack is a straightforward realization of finding adversarial perturbations in the FoolBox toolbox [211]. It optimizes pixel values of an original image to minimize the ground truth label confidence in a single step based on the gradient values instead of their sign proposed in the FGSM method.

One pixel attack [255] is a semi-black box approach to compute adversarial examples using an evolutionary algorithm [252]. The algorithm is not a white box since it does not need the gradient information of the classifier. However, it is not a full black box as it needs the class probabilities. The iterative algorithm starts with randomly initialized parent perturbations. The generated offspring compete with their parents at each iteration, and the winners advance to the next step. The algorithm stops when the ground truth label probability is lower than 5%.

DeepFool [188] is a white box iterative approach in which the closest direction to the decision boundary is computed in every step. It is equivalent to finding the corresponding path to the orthogonal projection of the data point onto the affine hyperplane, which separates the binary classes. The initial method for binary classifiers can be extended to a multi-class task by considering the task multiple one-versus-all binary classifications. After finding the optimal updates toward the decision boundary, the perturbation is added to the given image. The iterations continue estimating the optimal perturbation and apply it to the perturbed image from the last step until the network fails to predict the ground truth label.

Boundary attack is a black-box attack proposed by Brendel et al. in [30]. The algorithm starts with an adversarial image from another class compared with the target image and iteratively optimizes the distance between this image and the target image. It searches for an adversarial example with a minimum distance from the target image that keeps its original class throughout the optimization.

4.3.2 Feature Response Estimation

The technique of visualizing CNNs through feature responses is used to figure out which images’ region leads to the final prediction of a network. Therefore, computing feature responses enhances the explainability of the classifiers. This chapter demonstrates how to use this visualization tool to trace the effect of adversarial attacks on CNNs’ predictions and detect perturbed examples automatically.






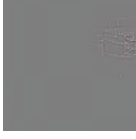

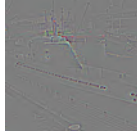



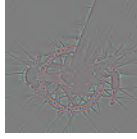
One pixel attack [255]: Predictions:				
	Eskimo dog	Feature response	Thimble	Feature response
FGSM [89]: Predictions:				
	Submarine	Feature response	Traffic light	Feature response
DeepFool [188]: Predictions:				
	Disc brake	Feature response	Dome	Feature response

Table 4.1: Effect of adversarial attacks on feature responses: (left) original images, and their feature responses, (right) perturbed versions, and their feature responses (figure adopted from [15]).

Erhan et al. [72] used backpropagation for visualizing feature responses of CNNs. They evaluate an arbitrary image in the forward pass and retain the activation values; then backpropagate from the last convolutional layer to the original image. As a result, the feature response maps have higher intensities in the regions that cause larger network activation values (see Figure 4.1). Moreover, the information on max-pooling layers in the forward pass can further improve the quality of these visualizations. Zeiler et al. [316] proposed computing “switches”, to consider the position of maximum values in all pooling layers, and then construct the feature response maps using transposed convolutional [68] layers.

Ultimately, Springenberg et al. [247] proposed a combination of both methods called “guided backpropagation”. In this approach, the information of “switches” (max-pooling spatial information) is kept, and the activations are propagated backward with the guidance of the “switch” information. This method leads to the best performance in the visualization of the inner workings of the network.

Therefore, guided backpropagation is used for computing feature response maps in this chapter.

4.4 Explainable Adversarial Attacks Detection

After reviewing the necessary background in the last section, this section presents this thesis's contribution to tracing adversarial examples in feature response maps, which inspired the novel approach to the automatic detection of adversarial perturbations in images. In this manner, visual representations of neural networks' inner workings also provide expert human guidance in developing CNNs that have increased reliability and explainability.

4.4.1 Tracing Adversarial Attacks

The research question followed in this chapter is whether explainability methods can provide insight into the reasons behind the misclassification of adversarial examples. The effect of adversarial attacks in the feature response maps of CNNs is traced in Figure 4.1. The general phenomenon observed in all images is that the feature response maps' active region for adversarial examples is widely spread. In contrast, Figure 4.1 demonstrates that the network looks at a smaller region of the image, i.e. is more focused, in the case of not manipulated samples.

The adversarial images are visually very similar to the original ones. However, they are not recognizable by deep CNNs. The original idea behind this study is that the focus of CNNs changes when facing an adversarial attack, leading to incorrect predictions. Conversely, the network makes the correct prediction once it focuses on the right region of the image. Visualizing the feature responses provides this and other exciting information regarding decision-making in computer vision models. For instance, the image of the submarine in Figure 4.1 can be considered a good candidate for an adversarial attack since the CNN is making the decision based on an object in the background (see the feature response maps of the original submarine in Figure 4.1).

4.4.2 Detecting Adversarial Attacks

Experiments tracing the effect of adversarial attacks on feature response maps thus suggest that a CNN classifier focuses on a broader region of the input if it is deliberately perturbed. Figure 4.1 demonstrates this connection for models' decision-making in the case of clean inputs compared with manipulated ones. The effect of adversarial manipulation is even more visible in the local spatial entropy of the grayscale feature response maps (see Figure 4.2). The feature response








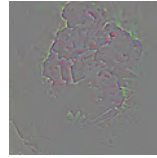
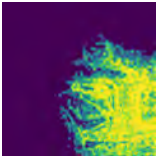
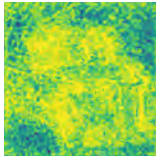
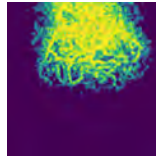
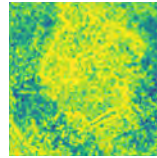
	Original	Adversarial	Original	Adversarial
Image:				
Feature response:				
Local spatial entropy:				

Table 4.2: Input, feature response maps, and local spatial entropy for clean and perturbed images, respectively (table adopted from [15]).

maps are initially converted to grayscale images, and local spatial entropies are computed based on the grayscale feature response maps as follows [37]:

$$S_k = - \sum_i \sum_j \mathbf{h}_k(i, j) \log_2(\mathbf{h}_k(i, j)) \quad (4.3)$$

where S_k is the local spatial entropy of a small part (patch) of the input image and \mathbf{h}_k represents the normalized 2D histogram value of the k^{th} patch. The indices i and j scan through the height and width of the image patches. The patch size is 3×3 , the same as the filter size of the first layer of the VGG19 model used. The local spatial entropies of corresponding feature responses are presented in Figure 4.2, and their difference for clean and adversarial examples suggests a likely chance of detecting perturbed images based on these maps.

Accordingly, the proposed method in this chapter uses the average local spatial entropy of an image as the final single measure to decide whether an attack has occurred or not. The average local spatial entropy \bar{S} is defined as:

$$\bar{S} = \frac{1}{K} \sum_k S_k \quad (4.4)$$

where K is the number of patches on the complete feature response maps and S_k shows the local spatial entropy as defined in Equation 4.3 and depicted in the last row of Figure 4.2. The proposed detector makes the final decision by comparing the average local spatial entropy from Equation 4.4 with a selected threshold to measure the spatial complexity of the feature response maps.

4.5 Experimental Results

The first experiments visually compare the approximated distribution of the averaged local spatial entropy of feature response maps in clean and perturbed images to evaluate the value of the explained metric in Equation 4.4. The validation set of ImageNet [223] with more than 50,000 images from 1,000 classes is the subject of this study, and feature response maps are computed for the VGG19 model [243]. Perturbations for this experiment are computed using several different methods, and the distribution of average local spatial entropies is depicted for the fast gradient sign attack (FGSM). Figure 4.2 shows that the clean images are distinguishable from perturbed examples, although there is some overlap between the distributions.

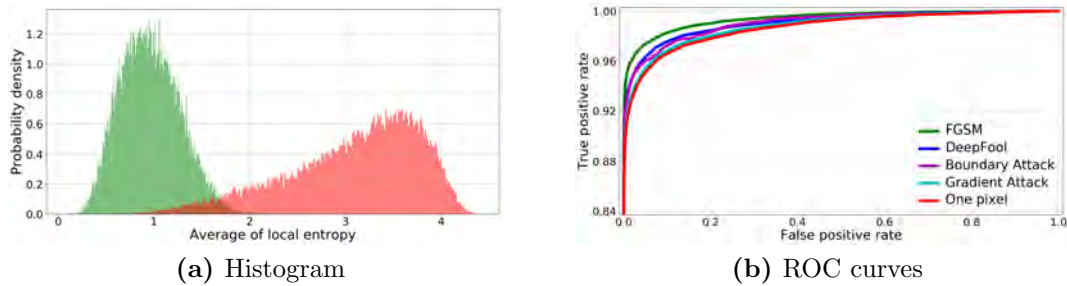


Figure 4.2: a) Distribution of average local spatial entropy in clean images (green) versus adversarial examples (red) as computed on the ImageNet validation set [223]. b) Receiver operating characteristic (ROC) curve of the performance of the detection algorithm on different attacks (figure adopted from [15]).

Computing adversarial perturbations using evolutionary and iterative algorithms is demanding in terms of time and computational resources. To apply the proposed detector to a wide range of adversarial attacks, many images are randomly drawn from the validation set of ImageNet for each attack to evaluate the detection performance of the presented method. The selection of images for each attack is made sequentially by class and filename and comprises only the number of images per method that can be computed within a reasonable amount of time, using a reasonable number of resources (see Table 4.3). The experiments are based on the FoolBox benchmarking implementation², and the attacks are computed using TitanX graphics processing unit (GPUs).

Figure 4.2b presents the receiver operating characteristics (ROC) of the proposed detector; numerical evaluations are provided in Table 4.3. Our detection method performs better for gradient-based perturbations compared to the single-pixel attack. Furthermore, Table 4.3 suggests that the best adversarial attack detection

²<https://github.com/bethgelab/foolbox>

performance is achieved for FGSM and boundary attack perturbations, where the network confidences on the ground truth labels are changed the most after manipulating the images. This observation suggests that the proposed detector is more sensitive to stronger attacks in fooling the network with a more drastic effect on target class confidence. Using feature response maps, the proposed method detects more than 91% of the perturbed images with a low false positive rate (1%).

Adversarial attack	#Images (run time [days])	Success rate	Ground truth confidence	Target class confidence	False positive rate		
					1%	5%	10%
FGSM [89]	50,014 (3)	0.925	0.022	0.588	0.954	0.974	0.983
Gradient attack [211]	50,014 (15)	0.499	0.052	0.371	0.922	0.954	0.969
One pixel attack [255]	50,014 (32)	0.620	0.037	0.463	0.917	0.951	0.966
DeepFool [188]	47,858 (42)	0.606	0.041	0.446	0.936	0.963	0.976
Boundary attack [30]	4,013 (17)	0.940	0.023	0.583	0.934	0.960	0.972

Table 4.3: The table describes the numerical evaluation of detection performance on different adversarial attacks. Column two gives the number of tested images and approximate elapsed run time. The success of an adversarial attack is defined if a perturbation changes the prediction. Columns four and five show average confidence values of the true (ground truth) and wrong (target) classes after a successful attack. Finally, the last columns show detection rates for different false positive rates (table adopted from [15]).

Method	Dataset	Network	Attack	Performance		
				Recall	Precision	AUC
Uncertainty density estimation [74]	SVHN [139]	LeNet [147]	FGSM	-	-	0.890
Adaptive noise reduction [153]	ImageNet (4 classes)	CaffeNet	DeepFool	0.956	0.911	-
Feature squeezing [307]	ImageNet-1000	VGG19	Several attacks	0.859	0.917	0.942
Statistical analysis [93]	MNIST	Self-designed	FGSM ($\epsilon = 0.3$)	0.999	0.940	-
Feature response (our approach)	ImageNet validation	VGG19	Several attacks	0.979	0.920	0.990

Table 4.4: This table describes the performance of similar state-of-the-art adversarial attack detection methods. The Area Under Curve (AUC) is the average value of all attacks in the third and last row (table adopted from [15]).

In general, it is difficult to directly compare different studies on attack detection since they use a wide variety of neural network models, datasets, attacks, and experimental setups. Table 4.4 presents a short overview of the performances of current detection approaches. The approach presented in this chapter is most similar to the methods of Liang et al. ([153]) and Xu et al. [307]. The detector proposed in this chapter outperforms both of the aforementioned methods based on the presented results in their work. However, it is not fully known if they used identical implementations and parameterizations of the attacks (e.g., a subset of images and learning rates for optimizing the perturbations). Similarly, adaptive noise reduction in the original publication [153] is only applied to four classes of the ImageNet dataset, and the method presented to detect adversarial attacks is

implemented based on CaffeNet deep learning framework, which differs from our experimental setup.

4.6 Discussion and Conclusion

The results presented in this chapter demonstrate the relevance and importance of adversarial attacks and the necessity to improve the robustness of CNNs against such perturbations. However, preliminary experiments on binary (cat versus dog [200]) and ternary (among three classes of cars [137]) classification tasks suggest that it is more challenging to find adversarial examples where marginal changes in the images are invisible to humans in such a setting. These tasks are proxies for the kind of few-class classification settings frequently arising in practice. Figure 4.3 illustrates these experimental results.



Figure 4.3: Successful adversarial examples created by DeepFool [188] for binary and ternary classification tasks are only possible with noticeably visible perturbations (figure adopted from [15]).

This chapter offers an approach to detect adversarial attacks based on human-explainable feature response maps. The proposed method traces the effect of adversarial perturbations on the networks' focus region in original images, which inspired a simple yet robust approach for detecting adversarial attacks automatically. The proposed method is based on thresholding the averaged local spatial entropy of the feature response maps and detecting at least 91% of state-of-the-art adversarial attacks with a low false positive rate on the validation set of ImageNet. However, the results are not directly comparable with state-of-the-art methods because of the diversity in the experimental setups and attack implementations.

Experimental results verify that feature response maps are informative in detecting specific failure cases in deep CNNs. Furthermore, the proposed detector uses the explainability of neural network decisions, an increasingly important topic for developing robust and reliable vision models. Future work, therefore, will concentrate on developing reliable and explainable image classification methods for practical use cases based on these preliminary results.

5 Motion Compensation in Computed Tomography Using CNNs

This chapter presents this thesis’s main applied contribution, motion artifact reduction, which enhances the quality of cone-beam computed tomography (CBCT) scans using 3D convolutional neural networks (3D-CNNs). This application is relevant and exciting for two main reasons: 1) There is no analytical solution to the problem of motion compensation since reconstruction algorithms are developed based on the assumption of measurements from a constant volume. 2) Motion artifacts are relevant in CBCT scans because of their long acquisition time, and using CBCTs demonstrates improving cancer therapy via adaptive dose calculation and patient posing.

This chapter offers a novel deep-learning (DL) based approach that significantly reduces motion artifacts and improves scan quality.

Because motion artifact reduction has no analytical solution, 3D deep convolutional neural networks (3D-CNNs) are employed as pre-and/or post-processing steps during CBCT reconstruction to target motion artifacts using a data-driven approach. The method described in the following paragraphs is performed either using the analytical Feldkamp-Davis-Kress (FDK) with filtered backprojection (FBP) reconstruction method or using the iterative algebraic reconstruction technique (iCBCT/ART). Based on refined UNet architectures, the neural networks are trained end-to-end via supervised learning. The dataset used in this chapter is generated from 4D computed tomography (CT) scans of lungs containing ten motion phases between inhalation and exhalation and patients’ breathing curves with negligible motion artifacts. The training ground truth volumes are the averaged volume over all phases in 4D-CT or the volume at the average amplitude of the breathing curve. The trained networks are validated using real-world CBCT scans and quantitative image quality metrics. In addition, a qualitative evaluation from clinical experts is performed.

The novel approach in this chapter can generalize to unseen data and yield notice-

able reductions in motion-induced artifacts and improvements in image quality compared with state-of-the-art CBCT reconstruction algorithms (up to 6.3dB and 0.19 improvements in PSNR and SSIM, respectively). The experimental findings from the simulation are confirmed by a clinical evaluation of real-world patients' scans (clinical experts reported at least a noticeable change in motion reduction over standard reconstruction in 74% of cases). To the best of our knowledge, this is the first time that inserting deep neural networks as pre- and post-processing plugins in the existing CBCT reconstruction pipeline and end-to-end training demonstrates significant improvement in imaging quality and reducing motion artifacts in CBCT scans. This chapter is mainly adopted from the research published in [12], licensed under CC BY-NC-ND 4.0 ¹.

5.1 Introduction

Computed tomography (CT) has become a versatile radiology and radiation therapy imaging technique. It obtains detailed 3D scans of the human body for diagnostics and planning therapies. Cone-beam CT (CBCT), in particular, is used for reconstructing scans from measurements of radiation therapy treatment devices (linear accelerators). CBCT reconstruction techniques in image-guided radiation therapy (IGRT) [113] and interventional radiology provide high spatial resolution in a cost-efficient manner [70]. IGRT treatment is performed in up to 40 fractions. For each fraction, it is necessary to obtain the image of the day in order to position the patient accurately. Novel applications of CBCT imaging in IGRT, such as online adaptive replanning [313] or daily treatment planning and dose calculation [114], are very well-known in the scientific community.

There are two leading families of reconstruction algorithms used in modern CT scanners: (i) analytical techniques and (ii) algebraic reconstruction. The first group is based on filtered backprojection (FBP), most prominently represented by the Feldkamp-Davis-Kress (FDK) method [75]. The second group consists of the algebraic reconstruction techniques (ART) family, which is based on reformulating and solving the reconstruction problem through an iterative optimization technique. Although the development of algebraic methods started in the late 1960s [103], they have only been deployed on CT scanners in the last 15 years [91] mainly because of their high computational cost. In recent years, this problem has disappeared due to the high availability and relatively low cost of GPUs.

Iterative CBCT (iCBCT) reconstruction algorithms based on the ART family introduced in [202] for Varian's Halcyon and TrueBeam addressed the need for superior image quality in terms of better noise suppression and improved contrast as compared with FDK and demonstrated in [82, 133, 174, 294]. However, there

¹© 2023 The Authors: <https://creativecommons.org/licenses/by-nc-nd/4.0>

is still room for improvement in these methods, if they are to tackle real-world artifacts which are not a part of theoretical and analytical solutions.

Imaging artifacts [232] are still a prevalent feature in CBCT reconstruction. The main sources of artifacts are (i) electrical and photon count noise, (ii) photons from scattered X-rays, (iii) extinction and beam hardening effects (e.g., due to metal implants), (iv) approximations in the reconstruction (due to finite beam width and detector pixel size), (v) aliasing (due to finite pixel size and cone beam divergence), (vi) ring artifacts (due to defect or miscalibrated detector elements), and (vii) patient motion.

Motion artifacts arise since the reconstruction methods assume that the scanned patient is stationary. However, periodic respiratory and cardiac motions and non-periodic motions, such as gas bubbles in the abdomen caused by the digestive system, lead to acquiring projections from different states of motion in the body. Patients' motion leads to the appearance of evident and undesirable, typically streak-shaped artifacts in reconstructed scans.

The following motion compensation strategies are used so far in IGRT clinical routine: (i) 4D or gated CBCT based on an external breathing signal [61], (ii) breath hold CBCT based on an external breathing signal and potential patient feedback, (iii) assisted breathing based on a ventilator system [205], (iv) abdominal compression devices applied to the patient [51], (v) internal breathing signal extraction [7].

This chapter presents a novel approach to mitigate motion artifacts in CBCT reconstruction using deep learning (DL). The CBCT reconstruction is integrated within a DL pipeline, where convolutional neural networks are employed as pre- and/or post-processing steps. These networks act on either the 2D X-ray projections (preprocessing), the reconstructed 3D volume (post-processing), or both. Next, the models are trained end-to-end in a supervised fashion using CBCT scans containing simulated motion artifacts and motion-free scans as ground truth. Finally, the trained models are validated quantitatively using various scan quality-related numerical metrics, and on an independent real-world patient CBCT dataset developed through qualitative clinical feedback.

5.2 Related Work

Much work has been done [232, 25, 86] regarding the characterization and mitigation of the various kinds of artifacts that negatively impact image quality in CT and CBCT scans. In recent years, DL-based approaches have shown promising results, including applications for IGRT and adaptive radiation therapy [203].

Würfl et al. [303] mapped the components of the FBP algorithm into a neural network by introducing a novel DL-enabled cone-beam back-projection layer. A forward projection operation efficiently computes the backward pass of the back-projection layer. This approach thus permits joint optimization and correction in the projection and volume domain. Moreover, Maier et al. [171] argued that implementing prior knowledge (such as the back-projection operation) in the form of (differentiable) known operators into DL algorithms reduces training error bounds while reducing the number of free parameters.

Limited-angle CT is employed to reduce acquisition time and decrease the radiation dose, which leads to a degradation of image quality and introduces sparseness artifacts. To overcome these issues, Wang et al. recently presented an encoder-decoder architecture based on the UNet model [219] to reconstruct high-quality scans with fewer projections. A UNet processes scans reconstructed by the simultaneous algebraic reconstruction technique (SART) to improve the imaging quality [18]. Experiments on chest and abdomen CT scans demonstrated the superiority of the proposed methods over existing approaches. Similarly, Schnurr et al. proposed UNet-based networks to correct limited-angle artifacts in circular tomosynthesis scans [231].

DL-based approaches demonstrate success in metal artifact reduction (MAR) [199, 319]. Lin et al. introduced a dual-domain architecture (DuDoNet) to jointly compensate for metal-induced artifacts in both projection and volume domains [157]. Experimental results on the DeepLesion CT dataset [309] showed that the proposed method outperformed both traditional and other DL-based approaches. An improved model, DuDoNet++, was proposed to compensate for over-smoothed and distorted image reconstruction and led to improved artifact correction, especially for large metallic objects [168]. Furthermore, there have been recent efforts in MAR using unsupervised approaches, as explained in [154]. The U-DuDoNet model [167] directly models the artifact generation and compensation process in both the projection and volume domains. More recently, the interactive [286] and interpretable [292] versions of DuDoNet have been introduced to improve the interpretability and enhance the interaction between projection and volume domains.

DL-based approaches have been employed to improve sparseness artifacts generated by low-dose CT reconstruction [95, 115, 321, 136]. Chen et al. present the AirNet model to fuse analytical and iterative CT reconstruction and integrate

them into DL to improve sparse-data 3D and 4D CBCT reconstruction [38, 39]. In the projection domain, DL-based models can also correct signal degradation caused by X-ray photons scattering within the patients' body [172, 71].

Finally, motion artifact compensation using DL has received comparatively less interest. Paysan et al. present an initial study of a UNet-based artifact reduction method, but only in the volume domain [204]. Su et al. used UNet architectures to reduce simulated motion artifacts in head CT scans based on simple simulated rigid (translations, rotations, oscillations) transformations. Finally, Lyu et al. used recurrent neural networks for cardiac motion artifact reduction in magnetic resonance imaging (MRI) [166].

5.3 Materials and Methods

This section presents the preliminary knowledge and the related theory which lays the foundation of this chapter's main contributions and findings. First, this section starts with a more detailed explanation of the CBCT reconstruction techniques used in this chapter. Secondly, the motion simulation framework is explained, followed by a discussion on the simulated and real-world datasets and the clarification of their differences.

5.3.1 CBCT Reconstruction

Both analytical and iterative methods are considered for the reconstruction of 3D CBCT volumes from 2D cone-beam projections in this research work. *Feldkamp-Davis-Kress* [75] (FDK) is an analytical reconstruction method based on filtered back-projection (FBP). Although the *Tuy* data-sufficiency conditions [273] are not met for circular trajectories of a cone-beam source, FDK provides a fast and reliable analytical approximation of the inverse Radon transform, and it has become a golden standard for 3D CBCT reconstruction [34]. *Ram-Lak* filter compensates for the radial non-uniformity of the sampling density in FDK. Moreover, *half-fan weighing* is necessary to avoid data duplication for the datasets acquired with half-fan geometry. The projections are acquired from the full 360° trajectory. However, the detector is shifted against the gantry to one direction to increase the field of view in half-fan geometry. Half-fan weighing is followed by *cosine weighting* to decrease the longitudinal full-off effect due to the cone-beam geometry. Finally, the projections are down-sampled so that their resolution matches the cut-off frequency requirement given by the target resolution of the reconstructed volume.

Besides FDK, in this chapter, the iterative *algebraic reconstruction technique* (ART) originally based on Kaczmarz algorithm [119] is used for iterative CBCT

(iCBCT) reconstruction. This method approximates the volume \mathbf{f} by an iterative optimization of the data-fidelity cost function: $\|\mathbf{A}\mathbf{f} - \mathbf{p}\|^2$, where \mathbf{A} and \mathbf{p} represent the forward-projection operator and projections in the attenuation space, respectively. In each iteration k , an update of the actual volume estimation is computed through back-projecting the gradient of the cost function, i.e. $\sum_{\alpha} \mathbf{A}^{\top}([\mathbf{A}\mathbf{f}_k]_{\alpha} - \mathbf{p}_{\alpha})$ where \mathbf{p}_{α} and $[\mathbf{A}\mathbf{f}_k]_{\alpha}$ denote the projection under angle α and corresponding forward-projection of actual estimated volume \mathbf{f}_k , respectively, and \mathbf{A}^{\top} represents the back-projection operator. One of the advantages of the iterative methods is allowing a straightforward integration of prior knowledge into the reconstruction process through a regularization term to augment the cost function during optimization. The implementation used in this chapter employs the edge-preserving *total variation* regularization, which helps to reduce the noise and cone-beam artifacts in the areas far from the iso-center.

In order to significantly reduce the computational cost, the GPU implementation of ART is further accelerated through the following approaches: First, the version of ART known as *simultaneous ART* (SART) is used where the volume is updated in parallel for each input projection. Next, *ordered subsets* (OS) [132] and Nesterov *momentum method* [190] are employed. Finally, a destination-driven approach [125] is employed in the forward projection of ART and backward projection of both ART and FDK. Further details about the TV-regularized OS-SART with momentum can be found in [207], where the method is presented as a part of the iCBCT algorithm deployed clinically in Varian products.

5.3.2 Motion Simulation

It is necessary to simulate motion for volumes with available ground truth to train the models using supervised learning; hence, the motion simulation method aims to generate synthetic datasets of CBCT volumes with motion artifacts. The motion simulation method starts from the phase-gated 4D CT scans described in Section 5.3.3 and a set of recorded breathing curves. The method profits from the Deeds [97] algorithm to perform a deformable registration between CT volumes of the end-inhale and end-exhale phases and to create a patient-specific deformation vector field (DVF).

A reconstructed CT scan, its DVF, and the patients' breathing curve are sufficient requirements to simulate the motion during the CBCT acquisition. The simulation method deforms the CT volumes by interpolating the DVFs according to the breathing amplitude. It creates a forward projection at each angular step by matching its timestamp with the relevant amplitude in the breathing curve. Each projection acquired through the described motion simulation method corresponds to a different respiratory state. Then, the CBCT volume is reconstructed using either the FDK or iCBCT reconstruction algorithms. Motion artifacts are

evident in the volumes reconstructed from the explained motion-simulated CBCT projection acquisition technique. Figure 5.1 shows an example of typical motion artifacts created by patient motion in real-world (clinical) CBCT dataset (test dataset, see Section 5.3.3) side-by-side with the emulated motion artifacts from our motion simulation.

The supervised learning approach presented in this chapter requires ground truth volumes without motion artifacts. The ground-truth volumes correspond either to a fixed motion state (average amplitude) or the average of all deformed volumes (average volume). Moreover, data augmentation is a crucial component of supervised learning pipelines to enhance optimization performance. Augmentation is realized by using different breathing curves to simulate motion with the same dataset of CT scan and DVF in this research work.

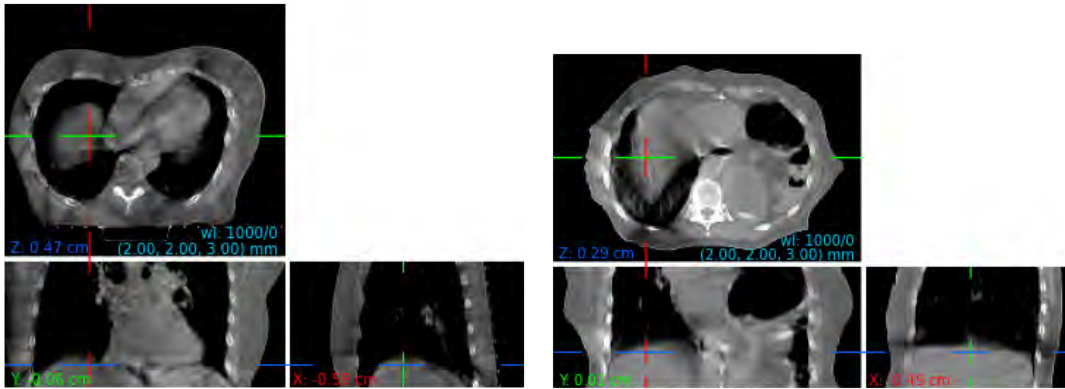


Figure 5.1: Motion Artifacts. Left: CBCT scans with motion artifacts from the test dataset. Right: Scan with artificially produced motion artifacts from the motion simulation. The scans are presented in HU with W/L=1000/0 (figure adopted from [12]).

5.3.3 Datasets

For the training and validation of the different methods, a dataset of thoracic 4D CT scans from 80 patients is split into fractions of 60%, 20%, and 20% for training, validation, and testing, respectively. They were provided as input to the motion simulation described in Section 5.3.2. The patient-specific anatomical correct deformations were extracted from the end inhale and exhale out of the ten breathing phases. To simulate plausible motion patterns during a virtual CBCT acquisition and to augment the training dataset, 400 recorded breathing curves were obtained via Varian’s real-time position management (RPM) system.

For testing the developed methods on real-world (clinical) patient CBCT scans, a dataset of 77 Halcyon scans was employed. All pre-processed projection data and reconstructed volumes were given at the same size, resolution, and geometry to ensure consistency: The projection size is 320x80 pixels (with a resolution of

1.344 × 4.032 mm), and the volume size is 256 × 256 × 48 voxels (2 × 2 × 3 mm). The source-to-imager distance is 154 cm with a detector offset of 17.5 cm.

5.4 Supervised Learning for Motion Artifact Reduction

This section presents the necessary underlying basics to start with supervised learning. First, the DL-enabled framework for reconstruction and refinement models, including UNets in the projection- and volume domains, is explained. Second, evaluation metrics for numerical analysis of simulated data are presented, and the section concludes by describing the experimental setup, including the hardware.

5.4.1 DL-Enabled CBCT Reconstruction

Motion leads to inconsistencies in the acquired projections, which appear as artifacts in the volume domain after reconstruction. Therefore, motion corrections can be, in principle, applied before and/or after reconstruction. The models, estimating these correction steps, are implemented as trainable neural network architectures derived from 3D refined UNet architectures.

The reconstruction algorithm used is either FDK or iterative CBCT (iCBCT) reconstruction, as discussed in Section 5.3.1. These algorithms are implemented based on forward- and back-projection layers implemented with custom compute unified device architecture (CUDA) code and interfaced as PyTorch modules. The analytical solution using filtered back-projection, inspired by FDK, is differentiable. Therefore, it is possible to back-propagate the gradient through this module and simultaneously optimize in both projection and volume domains, called dual-domain optimization. Dual-domain optimization requires a differentiable reconstruction method such as FDK and is not practical for iterative techniques.

The supervised learning approach uses the simulated motion dataset (Section 5.3.2) for training the motion compensation networks, where the loss is calculated in the volume domain. The ground truth is either calculated as the motion-averaged volume (“average volume”) or given as the volume corresponding to the fixed motion state matching the average breathing signal amplitude (“average amplitude”). The networks are validated on the validation and test portions of the simulated motion dataset and an independent real-world test dataset containing real-world clinical patient CBCT scans (see Section 5.3.3). In detail, the reconstruction pipeline consists of the following components:

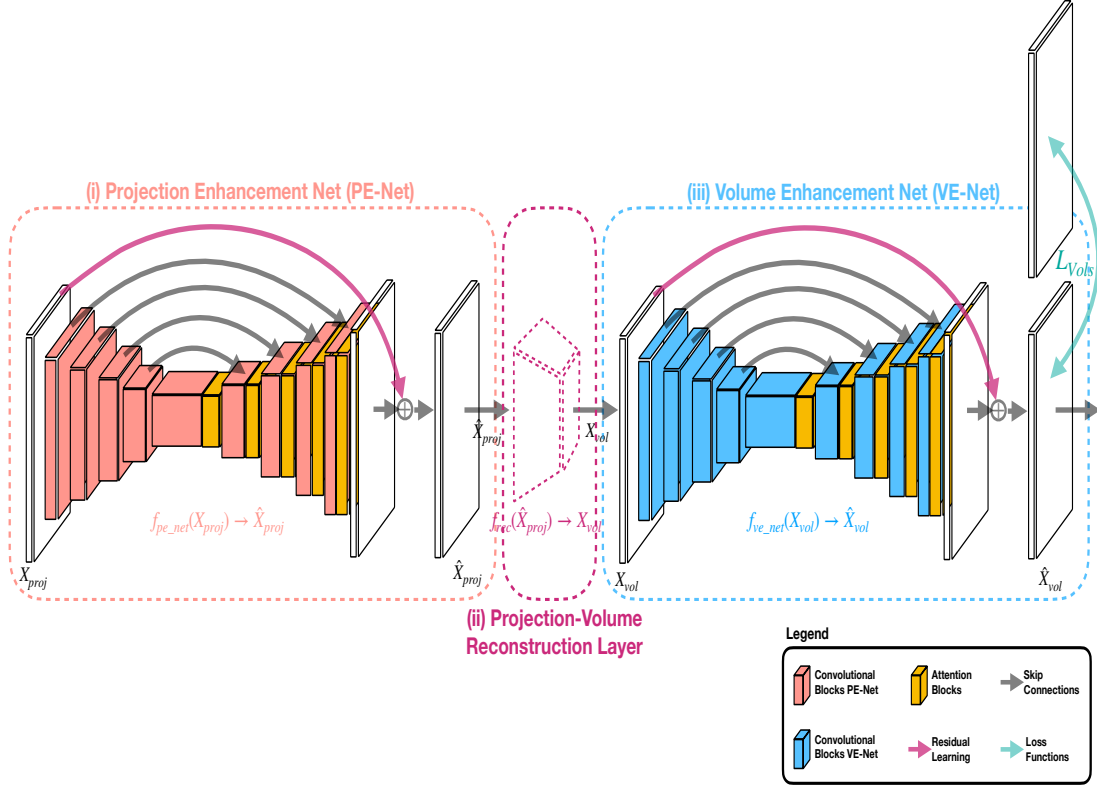


Figure 5.2: The architecture of the proposed dual-domain model for end-to-end optimization consists of the following components: (i) a projection enhancement network (PE-Net), (ii) a projection-to-volume reconstruction layer, and (iii) a volume enhancement network (VE-Net) (figure adopted from [12]).

Projection Enhancement Network (PE-Net): A convolutional neural network based on UNet architectures, explained in more detail in the next section, is deployed to mitigate motion-induced artifacts in the projection space. PE-Net receives as input the acquired projections $\{\mathcal{X}_{proj} \in \mathcal{R}^{H_p \times W_p \times C_p}\}$, and enhances these projections $\{\hat{\mathcal{X}}_{proj}\}$, i.e. $\hat{\mathcal{X}}_{proj} \approx f_{pe_net}(\mathcal{X}_{proj})$ to remove motion artifacts in the projection space. Here, $H_p \times W_p \times C_p$ denotes the projection dimensions in terms of height, width, and number of projections.

Projection-to-Volume Reconstruction Layer: The projection-to-volume reconstruction layer $f_{rec}(\cdot)$ receives as input the (enhanced) projections $\{\hat{\mathcal{X}}_{proj}\}$ and outputs a reconstructed volume $\{\mathcal{X}_{vol} \in \mathcal{R}^{H_v \times W_v \times C_v}\}$, i.e. $f_{rec}(\hat{\mathcal{X}}_{proj}) \rightarrow \mathcal{X}_{vol} : \mathcal{R}^{H_p \times W_p \times C_p} \rightarrow \mathcal{R}^{H_v \times W_v \times C_v}$, where $H_v \times W_v \times C_v$ represents the volume's height, width, and number of slices. This layer corresponds to the regular FDK or iCBCT reconstruction (Section 5.3.1).

Volume Enhancement Network (VE-Net): The VE-Net $f_{ve_net}(\cdot)$ is respon-

sible for enhancing the reconstructed volume and compensating motion artifacts in the volume domain. As output, the VE-Net produces an enhanced volume $\{\hat{\mathcal{X}}_{vol} \in \mathcal{R}^{H_v \times W_v \times C_v}\}$, i.e. $\hat{\mathcal{X}}_{vol} \approx f_{ve.net}(\mathcal{X}_{vol})$.

Our proposed dual-domain (end-to-end) model, shown in Figure 5.2, combines the above components for motion correction in both projection- and volume domains. It consists of three different modules: (i) a projection enhancement network (PE-Net), a (ii) projection-to-volume reconstruction layer, and a (iii) volume enhancement network (VE-Net).

The following paragraphs describe the different model blocks of the proposed architecture. Note that these blocks are used in both the projection enhancement (PE-Net) and volume enhancement (VE-Net) networks.

Encoder Blocks: The encoder blocks of the presented architecture in Figure 5.2 consist of four similar submodules including 3D a convolutional layer with filters of size $3 \times 3 \times 3$, followed by an instance normalization [274], the Swish activation function [210] and a 3D max-pooling layer of size $2 \times 2 \times 2$. The number of convolutional filters in the first block is doubled for every next layer. Hence, the latent representations of the input volume have a larger number of channels but a smaller spatial size with a higher receptive field after the first layer.

Decoder Blocks: The decoder block aims at computing the motion corrections from latent representations and has four submodules starting with a trilinear upsampling followed by 3D convolutions of size $3 \times 3 \times 3$, instance normalization, and the Swish activation function. The number of convolutional filters is halved after each layer to make the entire model’s architecture symmetric.

Attention mechanisms: To further compensate for motion artifacts, the models rely optionally on attention mechanisms. More precisely, as part of the bottleneck and decoder blocks of both PE-Net VE-Net, there are channel-wise and spatial-wise attention layers [302] in 3D. The corresponding input feature maps are multiplied at each decoder layer with the generated attention maps to refine the original features. The model can focus on more relevant features using these attention layers. Models including attention layers, are denoted by “Attn.” in Table 5.1.

Residual Learning and ResUNet: Using residual learning is crucial to simplifying the learning task and improving the convergence speed. The architecture depicted in Figure 5.2 uses two components to enhance the gradient flow and simplify the learning task. First, the proposed architecture profits from a direct residual connection from input to output (“residual learning”) to optimize the required correction instead of reconstructing the ground truth. The proposed architecture optionally includes internal residual connections between the input and output of the convolutional layers to improve the gradient flow as described in [320]. Networks including “ResUNet” layers are labeled as such in Table 5.1.

5.4.2 Evaluation Metrics

The experimental results of motion compensation on the simulated dataset are reported based on numerical performances using several quantitative metrics [293] sensitive to the similarity of pairs of volumes (x, x') . The evaluation metrics are computed by summing up the differences over all components, voxels in volumes, as follows:

- root mean squared error: $\text{RMSE} = \sqrt{\text{MSE}}$
where $\text{MSE}(x, x') = \frac{1}{N} \sum_i \|x_i - x'_i\|^2$
- peak signal-to-noise ratio: $\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right)$
- structural similarity index (SSIM)[293]

In addition, Table 5.1 reports the mean and standard deviation of the difference image $x - x'$ used for reducing (correcting) the motion artifacts. All metrics are calculated in Hounsfield units (HU) from pairs of uncorrected or corrected body-masked volumes and their corresponding ground truth counterparts.

5.4.3 Experimental Details

This section describes the experimental setup, architectural variants, optimization settings, and implementation details used for training based on a motion-simulated dataset.

Experimental Setup: The volume size is $256 \times 256 \times 48$ voxels based on the neural network architectures to optimize computational and memory allocation costs. Based on the training dataset discussed in Section 5.3.3, 720 projections are used per scan for training, and motion artifacts in CBCT scans are computed using motion simulation introduced in Section 5.3.2. The reconstruction and forward-projection geometry are selected to match the real-world test dataset, which is used for clinical evaluation in Section 5.5.

Data Augmentation: Five different patient breathing curves per CT scan are added for motion simulation from each original CT scan in the training dataset. Data augmentation through various breathing curves led to a considerable boost in the final performance of our motion correction models.

Model Architecture: The baseline model, initially considered for motion correction, is a UNet with residual learning from input to output as depicted in Figure 5.2. A plain UNet [219] architecture without residual connections is already sufficient for correcting the artifacts in volume space; however, residual learning is necessary for the more complicated tasks, including correcting the

projections and dual-domain optimization. Therefore, all of our models include residual learning. The baseline UNet model has a depth of 4 and has 32 filters in the first layer. The number of filters doubles after every layer up to the middle (model’s bottleneck), and the architecture is reverted afterward. The same architecture is used for both PE-Net and VE-Net. For dual-domain optimization, two such models form the architecture together. For PE-Net, the models process the projections in chunks of 192 due to memory limitations. PE-Net and VE-Net are extended with internal residual connections (“ResUnet”) and/or channel-spatial attention (“Attn.”) for different experiments presented in the next section.

Implementation and Optimization Settings: The models used for motion compensation are implemented using the PyTorch [201] framework. The experiments were performed on NVIDIA V100 (A100) GPUs with 32 (40) GB of VRAM. Both projections and volumes are normalized using constant coefficients per dataset to the approximate range of zero and one before optimization. The loss function for optimizing the models is the difference between the predicted and reconstructed volume as computed by the $\ell_1 - norm = \sum_i \|x_i - x'_i\|$. The AdamW [162] optimizer is used with a learning rate of $1.41 \cdot 10^{-6}$ and weight decay of $1.87 \cdot 10^{-8}$ in the projection domain, and a learning rate of $1.11 \cdot 10^{-4}$ and weight decay of $1.39 \cdot 10^{-8}$ in the volume domain. These parameters are the results of a joint hyperparameter sweep with other parameters, such as a number of convolutional filters, kernel size, and convolutional dilation. This experiment’s batch size is 1 (due to GPU memory limitations), and training continues for a total number of 300 epochs. The model that reduces the validation loss the most during the training is selected for testing.

5.5 Experimental Results

This section presents the experimental quantitative and qualitative results obtained by applying DL-based motion reduction techniques using 3D convolutional neural networks. First, the quantitative results obtained with the test portion of the simulated dataset are presented. Second, the qualitative results based on a clinical evaluation of the real-world test dataset are discussed.

5.5.1 Quantitative Results

In order to train the model architectures (see Figure 5.2) in a supervised scenario, only the simulated motion dataset (Section 5.3.3) is relevant. The training set is used for training the models, while the validation set results guide the optimization to select the best models and hyperparameters. Experimental results in this section consist of the final performance on the left-out test set during the training

Model Architecture	RMSE ↓	PSNR (dB) ↑	SSIM ↑	Mean±stdev
Baseline Performance of Average Volume				
FDK	77.8875	28.3802	0.8086	-
iCBCT	76.2560	28.6741	0.8701	-
Baseline Performance of Average Amplitude				
FDK	86.9695	27.5059	0.7992	-
iCBCT	106.5914	25.6087	0.7304	-
Volume Domain (Average Volume)				
3D-UNet (FDK)	38.27(-39.62±9.06)	34.72(6.34±1.45)	0.9585(0.1499±0.0412)	0.0154±38.2148
3D-ResUNet (FDK)	39.86(-38.03±10.53)	34.32(5.94±1.63)	0.9495(0.1410±0.0457)	-8.2486±38.8685
3D-ResUNet+Attn.(FDK)	39.65(-38.24±8.58)	34.35(5.97±1.17)	0.9559(0.1473±0.0406)	-1.9394±39.5164
3D-UNet (iCBCT) [†]	44.20(-32.05±14.65)	33.32(4.65±1.79)	0.9481(0.0780±0.0400)	-3.7927±43.9936
3D-ResUNet (iCBCT)	44.80(-31.46±14.67)	33.22(4.54±1.80)	0.9464(0.0763±0.0385)	-1.9903±44.7111
3D-ResUNet+Attn.(iCBCT)	45.75(-30.50±15.01)	33.05(4.37±1.89)	0.9377(0.0676±0.0406)	-6.0158±45.2901
Volume Domain (Average Amplitude)				
3D-UNet (FDK)	51.67(-35.30±11.08)	32.10(4.59±1.10)	0.9410(0.1418±0.0431)	-3.5407±51.4552
3D-ResUNet (FDK)	51.28(-35.69±11.87)	32.14(4.63±1.16)	0.9417(0.1425±0.0432)	-2.9049±51.1370
3D-ResUNet+Attn.(FDK)	51.87(-35.10±11.78)	32.03(4.52±1.15)	0.9326(0.1335±0.0456)	-6.9976±51.2475
3D-UNet (iCBCT) [†]	55.42(-51.17±11.50)	31.42(5.81±1.33)	0.9300(0.1996±0.0656)	0.7139±55.2177
3D-ResUNet (iCBCT)	55.76(-50.83±12.06)	31.35(5.75±1.39)	0.9282(0.1979±0.0634)	-4.0567±55.4900
3D-ResUNet+Attn.(iCBCT)	58.78(-47.81±11.28)	30.88(5.27±1.28)	0.9131(0.1828±0.0598)	-11.9311±57.1327
Projection Domain (Average Volume)				
3D-UNet (FDK)	73.88(-4.01±1.88)	28.89(0.51±0.33)	0.8654(0.0569±0.0165)	3.8085±73.5703
3D-ResUNet (FDK)	67.91(-9.98±4.86)	29.68(1.30±0.78)	0.8931(0.0845±0.0224)	-1.2820±67.7729
3D-ResUNet+Attn.(FDK)	67.68(-10.21±7.28)	29.71(1.33±0.98)	0.8940(0.0855±0.0232)	-1.5657±67.5189
Dual-Domain (Average Volume)				
3D-UNet (FDK)	49.19(-28.70±6.19)	32.43(4.05±0.62)	0.9377(0.1292±0.0349)	-0.2131±48.9999
3D-ResUNet (FDK)	45.51(-32.38±8.13)	33.07(4.69±0.73)	0.9425(0.1339±0.0406)	-8.9502±44.4396
3D-ResUNet+Attn.(FDK)	45.65(-32.24±9.07)	33.00(4.62±0.82)	0.9396(0.1311±0.0425)	-9.7962±44.3982

Table 5.1: Presented are the quantitative results of DL-based motion correction for CBCT data with simulated motion. The table presents the performance of the proposed motion reduction framework based on the RMSE, PSNR, and SSIM metrics and reports the mean and standard deviation of the body-masked difference (correction) volumes. The metrics are calculated between the reconstructed and ground truth volumes, converted to HU with slope and intercept of 48200 and -1106 , respectively. All numerical values are averaged over the test set. The table shows the average metric together with the average gain (or loss) and the latter’s standard deviation to clarify the contribution of the motion correction. For example, in the last row, the average PSNR is reported as 33.00 dB, corresponding to an average improvement of 4.62 dB, with a standard deviation of 0.82 dB. The models noted by \dagger are used for clinical evaluation (Section 5.5.2) (figure adopted from [12]).

and parameter optimization.

Table 5.1 presents the numerical performance of the various architectures discussed in Section 5.3 for two reconstruction methods FDK and iCBCT, with two different sets of ground truth volumes (“average volume” or “average amplitude”). Three different neural network architectures are investigated: “3D-UNet” (base architecture), “3D-ResUNet” (UNet-based enhanced with ResUNet), and “3D-ResUNet+Attn.” (enhanced using both ResUNet and attention blocks). The ground truth volumes with average amplitude differ more from their corresponding uncorrected volume with motion artifacts than the ones with averaged volume. Therefore, the baseline RMSE is larger for average amplitude, and lower baseline performances in terms of PSNR and SSIM are reported in Table 5.1. Since computing the gradients in the backward pass of the reconstruction algorithm, which is required for training models in the projection domain, is only practical for the FDK reconstruction, Table 5.1 does not present results based on iCBCT for optimizing in the projection domain and dual-domain. The numerical results are reported based on computing the metrics as introduced in Section 5.4.2 between the body-masked ground truth and reconstructed volumes, converted to HU.

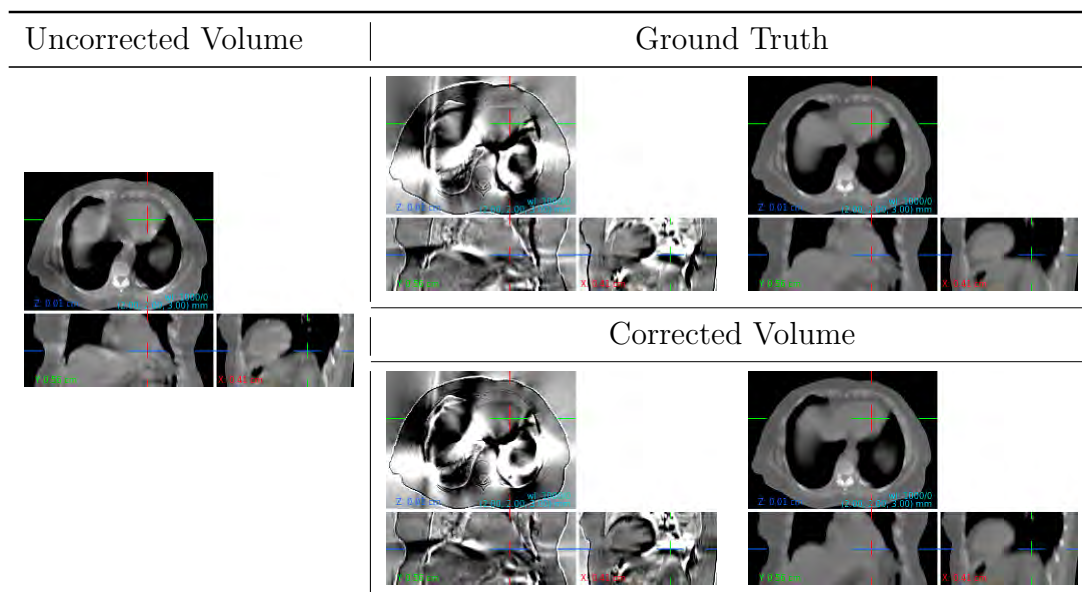


Figure 5.3: Example results for FDK reconstruction (volume domain optimization). Presented is the uncorrected volume using default reconstruction (left), the ground truth volume, both as difference and absolute image (“average volume”, top right), as well as the corrected volume (bottom right). Images are presented in HU with $W/L=1000/0$ (figure adopted from [12]).

The numerical evaluation demonstrates that training 3D-CNNs is consistently successful in compensating motion for both projection and volume domains, with

the best performance being achieved in the volume domain. Numerically, it corresponds to a rise of 6.34 dB in PSNR and 0.1499 for SSIM for FDK with “average volume” ground truth. The highest improvement reported for iCBCT is 5.81 dB of PSNR and 0.1996 in SSIM with “average amplitude” ground truth. Table 5.1 reports a very competitive performance in dual-domain optimization. However, most of the motion correction performance in the dual-domain setting is based on the volume domain corrections. The maximum average gained PSNR in the case of pure projection domain optimization turned out to be 1.33 dB. These results represent the first successful attempt at reducing motion artifacts in CBCT scans using deep neural networks.

The method proposed reduces motion artifacts for two reconstruction techniques (FDK and iCBCT) with several different architectures, including variants with added internal residual connections and/or channel-spatial attention. The motion compensation performance shows a small but consistent variance with the details of the neural network architecture. Reducing the motion artifacts in the projection domain is a subject for further research and optimization due to the more challenging optimization settings. Optimization in the projection domain relies on gradients propagated all the way through the CBCT reconstruction layer and suffers from the large volume of data in the projection space and current GPU memory limitations.

Comparing the two CBCT reconstruction algorithms, iCBCT shows more robustness against motion during acquisition time, and a slightly lower drop in baseline performances is reported. In addition, artifact reduction using 3D-CNNs in the volume domain for iCBCT reconstruction is successful and shows the same results as FDK. Figures 5.3 and 5.4 present example visualizations of the observed motion artifact improvements seen in volume domain learning on top of the FDK and iCBCT reconstructions, respectively.

5.5.2 Clinical Evaluation

To validate the quantitative results of the previous section in a clinical setting, the trained motion compensation CNN models are applied to a real-world test dataset (see Section 5.3.3 and Figure 5.5). Finally, the performance of the motion correction models is evaluated based on the feedback obtained from clinicians.

The real-world CBCT scans used in this study are sufficiently different from the simulated training datasets, e.g., the projection count and HU calibration, to objectively judge the models’ generalization capabilities. The attenuation values of the real-world test dataset are rescaled to match the scale of the training dataset to compensate for the different calibrations.

The expert feedback was collected from a study where clinicians visually inspected 30 pairs of iCBCT reconstructed and motion-corrected volumes, 15 each for ei-

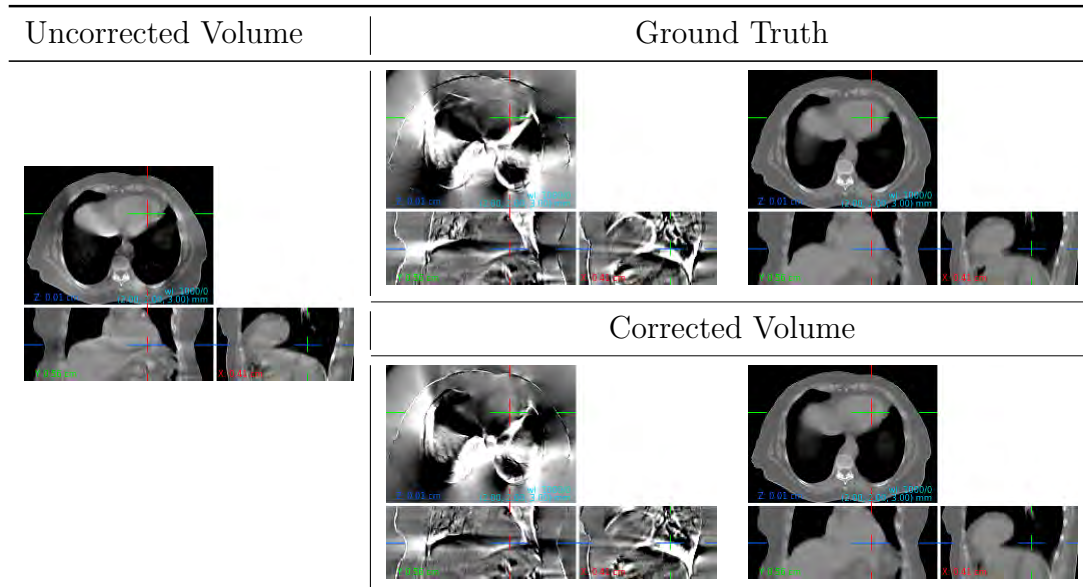


Figure 5.4: Example results for FDK reconstruction (volume domain optimization). The uncorrected volume using default reconstruction (left), the ground truth volume, both as difference and absolute image (“average volume”, top right), as well as the corrected volume (bottom right) are depicted in the table. Images are presented in HU with W/L=1000/0 (figure adopted from [12]).

ther a model trained using average amplitude or average volume ground truth. The best performing CNN architectures from Table 5.1, UNet in volume domain without residual connections or attention, were used for clinical validation. Subsequently, 20 clinicians, including radiation oncologists, medical physicists, radiation technologists, and physicians, answered several questions about their preferences for using CNN models to reduce motion artifacts compared with the standard reconstruction. Each of the clinicians identified themselves as one of three general categories: physician (26%), medical physicist (37%), and dosimetrist/radiation technician (37%).

Initial feedback on the iCBCT datasets indicated the presence of severe and mild unavoidable real-world artifacts besides motion in 34% and 20% of the scans, respectively. The study participants specified their level of agreement or preference concerning (a) a reduction of the observed motion artifacts and (b) the use for various applications, including dose calculation, patient positioning, and segmentation.

This clinical evaluation, the first of its kind to the best of our knowledge, faced the challenge of subjective assessments from experts with different clinical backgrounds. For example, physicians reported a noticeable or strong improvement in CNN-based motion artifact reduction using average volume ground truth in

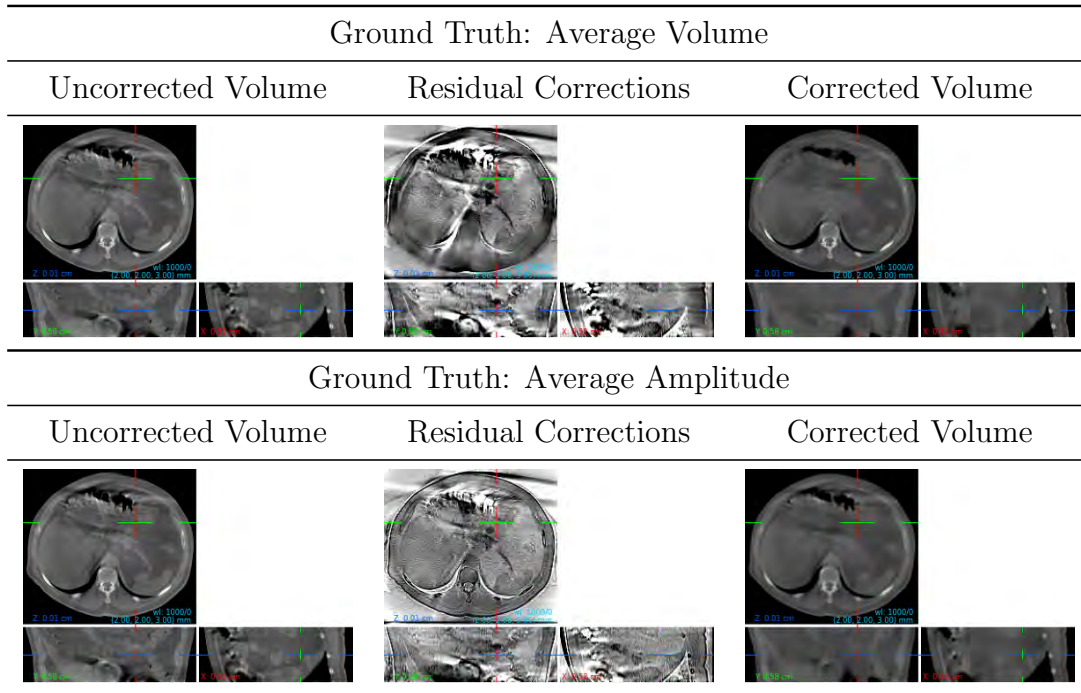


Figure 5.5: The table shows example results for iCBCT reconstruction for real-world test dataset, using the two options for the choice of ground truth. The uncorrected volumes using default reconstruction (left), the residual corrections (middle), as well as the corrected volumes (right) are presented (figure adopted from [12]).

80.00% of scans, while medical physicists only reported this in 65.83% of the scans. Nonetheless, medical physicists preferred CNN-corrected volumes in 63.33% of the cases for dose calculation, while the physicians reported this in only 30.67% of the cases.

Table 5.2 presents the average overall votes and the final clinical evaluation results. Despite the differences in the improvements reported by the different experts, there is a clear positive trend showing that the proposed CNN models are indeed able to reduce motion artifacts successfully. In addition, clinicians reported a tendency toward using CNN-corrected images (using average volumes ground truth) for plan adaptation and dose calculation. One area where clinical experts preferred to use images without CNN-based reconstruction is for soft-tissue-based patient positioning and manual or automatic tissue segmentation, as these images are typically sharper compared with the CNN-corrected ones.

Nevertheless, quantitative evaluation to compute the level of agreement when applying an automatic segmentation algorithm using CBCT images with and without motion artifact correction leads to overwhelmingly positive results. The

Ground Truth → ↓ Application / Preference →	Average Volume			Average Amplitude		
	CNN (%)	Equal (%)	Standard (%)	CNN(%)	Equal(%)	Standard(%)
Motion artifact reduction	74.00	26.00	-	58.33	41.67	-
Plan adaptation and dose calculation	49.33	22.00	28.67	26.33	17.33	56.33
Soft-tissue-based patient positioning	23.00	12.67	64.33	13.00	7.00	80.00
Manual and automatic tissue segmentation	24.33	14.67	61.00	13.00	10.33	76.67

Table 5.2: Results of the clinical evaluation. This table shows the preferences for CNN-based or default iCBCT reconstruction when using CNN models trained using either average volume or average amplitude ground truth concerning motion artifact reduction and potential applications such as plan adaptation and dose calculation, patient positioning and segmentation (table adopted from [12]).

average dice score measures the automatic segmentation contours in original and motion-corrected volumes. This score is averaged over 18 organs or tissues, visible in most CBCT scans, including pulmonary arteries, breast, chest wall, lung, ribs, and spinal canal. The high dice score of 0.89 (0.88) when using average volume (average amplitude) ground truth demonstrates a very high level of consistency between the obtained segmentation contours despite the low preference reported by clinical experts for using the motion-corrected images for segmentation.

5.6 Discussions and Conclusions

This chapter presents the first DL-based method for globally reducing motion artifacts in reconstructed 3D CBCT images, built on top of the two reconstruction algorithms FDK and iCBCT. Neural network architectures which act either on the reconstructed CBCT volumes, the input X-ray projections or both were trained in a supervised way using a motion simulation framework to provide motion-free ground truth. The experimental results demonstrate that DL-based architectures can correct motion artifacts. So far, the best results have been obtained in the volume domain through the implementation of a refined U-net architecture.

The quantitative evaluations demonstrate that using DL through deep neural network architectures yields significant improvements in image quality and reduces motion-induced artifacts in CBCT scans. In addition, a clinical evaluation was performed, in which clinical experts confirmed the principal quantitative results for motion artifact reduction using a real-world test dataset. Clinicians confirmed that artifacts are reduced and expressed a preference for using CNN-corrected CBCT images for dose calculation. However, for patient positioning or segmentation, this could not yet be demonstrated.

There are several related avenues that could be explored in future research:

First, the presented results show promising improvements, mainly in the volume

domain, independent of the acquisition parameters and reconstruction technique. However, there is room for improvement in the projection and dual-domain setting. One potential reason is the processing of the projections in batches due to GPU memory limitations, which leads to a loss of correlation between different projection batches separately processed by the neural network. In addition, great care is necessary to ensure the backpropagation of gradients through the CBCT reconstruction layer to provide a meaningful and noise-free learning signal in the projection domain.

Second, models trained using supervised learning typically suffer from generalization to data acquired in entirely different settings. Although the calibration technique used in this study successfully reduced the performance gap between the performance of the models on simulation and real-world datasets, generalization to highly different acquisition setups and other anatomies is not certain. This provides motivation for further investigation of unsupervised learning and/or domain adaptation techniques.

Third, the current motion simulation only simulates respiratory motion and does not include other effects, such as cardiac motion. Therefore, tackling cardiac motion in chest CBCT scans combined with respiratory motion remains an open problem. This method could also be extended to handle abdominal CBCT scans, including different motion effects.

In conclusion, while the initial results are very promising, future research will aim to improve adaptive treatment capabilities in IGRT, including patient positioning and tumor targeting, auto-segmentation, and dose calculation applications directly on the radiotherapy device.

6 Applications in Affective Computing, Medical Imaging and Beyond

This thesis presents several contributions to diverse and interdisciplinary research niches among many applications of machine learning (ML) and deep learning (DL). Although the original papers in this section have much more extensive content, this chapter summarizes the most scientifically thrilling findings and lessons learned from applying ML and DL in the real world. This chapter is broader in terms of the various applications discussed, more diverse than previous chapters, more straightforward, and more interesting because of its diversity.

Despite the brief overviews, this chapter presents many interesting practical findings and insights that are beneficial in applied research and tuning ML and DL methods to their best performances. Furthermore, this chapter describes several practical problems in ML and DL, such as affective computing, pain estimation, and data homogenization, and identifies initial solutions to this particular research area. Finally, similar to the previous chapters, the remainder of this chapter discusses some exciting directions for future research.

The chapter is organized as follows: discussing solutions for facial expression estimation, emotion recognition, and findings on automated ML and DL focusing on bringing neural networks to their best performances. The chapter then presents two medical applications targeting signal processing for pain estimation and image processing for data homogenization for DL pipelines. The last two sections of this chapter introduce and elaborate on two well-known challenges of DL: fairness and robustness.

6.1 Affective Computing

This section presents two applications of machine learning in affective computing. It begins with explaining the application of support vector machines in facial expression estimation, followed by emotion recognition using audio-visual features.

6.1.1 Facial Expression Estimation

Human facial expressions can reveal information about their affective states [311] or cognitive load[134], which are crucial in human-computer interaction (HCI). There is extensive literature, and several surveys have been developed around facial expression estimation due to its importance. Researchers have divided the human face into several regions called action units (AUs) to quantify facial expressions. Figure 6.1 shows a few such action units defined in the literature to measure facial expressions. The activity of AUs can be measured based on binary occurrence labels (active/deactive) or quantified in terms of intensity in discrete activation levels from zero to six. The predictions of AU intensities per frame can be considered a time series. One can compare measures such as root mean square error (RMSE) and Pearson correlation coefficient (PCC) between predictions and ground-truth labels. Since the labels are discrete, it is also possible to compute the mean intraclass correlation coefficient (ICC) as a performance measure [240].

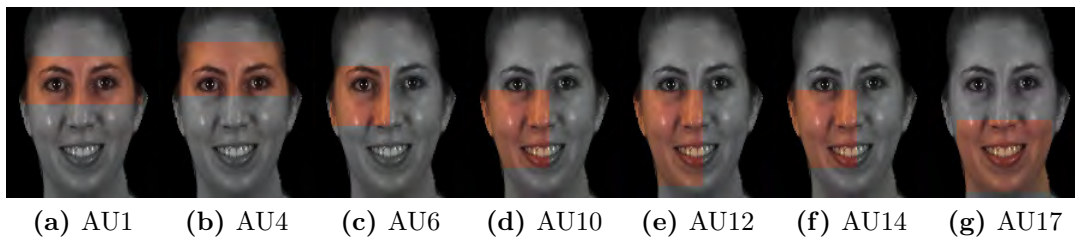


Figure 6.1: Several action units (AUs) used for facial expression estimation (figure is adopted from [8]).

According to the usage of the ground-truth labels, AU activities can be predicted in binary occurrence or by estimating the intensity of a specific AU. Different types of neural networks can be used based on the goal of the facial expression task. For instance, a binary classifier can predict the occurrence of activation in an AU, or its level can be predicted using a multi-class classifier or a regression model.

The presented method in this section for AU intensity estimation consists of multiple steps, starting with preprocessing for face alignment followed by training facial expression templates from data using K-SVD dictionary learning (see Figure 6.2). Then, each input image's features are computed based on their projec-

tion onto dictionary elements to compute the facial features. Support vector machine (SVM) based classifiers and regression models are trained for AU occurrence and intensity estimation based on the computed features. Dictionary-learned features improved the baseline results using conditional random fields [276] by 35%. However, the DL-based method achieved superior performance on an unseen test dataset reported in the 3D facial expression recognition and analysis challenge (FERA 2017) [324].

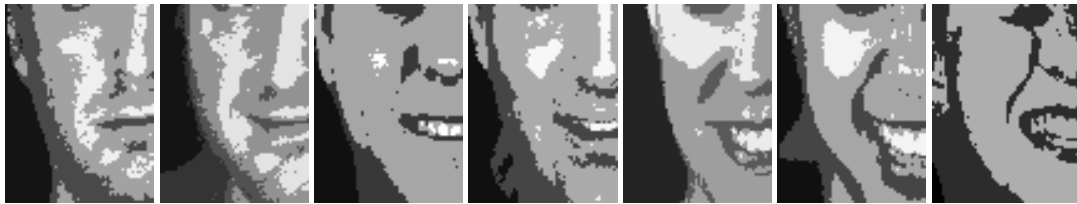


Figure 6.2: Several dictionary-learned facial templates used for facial feature extraction (figure is adopted from [8]).

The main practical insights of this section can be summarized as follows: 1) Pre-processing steps such as face alignment (especially when images are from different head poses) are vital for ML and DL-based approaches. 2) The performance of deep learning models has exceeded the ML-based methods for a long time; however, training regression models instead of classifiers imposes much more optimization effort and hyperparameter tuning overhead in DL than in ML. 3) Not all classes are located at an equal distance in the embedding space. Hence, using a hierarchical classifier to identify the occurrence and then predict the intensity of an AU improves the performance.

More details on implementing dictionary-learned feature extraction techniques used with a support vector regression model for facial expression recognition are presented in [8].

6.1.2 Emotion Recognition

Affective computing and, more specifically, human emotion recognition is an interdisciplinary field linking cognitive sciences, psychology, and computer science. It has recently attracted more attention in the context of human-computer interaction (HCI) to interpret and understand human behavior and emotions. The idea of automatic emotion recognition is useful for making various sensory measurements, including facial video, audio, and physiological signals to predict human affective status and emotions according to the changes in the sensory information.

Besides discrete emotional categories such as happiness, surprise, fear, anger, and disgust, there are two continuous dimensions of arousal and valance which

can express human emotions. Arousal shows a human's level of activeness and engagement in a specific situation, while valance indicates the positiveness of a human's feelings. Arousal and valance levels can define different human feelings and affective statuses. For instance, happiness and excitement share a positive valance level with low and high arousal levels, respectively. Similarly, sadness and anger are positioned on the negative side of the valance scale with low and high arousal levels [209]. Furthermore, researchers have also figured out that the affective status of humans in social interaction has another dimension described as dominance [126].

Based on the briefly explained theory, Ringeval et al. designed an experiment and created the RECOLA multimodal corpus for emotion recognition [215]. The human subjects participating in this experiment were divided into two groups for interactive sessions. Then, they were presented with a task, such as surviving an air airplane crash in the middle of a jungle. The participants were given a list of tools and asked to rank the list of tools according to their preference independently. Then, the participants of each group were connected via video call to discuss their solutions and the organizers recorded the audio-visual information and physiological signals during this interactive session. According to the progress of the discussions, participants experienced different types of natural emotions with various intensities. After the interactive session, six psychologists rated the participants' emotions based on two dimensions: arousal and valance. Then, a gold standard label was computed based on the inter-rater agreements of the psychologist's ratings at every time step. Next, audio-visual and bio-physiological information processing occurred to quantify the participants' emotions automatically through the reproduction of the gold standard labels.

Valstar et al. processed the raw audio-visual and bio-physiological data of the RECOLA dataset into features that are ready for ML-based pipelines for prediction [275]. They extracted local Gabor binary patterns from three orthogonal planes (LGBP-TOP) [4] for appearance features and extracted facial landmarks to evaluate the face geometry [305]. Moreover, they used the COVERED toolbox to extract voice quality and spectral features from audio [56]. Bio-physiological recordings include electrocardiogram (ECG) and electrodermal activity (EDA) signals. All bio-physiological signals pass through band-pass filtering in the pre-processing step. The ECG signals are processed to extract heart rate (HR) and its measures of variability (HRV) as features for emotion recognition [214]. EDA signals are also decomposed to their rapid and transient component called skin conductance response (SCR) as well as a slower basal drift denoted as skin conductance level (SCL) [52]. Valstar et al. computed four statistical features from both EDA components and their first derivative to use in the emotion recognition pipeline [275].

This thesis offers an ML-based pipeline for processing the features' information

and creating predictions from multiple data modalities. Figure 6.3 demonstrates this pipeline, including random forests (RF [29]) for training regression models, which predict the continuous labels in two dimensions of arousal and valance based on extracted features for the four data modalities of audio, video, ECG and EDA. Furthermore, since the psychologists evaluated the emotional status of the participants based on their audio-visual signals, these modalities contain more information, and RF models have created more precise and less noisy predictions using audio-visual features.

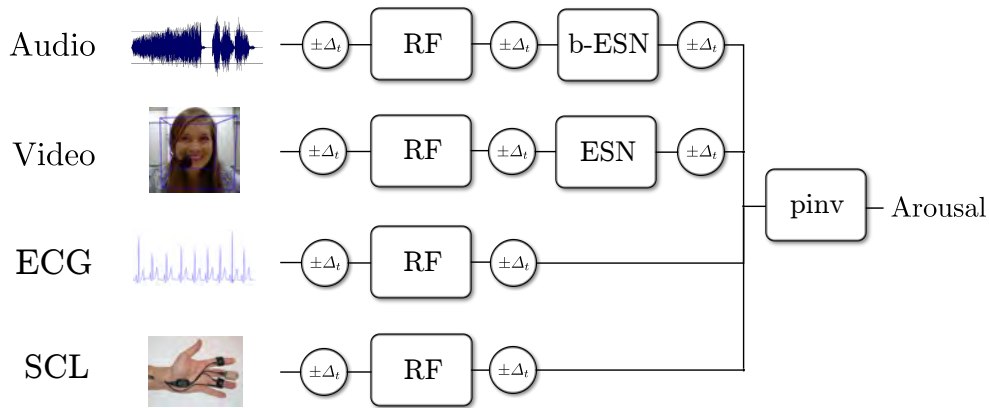


Figure 6.3: Presented is the proposed sequence of blocks for the automatic fusion of audiovisual and biophysiological information to predict arousal levels (figure is adopted from [10]).

Due to the precise prediction from audio-visual features, it was possible to train a recurrent model based on echo state networks (ESNs [112]) to refine the prediction and consider temporal information. The idea of reservoir computing inspires ESNs. They are models with time series as input (here, the sequence of arousal or valance predictions) with internal states connected by random weights. The models' output weights are trained to minimize the loss function between the internal state of the reservoir and ground-truth labels (see Figure 6.4). The emotion recognition pipeline described in this section uses a bi-directional ESN for audio and a standard ESN for video modalities which were selected based on their performance on each data modality. The internal state of the reservoir in bi-directional ESNs depends not only on previous samples in the time series but also on the samples that follow. An alternative to ESNs are long short-term memory (LSTM [100]) models as well as their bi-directional version [233].

Multimodal information fusion is the final and essential component of the emotion recognition pipelines. Different modalities contain various levels of information for arousal and valance tasks. For instance, the arousal predictions from audio are more accurate, and video modalities include more information for the valance. Hence, averaging the predictions of all modalities does not lead to opti-

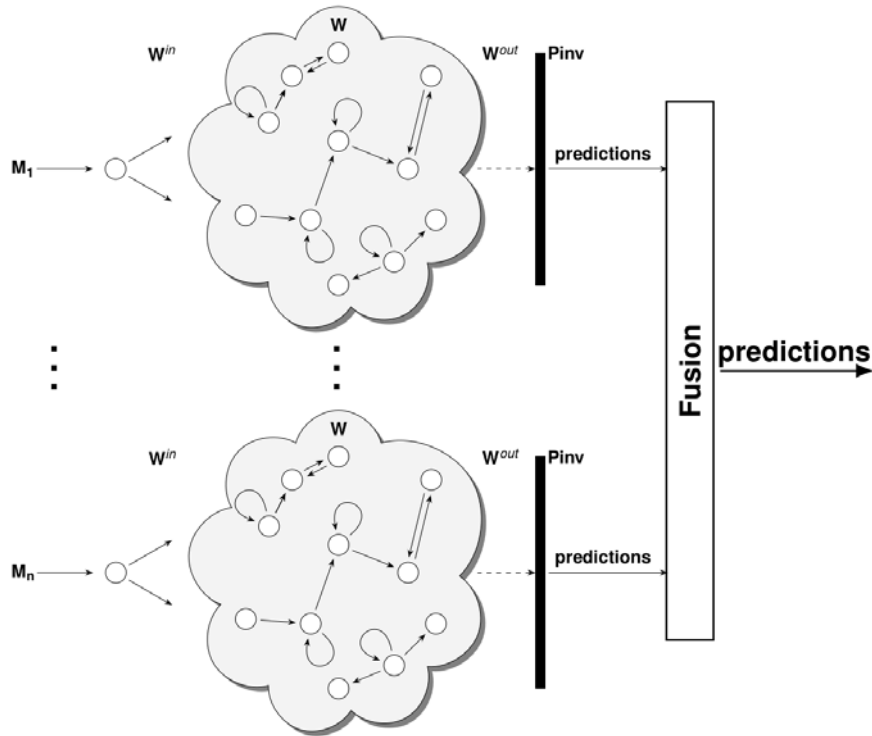


Figure 6.4: Presented are the echo state networks (ESNs) based architectures for modeling temporal information dependencies and fusing multimodal information. One ESN is trained for each modality, and the predictions of all modalities are combined using precomputed weights according to the importance of modalities per task.

mal predictions. The pipeline presented in Figure 6.4 uses Moore pseudo-inverse to minimize the mean square error of multimodal information fusion predictions and gold standard labels. Thus, the final information fusion block is a linear combination of the predictions from each modality. More details about numerical results and methodology are presented in a paper published in conjunction with the audio-visual emotion recognition challenge (AVEC) [275].

The most intriguing path for future research in emotion recognition is multimodal and temporal information fusion, in addition to multi-task learning. The information fusion technique presented in this section used universal weights to combine the modalities. However, the importance of different modalities can differ from time to time based on events in the audio-visual and bio-physiological signals. Thus, developing adaptive fusion techniques with attention mechanisms can significantly improve the fusion's performance. Furthermore, arousal and valence are predicted separately in this work. These tasks can be combined into a multi-task learning problem to train models that fine-tune the predictions of one emotion dimension by being aware of the other. The ratings from psychologists and gold standard labels are not necessarily temporally synchronized with

physiological and audio-visual information. Tracing such temporal mismatches between features and labels is still cumbersome, which we tackle by finding the optimal shifts as a hyperparameter (see Figure 6.5). Developing more sophisticated methods to model such temporal dependencies in high-dimensional features or video is another thrilling venue for research.

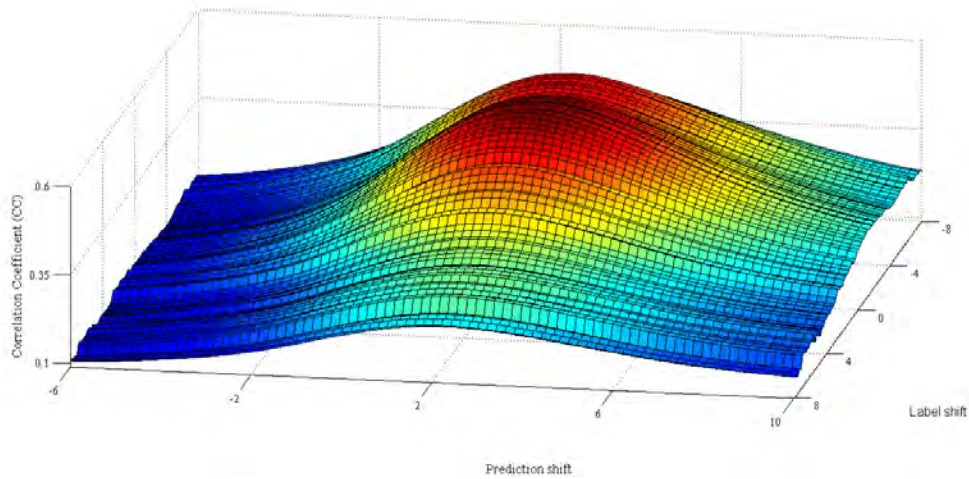


Figure 6.5: The figure depicts the temporal mismatch between audio features and gold standard labels. The average performance of predicting the arousal level of participants from audio increases considerably when features are aligned with gold-standard labels (figure is adopted from [10]).

6.2 Automated Data Analysis

The breakthroughs in ML and DL provide the opportunity to leverage an immense amount of data to approximate almost any function and draw arbitrary decision boundaries for classification and regression. However, the main consequence of such a large degree of freedom was the very challenging task of selecting suitable models with optimal generalization capabilities and a set of hyperparameters for a given dataset. Researchers used to tune the parameters empirically and put their inductive biases into models. With the rise of computing resources, searching architectures and hyperparameter spaces became feasible and popular. The body of literature focusing on automated ML (AutoML) and automated DL (AutoDL) is massive, and this section offers several insights found whilst researching these two subjects.

6.2.1 Automated Machine Learning

The goal of automated machine learning is defined as solving the combined model selection and hyperparameter (CASH) optimization problem. The main goal is to develop search algorithms that can adapt to the tasks based on new trials or even further leverage previous experience from other datasets (referred to as meta-learning in literature [278]). The AutoML challenge series provided a framework to compare methods targeting the CASH problem [94]. The idea of this challenge is to provide two independent sets of benchmark datasets for training and testing with different tasks such as regression and various types of classification. The performance of methods developed for AutoML has been evaluated with strict limitations on time and resources. Several strategies have been adapted and used in the literature to solve the CASH problem automatically. The most straightforward strategy to target the CASH problem is the random search method. Although it is a naive search strategy, random search can achieve competitive results for a new task when no similar dataset or problem is available (see Table 6.1). Furthermore, it is possible to use evolutionary selection for tuning the choices of models and hyperparameters for the target datasets. The idea is to choose the subsequent models and hyperparameters based on the previous best-performing ones. Olson et al. proposed an evolutionary algorithm for the CASH problem called the tree-based pipeline optimization tool (TPOT [196]).

Several strategies have been adapted and used in the literature to solve the CASH problem automatically. The most straightforward strategy to target the CASH problem is the random search. Although it is a naive search strategy, the random search can achieve competitive results for the new task when no similar dataset or problem is available (see Table 6.1). Furthermore, it is possible to use evolutionary selection for tuning the choices of models and hyperparameters to the target

datasets. The idea is to choose the subsequent models and hyperparameters based on the previous best-performing ones. Olson et al. proposed an evolutionary algorithm for the CASH problem called the tree-based pipeline optimization tool (TPOT [196]).

One method used to incorporate the previous experiences from other datasets is Bayesian optimization [77]. The idea is simple, hence practical. The idea is to consider functions of choice (commonly Gaussian processes) with free parameters defining the space drawn by hyperparameters and objective functions. Free parameters of the Gaussian processes are updated after every trial of a set of HPs, and the performance on a given dataset is computed. The parameters of Gaussian processes, in this way, learn the connections between hyperparameters and performance on the task. Models trained based on meta-learning can transfer knowledge from training datasets to new target datasets. Auto-Sklearn is an example of such a method that not only selects the best model and hyperparameters based on the target dataset but also learns from previous runs on different datasets [76].

Dataset	Task	Metric	Random-Search		Auto-Sklearn		TPOT	
			Test	Time	Test	Time	Test	Time
Cadata	Regression	R2 (coefficient of determination)	0.7119	55.0	0.7327	54.9	0.7989	54.6
Christine	Binary classification	Balanced accuracy score	0.7146	99.4	0.7392	99.3	0.7442	105.1
Digits	Multiclass classification	Balanced accuracy score	0.8751	201.2	0.9542	201.2	0.9476	207.2
Fabert	Multiclass classification	Accuracy score	0.8665	77.5	0.8908	77.4	0.8835	78.5
Helena	Multiclass classification	Balanced accuracy score	0.2103	190.2	0.3235	216.4	0.3470	197.5
Jasmine	Binary classification	Balanced accuracy score	0.8371	24.1	0.8214	24.0	0.8326	25.9
Madeline	Binary classification	Balanced accuracy score	0.7686	48.3	0.8896	48.2	0.8684	53.0
Philippine	Binary classification	Balanced accuracy score	0.7406	56.3	0.7634	56.2	0.7703	56.4
Sylvine	Binary classification	Balanced accuracy score	0.9233	28.9	0.9350	28.9	0.9415	29.0
Volkert	Multiclass classification	Accuracy score	0.8154	122.3	0.8880	122.2	0.8720	125.5
Average Performance			0.7463	90.31	0.7938	92.85	0.8006	93.26

Table 6.1: Performance of three automated machine learning algorithms with different paradigms on AutoML challenge datasets and their convergence time [94] (table adopted from [272]).

6.2.2 Automated Deep Learning

The mainstream research in AutoDL presented in Section 2.1.7 focuses on developing novel vision architectures mainly based on the ImageNet dataset. However, the other open research question with more relevance to practical problems is finding the optimal architecture and set of hyperparameters for a given dataset that is not necessarily large in terms of the number of images and classes. Searching for solutions to the AutoDL problem inspired the series of AutoDL challenges to find lightweight models with hyperparameters that can quickly adapt to new but small datasets [160]. The target of these challenges was the area under the learning curve (ALC) instead of the final or best performance. Hence, models

converging faster outperform those with slow learning and better final performance based on the final evaluation metric. This evaluation metric highly favors lightweight models, which can be fine-tuned for new datasets very quickly.

Deep convolutional neural networks (CNNs) outperformed the classical methods on AutoDL for vision. Due to their design, which is optimized to learn representations from a large dataset, pretraining on ImageNet is still an undeniable part of the model preparation. The performance of small models such as ResNet18 [96] and MobileNet-V2 [225] which have been trained on a small dataset, show consistency by changing their learning rate for a fixed pipeline; hence, a fixed learning rate can be used for different datasets (see Figure 6.6). However, regularization shows a more critical role in optimally fine-tuning the models to small new datasets. It is no wonder that the winning solutions of AutoDL contained the *fast auto-augment* method to learn augmentation strategies tailored for a given dataset.

Research in developing models for audio processing falls behind vision systems with respect to lightweight architecture searched models on large audio datasets. For example, a commonly used pretrained network for audio processing is VG-Gish [159] trained on Youtube-8m dataset [189], which is far from light-weight. Hence, searching for appropriate architectures is pivotal when searching for optimal models for small datasets. Similarly, augmentation strategies are not as well explored in audio processing, and a significant boost is expected with the development of more suitable or automated augmentation techniques.

Despite the differences between audio and video processing pipelines for research in AutoDL for audio-visual data, the block diagram used for pattern classification can be summarized with similar components as depicted in Figure 6.6. Preprocessing the data is the first step which computes the spectrogram for audio data, augmentation for images, or selects key frames from videos. Then, the raw information can be processed into latent representations using convolutional backbones. Information fusion in the following steps combines the information along the axis of time via convolutions or spatially using global pooling. The last layer is a fully connected classifier to predict the patterns from the models' final embeddings. The main advantage of such a similar architecture is the possibility of mid-level information fusion between audio-visual modalities in applications such as emotion recognition, explained in Section 6.1.2.

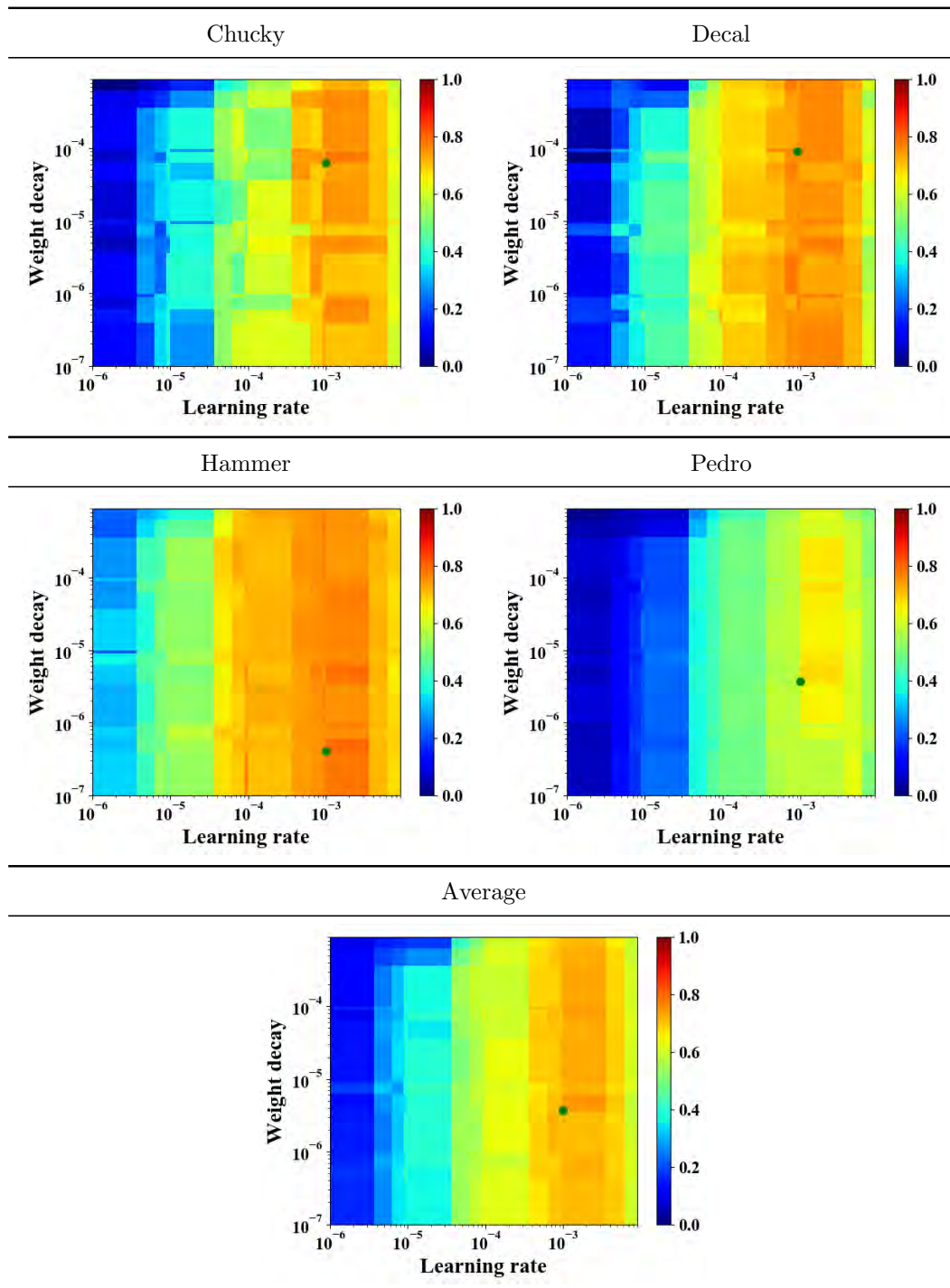


Figure 6.6: Performance of four different vision datasets in terms of ALC of MobileNetV2 as a function of weight decay and learning rate (top two rows) and averaged performance over all datasets (bottom row). The green dot indicates the best performance (figures are adopted from [271]).

6.3 Medical Diagnosis and Imaging

This section presents two medical applications for machine and deep learning methods. First, the application of machine learning in pain detection in medicine through bio-physiological signal processing is explained. Second, data homogenization for medical images with future applications in merging datasets and image preprocessing in federated scenarios is discussed.

6.3.1 Pain Estimation

ML and DL methods have been widely used in medical applications such as pain estimation. Pain is an evolutionary mechanism developed in human bodies to stop and prevent external damaging stimuli or harmful behaviors. However, pain also appears as a consequence of operations in clinical settings. Not all patients, such as neonates, unconscious patients, or patients with cognitive or communicative impairments, are capable of communicating the location and level of pain when seeking treatment. Hence, automatic pain detection and intensity estimation have become more popular among researchers.

Werner et al. introduced the Biovid heat pain database for automatic pain estimation from bio-physiological signals [282]. The idea of the experiment was to estimate stimulated pain using heat induced by a thermode. The experiment started with a calibration phase when the organizers measured the participants' pain perception and tolerance thresholds. Then, the experiment began with a cold thermode, which became increasingly hotter until the participant noticed the pain (perception threshold), and stopped when the heat became unbearable for a participant (tolerance threshold). Then, the temperature between these two thresholds was linearly divided into four levels, and the participants were stimulated with four pain levels during two parts of the experiment. Each part contains twenty episodes of pain stimulation with a duration of four seconds with a break of approximately eight seconds. The bio-physiological signals were recorded during the experiment for signal processing and automated pain estimation. The signals recorded in this experiment include several data modalities such as electromyography (EMG), electrocardiography (ECG), and electrodermal activities (EDA).

After data collection, bio-physiological signals are preprocessed for feature extraction. Multiple time and frequency domain statistical features are available and computed for pain detection and pain level estimation [118]. The key component improving the pain estimation accuracy in this stage is the extraction of the modality-dependent features, especially in electrodermal activity (EDA) signals which contain the most relevant information for pain estimation [9]. There needs for more research on bio-physiological pain estimation in order to be able to use

supervised DL methods to optimize features (embeddings) automatically instead of computing hand-crafted features. However, this shortcoming is to some extent addressed using the *unsupervised representation learning* for bio-physiological signals [266].

Pain estimation based on the Biovid heat database can be considered a classification or regression task. Feature preprocessing considerably affects the accuracy of pain level quantification, and normalization of the features based on their mean and variance improves the performance of classifiers and regression models. Further improvements are achieved by normalizing the features per participant based on their baseline level of bio-physiological signals (feature personalization [118]). Moreover, personalization extends to another level by clustering people into several groups using Kullback-Leibler (KL) divergence and finding the closest subjects to train the model based on their data [116]. This improvement in estimating health-related measures using personalization hints at the fruitful direction of personalized information processing and treatment for research in health care. Different classifiers and regression models such as random forests (RF [29]) and radial basis function networks (RBFs [31]) showed a similar performance after tuning, and it is possible to predict the confidence of the estimated level of pain by combining the predictions from several individual models [116].

6.3.2 Data Homogenization

Deep CNNs achieved great success in a wide range of computer vision tasks and improved state-of-the-art performances by a large margin; however, early on, they showed a weakness in generalization in the presence of a change in data distribution or concept drifts. This thesis offers an idea to deal with the changes in data distribution through data homogenization and merging multiple datasets. Deep learning and computer vision literature is full of attempts at domain adaptation [49, 288, 299] and style transfer research [83, 326]. However, merging a few datasets into a unified style using a preprocessing network is the novelty of the idea presented in this thesis.

The research presented in this section is conducted in the context of COVID-19 detection from 2D chest computed tomography (CT) scans. This thesis presents a preprocessing network (PrepNet [11]) aiming at data homogenization with minimum changes in the original images. Accordingly, the proposed techniques have two main components: an autoencoder and a dataset/technology classifier. (see Figure 6.7). The autoencoder-based CNN aims to find common ground for all the datasets and preprocesses them with minimal changes to fool the dataset classifier. The autoencoder and dataset classifier models are trained one after the other at each step of the training process. The preprocessing network aims at fooling the dataset classifier by erasing the differences between dataset samples,

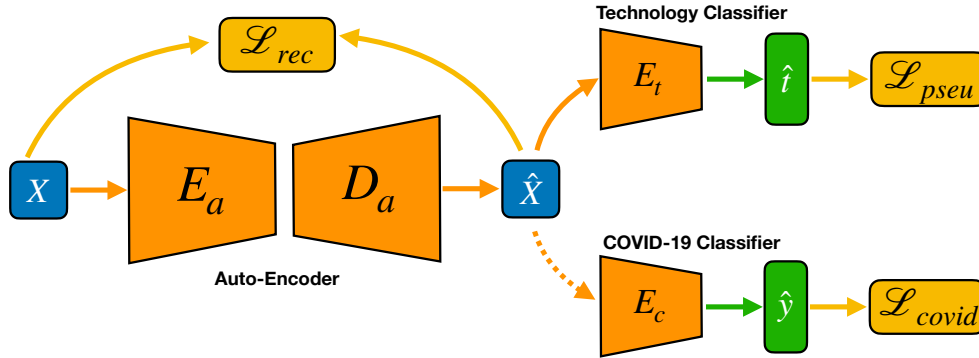


Figure 6.7: The proposed architecture for *PrepNet* model with three modules: (i) an auto-encoder aims at CT dataset homogenizer; (ii) a multiclass classifier to recognized CT-datasets; and (iii) a binary classifier for diagnosis (COVID-19). The loss functions of the dataset classifier and auto-encoder were trained adversarially against each other. The binary classifier for diagnosis (COVID-19) was trained independently using the preprocessed scans by auto-encoder (figure adopted from [11]).

while the dataset classifier learns the differences between the new preprocessed scans. Two networks compete against each other to improve their performance in a similar optimization as generative adversarial networks (GANs). After sufficient training with the correct set of hyperparameters, the auto-encoder learns to bring the datasets into a joint distribution that looks similar to the human eye as well as CNNs. During the optimization, we minimize the reconstruction loss of the auto-encoder to keep the scan as unchanged as possible and only focus on erasing the dataset differences and reducing probable generative artifacts.

Test dataset → Dataset portion	BA	SARS-COV-2			UCSD COVID-CT				Within Test Average	Cross-Dataset Average	Pre-trained encoder
		Sens	Spec	AUC	Test	Sens	Spec	AUC			
<i>COVID classifier</i>											
SARS-COV-2	0.8924	0.9292	0.7876	0.8584	0.4433	0.7835	0.1262	0.4548	0.8587	0.4159	Yes
UCSD COVID-CT	0.3295	0.3476	0.2743	0.3110	0.8250	0.7113	0.9320	0.8216	(baseline)	(baseline)	
<i>AutoEncoder</i>											
SARS-COV-2	0.8956	0.9907	0.6460	0.8183	0.4983	0.9175	0.0970	0.5073	0.8555	0.4836	Yes
UCSD COVID-CT	0.49405	0.6030	0.3008	0.4519	0.8154	0.7216	0.8846	0.8031	(−0.32%)	(+6.77%)	
<i>PrepNet</i>											
SARS-COV-2	0.9007	0.9353	0.7982	0.8668	0.5157	0.9175	0.1067	0.5121	0.8404	0.5343	Yes
UCSD COVID-CT	0.5545	0.6446	0.1858	0.4852	0.7800	0.8556	0.7087	0.7822	(−1.83%)	(+11.84%)	

Table 6.2: Test and cross-dataset performance of different methods. Using an adversarial loss to train a *PrepNet* improves the cross-dataset average performance (table adopted from [11]).

The evaluation method proposed for *PrepNet* not only measures the intra-dataset test performance, but also focuses on cross-dataset performance. The ultimate goal of *PrepNet* is to homogenize datasets so that the model trained on one can be used for diagnosis on the other datasets. Two public datasets for COVID

diagnosis from CT scans called SARS-COV-2[245] and UCSD COVID-CT[312] are the subjects of this study. Figure 6.8 depicts the performance of our proposed PrepNet and visually compares its results with classical auto-encoders and other preprocessing techniques for chest CT scans. Table 6.2 shows the performance of the models trained on the original dataset, solely preprocessed using an auto-encoder learned in a self-supervised manner on reconstruction loss and preprocessed using PrepNet. PrepNet achieved the best cross-dataset generalization amongst all the presented methods with a minor drop in intra-dataset test performance. However, the gap between cross-dataset performance and intra-dataset performance is still significant, and there is considerable room for improvement in future research.

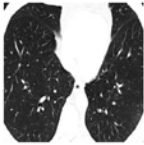
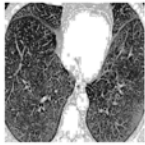
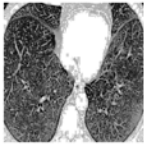
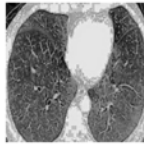
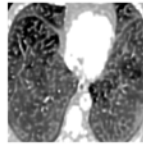



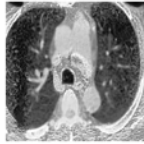
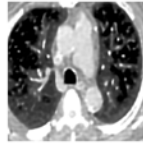



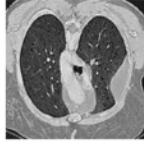
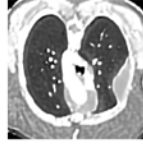

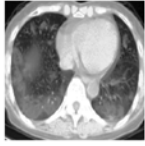
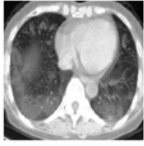
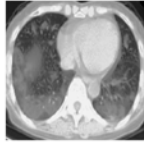
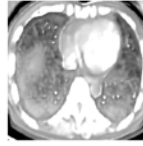
Dataset	COVID	Original	Histogram equalization	Normalization	Auto-encoder	PrepNet
SARS-COV-2	Negative					
SARS-COV-2	Positive					
UCSD COVID-CT	Negative					
UCSD COVID-CT	Positive					

Figure 6.8: Original images from the datasets with different preprocessing methods applied (figure adopted from [11]).

6.4 Face Recognition

Face recognition (FR) and face matching technologies, especially in surveillance applications, were probably the most controversial models developed with (CNNs) for computer vision. The idea of identity matching and verification using images was so appealing for many applications such as online banking or intelligent surveillance that the research literature around developing models and collecting datasets for FR expanded rapidly [178].

Researchers collected clean datasets for FR in academic research developments and datasets from real-world images [284]. Developing loss functions to compute generic embedding was a key component of extending FR to face matching on the face, which has not been seen in the training set. Triplet loss [257, 258, 259] and more modern loss functions such as large margin cosine loss [285] and arccos loss with angular margin [59] are amongst such developments. Despite the scientific successes in this research area, social activists raised issues concerning fairness because of biases in FR systems. This section describes the issue of fairness and presents scientific findings in the context of FR systems.

6.4.1 Algorithmic Bias in FR Systems

The research progress in FR technology was quick, and the models rapidly made their way into practical applications; however, multiple reports show some biases and inaccuracies against races that have fewer images in the training datasets [99, 169, 161]. These incidences attracted much negative feedback from society, which was reflected in the news¹². Another reaction followed this wave with companies starting to ban the FR technology³ [138]. As a result, using the FR technology started to be abandoned, and measuring algorithmic biases became more critical after these events [23].

6.4.2 Measuring Bias and Awareness

The main findings of our research are about methods of measuring and removing biases. After all the controversies regarding biases in FR technology, researchers quickly started to seek strategies for measuring the sources of such biases in FR. The pioneering research on collecting datasets with racial diversity rapidly exposed the gap in FR models' accuracy for different races, which causes limitations in service accessibility where FR technologies are involved and raises ethical

¹<https://www.washingtonpost.com/technology/2021/02/17/facial-recognition-biden/>

²<https://www.bbc.com/news/technology-48276660>

³<https://www.banfacialrecognition.com/>

issues regarding fairness.

Researchers introduced racial awareness as a proxy for measuring biases in FR models and research showed that the FR models distribute the faces based on their ethnicity in the embedding space [289]. Accordingly, several research works suggested adversarially removing the racial information as a solution to the problem of biases in FR systems [306, 131, 314]. However, our research in measuring the biases demonstrated that the intuitive idea that racial clustering in embedding space is correlated with biases is not always true [87]. Instead, the reason behind racial biases comes from how the faces of different races are distributed in the embedding space (see Figure 6.9). Similarly, blinding the FR technologies from racial information in the embedding space does not necessarily lead to decreasing the racial bias [295]. Hence, awareness and bias are two distinct though related issues in FR, and methods dealing with ethnicities individually, such as the research presented in [217] does, are more appealing based on these findings.

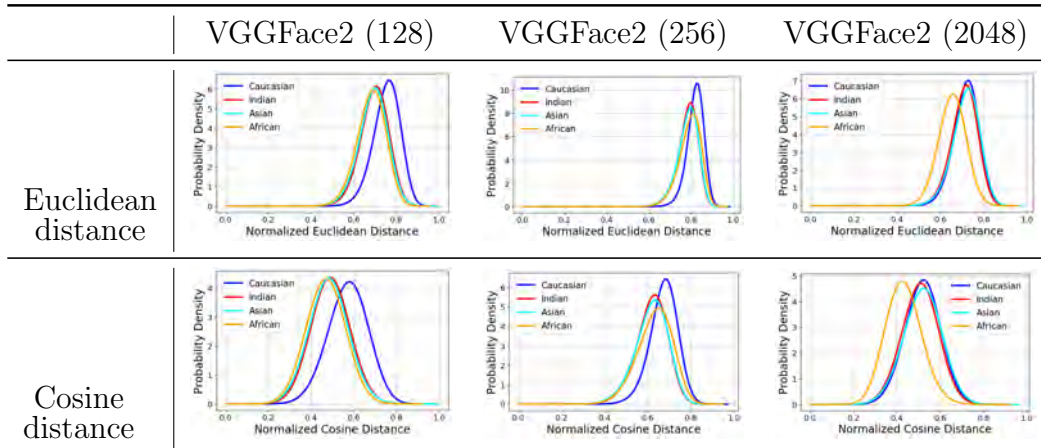


Figure 6.9: Probability density distribution of pairwise (Euclidean and Cosine) distances between test images' embeddings of different races. The embeddings are computed using the VGG model fine-tuned for face recognition (VGGFace2 [35]) with different embedding dimensionalities (128, 256 and 2048). The figure shows that the faces from the Caucasian race, which have the largest share of data samples, have a larger average distance than those of Africans, Asians, and Indians (figure adopted from [87]).

6.5 Rotation-Invariant Vision Transformers

Inductive biases such as translation invariance undeniably accelerated the rapid advances of modern vision models based on convolutions through parameter sharing and improving sample efficiency. However, state-of-the-art models can only partially incorporate rotation invariance. Recent attempts to develop rotation-invariant techniques mainly face the challenge of high memory requirements or limiting the original model capacity. This section proposes an embedding layer method for vision transformers to leverage the invariance of self-attention layers to the order of tokens and train robust models against local and global rotation. The proposed image embedding technique requires negligible memory overhead to train rotation invariance models on large datasets such as ImageNet[223]. Furthermore, the proposed method improves the robustness of vision transformers against rotation on the classification task.

6.5.1 Introduction and Problem Statement

The performance of vision models, more specifically vision transforms (ViTs), drops when the input images are not presented to the models in the original pose. Rotation and scaling are two transformations that researchers found to be a reason for the decline in the performance of vision models from early works⁴. This section presents a solution to rotation invariance in ViTs for object classification. Figure 6.10 shows the decline in the performance of a ViT-based classifier and segmentation model after different degrees of rotation. Besides compensating for the drop in accuracy to improve the robustness of vision models, developing rotation equivariant methods was very appealing to add another inductive bias to enhance the vision models' sample efficiency and convergence speed.

Several different techniques aim to improve the vision model's robustness against rotation. These methods can be divided into two categories: 1) Preprocessing data for training or evaluation. 2) Using equivariance or invariance inductive biases in vision models. The first group uses data augmentation, a conventional technique widely used in deep learning, to increase the size of the datasets artificially, improve the generalization, and train robust models [242, 228]. The second group of the research can be summarized for CNNs and vision transforms (ViT) as follows:

CNNs: Cohen et al. introduced group equivariant convolutional neural networks (G-CNNs) to learn equivariant representation for discrete symmetry groups of rotations [45]. Marcos et al. proposed the rotation of the convolutional filters instead of lifting the representation to the group and using the pooling and vector field representations of the input to achieve rotation invariant, covariant, and

⁴<http://yann.lecun.com/exdb/lenet>

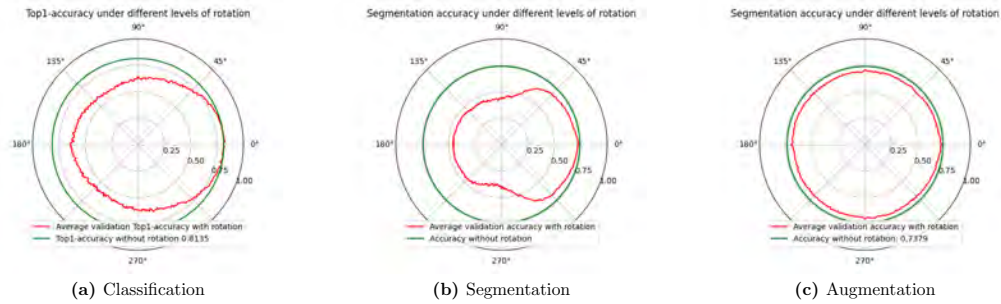


Figure 6.10: Classification and segmentation performance of vision transformers under different degrees of rotation. Augmentation improves the robustness of vision transformers against rotation; however, rotation invariance encoded in the method as inductive bias can improve the sample efficiency of the models.

equivariant features [176]. Lifting the representations or filters to the discrete Lie groups increases the memory consumption linearly with the group size. Alternatively, the convolutional filter can be designed to be equivariant to specific transformations. Esteves et al. train isotropic filters for rotation equivariant CNNs [73] and Weiler et al. proposed learning the models’ weights which are expansion coefficients for the steerable function space [296]. Wiersma et al. presented a surface harmonic network with both invariant and equivariant features [298]. The main disadvantage of optimizing equivariant filters is limiting the capacity of models for learning the data.

ViTs: Romero and Cordonnier used the group lifting concept to train equivariant vision transformers on a discrete group of image rotations [218]. Hutchinson et al. adapted a similar idea, generalized it to the continuous rotation and translation equivariant models, and applied their method to pattern recognition in point-cloud graphs, molecular property prediction, and chasing particle dynamics [111]. Finally, Su et al. demonstrated that rotary positional encoding enhances the performance of natural language processing models [254].

Next, this section reviews the main blocks and concepts used in ViTs. The explanation of rotation invariant and equivariant features is followed by self-attention and formulation under input rotation.

Rotation: Let \mathbf{X} be a vectorized patch of an image with a given size, for example, 16×16 . Then, we define a rotation matrix called \mathbf{R} such that the transformed version of the original image \mathbf{x}_θ can be computed as follows:

$$\mathbf{X}_\theta = \mathbf{R}\mathbf{X} \quad (6.1)$$

where θ shows the angle of rotation. The rotation transformation is defined using a rotation matrix, and it can be computed as follows:

$$\mathbf{R} = \mathbf{X}_\theta \mathbf{X}^\dagger \quad (6.2)$$

The goal of the roto-translation equivariant models with a self-attention layer is computing a representation ($\mathcal{L}(\mathbf{x})$) in which the representations rotates with the same degree as the input image:

$$\mathcal{L}(\mathbf{X}_\theta) \approx \mathcal{L}(\mathbf{X}) \quad (6.3)$$

Alternatively, we can consider the images as a spatial function in 2D space having three values (RGB vector) at every position and define the rotation on every pixel coordinate $((x, y))$ as follows:

$$r_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \quad (6.4)$$

Given the definition of the rotation matrix (r_θ) based on pixel coordinates, the inverse of the rotation operator is equal to its transpose ($r_\theta r_\theta^T = \mathbf{I}$), and the following properties hold accordingly:

$$\begin{bmatrix} x_\theta \\ y_\theta \end{bmatrix} = r_\theta \begin{bmatrix} x \\ y \end{bmatrix} \quad (6.5)$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = r_\theta^T \begin{bmatrix} x_\theta \\ y_\theta \end{bmatrix} \quad (6.6)$$

Rotation invariance, covariance, and equivariance: A representation ($\mathcal{L}(\cdot)$) of an input pattern (X) is invariant to rotation (R) if it does not change with the rotation of the input. The equivariance representations rotate similarly with the input's rotation; However, covariant representations change according to the original representations based on a constant function ($f(\cdot)$). These definitions can be shown in the following equations:

$$\begin{aligned} \text{Invariant:} \quad & \mathcal{L}(X_\theta) \approx \mathcal{L}(X) \\ \text{Equivariant:} \quad & \mathcal{L}(X_\theta) \approx \mathcal{L}_\theta(X) \\ \text{Covariant:} \quad & \mathcal{L}(X_\theta) \approx f(\mathcal{L}(X)) \end{aligned} \quad (6.7)$$

Self-Attention: The output of the self-attention layer for an image ($(X) \in \mathbb{R}^{N \times T}$) converted to N tokenized patches of length (T) can be written as follows:

$$\begin{aligned} \mathbf{Q} &:= \mathbf{XW}_q \\ \mathbf{K} &:= \mathbf{XW}_k \\ \mathbf{V} &:= \mathbf{XW}_v \\ \mathbf{A} &:= \mathbf{QK}^T \\ \mathbf{Y} &:= \text{SA}(\mathbf{X}) \\ &:= \text{softmax}(\mathbf{A})\mathbf{V} \end{aligned} \quad (6.8)$$

where \mathbf{K} , \mathbf{Q} and \mathbf{V} shot the key, query and value. The linear weights used to compute the representations are denoted by \mathbf{W}_k , \mathbf{W}_q and \mathbf{W}_v for keys, queries and values, respectively. \mathbf{A} shows the attention matrix and \mathbf{Y} denotes the self-attention (SA) layer. The softmax function, denoted by *softmax*, is defined as follows:

$$\text{softmax}(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i)}{\sum_j \exp(\mathbf{x}_j)} \quad (6.9)$$

Self-Attention with Rotation: Then, we can write the rotation invariant self-attention objective as follows:

$$\begin{aligned} \mathcal{L}(\mathbf{X}_\theta) &= \text{softmax}(\mathbf{X}_\theta \mathbf{W}_q (\mathbf{X}_\theta \mathbf{W}_k)^T) \mathbf{X}_\theta \mathbf{W}_v \\ &= \text{softmax}(\mathbf{X}_\theta \mathbf{W}_q \mathbf{W}_k^T \mathbf{X}_\theta^T) \mathbf{X}_\theta \mathbf{W}_v \\ &= \text{softmax}((\mathbf{R}\mathbf{X}) \mathbf{W}_q \mathbf{W}_k^T (\mathbf{R}\mathbf{X})^T) (\mathbf{R}\mathbf{X}) \mathbf{W}_v \\ &= \text{softmax}(\mathbf{R}\mathbf{X} \mathbf{W}_q \mathbf{W}_k^T \mathbf{X}^T \mathbf{R}^T) \mathbf{R}\mathbf{X} \mathbf{W}_v \\ &\approx \text{softmax}(\mathbf{X} \mathbf{W}_q \mathbf{W}_k^T \mathbf{X}^T) \mathbf{X} \mathbf{W}_v \\ &= \text{softmax}(\mathbf{X} \mathbf{W}_q (\mathbf{X} \mathbf{W}_k)^T) \mathbf{X} \mathbf{W}_v \\ &= \mathcal{L}(\mathbf{X}) \end{aligned} \quad (6.10)$$

6.5.2 Method and Experimental Results

The mathematical formulation of self-attention with rotation suggests that it is possible to constrain the key, query and value matrices to make self-attention equivariant. The necessary condition is that both rotation matrices (\mathbf{R} and \mathbf{R}^T) can commute⁵ through $\mathbf{X} \mathbf{W}_q \mathbf{W}_k^T \mathbf{X}^T$ and its softmax. This is the necessary condition to make self-attention equivariant ($\mathcal{L}(\mathbf{X}_\theta) = \mathbf{R}_\theta \mathcal{L}(\mathbf{X})$) which is more complicated than invariance, and it is also more appealing since equivariant features are useable in building invariant models. However, invariant models do not necessarily provide equivariant features.

The problem of equivariant self-attention is an open problem for future research. However, this thesis offers a solution to rotation-invariant ViTs based on a fundamental property of self-attention. The self-attention mechanism is invariant to the order of the tokens, meaning the representations do not change when the order of the tokens is different. Therefore, if we transform the image so that rotation only changes the order of the tokens, then the ViT based on self-attention will be invariant and robust against rotation.

The idea of rotation invariant ViTs can be realized using a radial tokenization technique presented in Figure 6.11. The idea is to take the tokens based on the polar coordinate and extract every token from the original image. The proposed

⁵Two matrices \mathbf{X} and \mathbf{W} called to commute if $\mathbf{X}\mathbf{W} = \mathbf{W}\mathbf{X}$

method uses pixel values on a circle's radius placed at the center of the image instead of turning patches of size 16×16 into tokens. Using this embedding method, only the order of the tokens changes with the input rotations, and the whole ViT stays invariant to rotation. This idea works for global rotation; however, it can also be implemented at the patch level to tackle the local rotation of images' elements, which is more critical for medical applications [143].

Steerable Convolutions and isotropic filters inspire this section's other patch embedding techniques. Figure 6.11 shows how isotropic patch embeddings turn the original image into patches. The idea here is to divide the original models into patches of size 16×16 and then sample them via circles around the center of the patch and project them into tokens afterward. Implementing the radial patch embedding technique at the patch level can train robust models against local rotations.

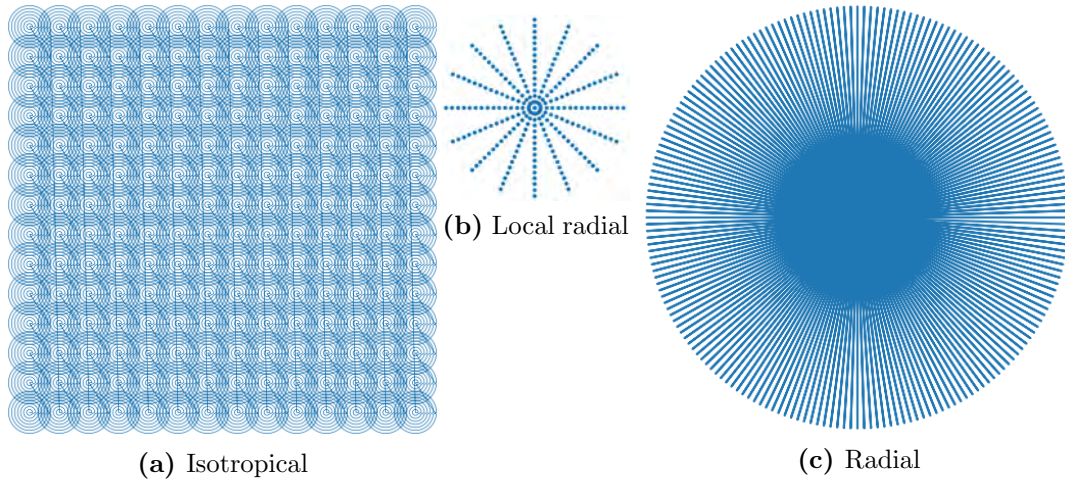


Figure 6.11: The proposed patch embedding methods for vision transformers: a) Isotropic patch embedding for the entire image. Every patch is sampled based on the circles around the center of the patch. b) Radial patch embedding in the patch level technique samples every patch based on the pixels on a circle radius positioned at the patch's center. c) Radial patch embedding for the entire image.

A ViT architecture based on Deit's baseline model[269] is optimized on the ImageNet dataset for pertaining, and initial evaluation shows the functionalities of the proposed methods. Figure 6.12 depicts the performance of the different patch embedding methods used to improve the robustness of the ViTs against rotation. Radial and isotropic patch embeddings demonstrate considerably higher robustness against rotation compared with the original transformer. However, it is notable that training a transformer using data augmentation is a very competitive solution to the presented problem. Table 6.3 shows that rotation invariant patch embedding models pretrained on ImageNet generalize to the other related

Dataset	Accuracy	Base Vision Transformer				Best (SOTA)
		Original	Isotropic	Radial	Local Radial	
Oxford-IIIT Pets [200]	top1	0.9305	0.6890	0.8575	0.8133	0.9710
	top5	0.9926	0.9302	0.9839	0.9725	-
Oxford Flowers [193]	top1	0.9167	0.8000	0.8510	0.8402	0.9976
	top5	0.9696	0.9186	0.9461	0.9461	-
FGVC Aircraft [173]	top1	0.7570	0.3378	0.5794	0.5425	0.9490
	top5	0.9355	0.6439	0.8599	0.8428	-
Caltech Birds [297]	top1	0.7960	0.5221	0.6933	0.6262	0.9548
	top5	0.9462	0.7886	0.8952	0.8714	-
ImageNet [223]	top1	0.7884	0.5698	0.7091	0.6276	0.9088
	top5	0.9370	0.7813	0.8919	0.8364	-

Table 6.3: The performance of rotation invariant vision transformers on several vision benchmark vision datasets. Rotation invariant patch embedding increases the robustness of ViTs at the expense of a decrease in performance.

object detection tasks with a drop in the performance compared to the original transformer method.

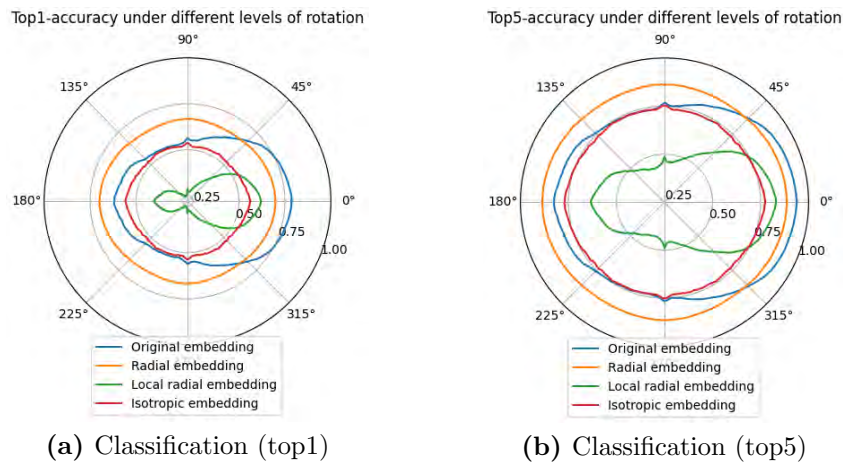


Figure 6.12: Robust training against rotation using rotation invariant patch embedding techniques.

This preliminary study shows that transformers can run in the rotation invariant mode without memory and compute overhead by adjusting the patch embedding techniques. Compared to similar methods such as group equivariant self-attention [218], the proposed method profits from higher memory efficiency and angular resolution. The research questions regarding evaluating rotation invariance to ViT based on the initial motivations and goal, namely sample efficiency and training speed, are still open for further investigation.

7 Conclusions

DL breakthroughs and computer vision models developed based on DL revolutionized the research areas of image, video, and information processing in the last decade. Deep CNNs have become so popular that it is incredibly cumbersome and rare to find cases in which classical approaches can still outperform CNNs on academic datasets. Despite the undeniable breakthroughs, DL methods have faced arduous challenges for deployment in practical applications. This thesis discussed many such challenges and presented scientific developments to tackle these challenges. Nonetheless, there is still considerable room for further research to accelerate the entrance of DL-based techniques into practical applications. This chapter briefly summarizes the thesis, revisits the research challenges such as trustworthiness, explainability, robustness, optimization, and fairness, points out this thesis's contribution, and draws an outline for future work.

7.1 Summary of Thesis

This thesis has been motivated by the hindrances of using DL, specifically vision models, in practical applications. After describing the challenges and laying the theoretical foundation in the first two chapters, the thesis presented an alternative to classical multilayer perceptrons (MLPs) and instead uses radial basis function networks (RBFs) as classifiers for convolutional neural networks (CNNs). RBFs have been in the scientific literature for a long time. However, they have not been optimized for CNNs before because of the complications in the optimization. This thesis offered theoretical breakthroughs to adapt RBFs for CNNs to improve the robustness and interpretability of the classification.

The interpretability of CNNs has been at the center of attention in many research works recently. However, methods such as guided-backpropagation [247] developed in this context have mostly been used to monitor the models' behavior [212]. The fourth chapter of this thesis extended the idea of understanding vision models and putting them into action for debugging CNNs and detecting adversarial

attacks with the hope of inspiring more such research in the future. This thesis's fifth and sixth chapters focused on ML and DL applications. Chapter 5 described how a problem without analytical solutions can be addressed using data-driven methods and simulation. It presented motion compensation in cone-beam computed tomography (CBCT) scans using 3D-CNNs. Chapter 6 reviewed several different applications of ML and DL in affective computing and health care, and pointed at the findings of this thesis targeting fairness in facial recognition systems and robustness of vision transformers (ViTs). The optimization process is crucial in bringing vision models to performance and affects their behavior in terms of robustness, generalization, and data requirements. Chapter 6 also offered findings in hyperparameter and model optimization gained by employing ML and DL in several applications and formulating the best practices and patterns in the automated search for best ML and DL models.

7.2 Future Research Work

Researchers' long-term vision of applying ML and DL in medical applications and autonomous driving systems is only feasible by establishing human trust in reliable and robust artificial intelligence (AI). Thus, *trustworthiness* and *reliability* are the overarching themes in the research community for practical AI applications with maximum performance and minimum negative impact [123]. It is intuitively clear that a model or algorithm used in applications involving human privacy or service access has to be reliable and trustworthy. Furthermore, trustworthiness is in demand in medical applications and autonomous driving systems involving human life and security.

Despite the demand for trustworthiness being intellectually evident, best practices for engineering trustworthy models for a specific application is an open problem and requires further investigation [237]. The importance of trustworthiness is also highly dependent on the application. For instance, robustness against spoofing or adversarial attacks is more relevant to person identification problems, while adaptation to the new vendors and image acquisition parameters emerge in medical image processing. Since the term trustworthiness is generic and includes many aspects, researchers break it down into several categories with more specific definitions where it is also possible to evaluate the performance based on acceptable common-sense explanations or mathematical metrics.

The long-term vision of AI research (reliability and trustworthiness) can be divided into smaller actionable blocks that current research addresses. Explainability, robustness, and fairness are the requirements of the trustworthy AI concepts investigated in this thesis. The remainder of this section explains this thesis's contributions to the components of trustworthy AI and opportunities for mid-term research in these areas.

Explanability: Answering the following three questions is the target of the research around explainable AI in computer vision: 1) How do models learn? 2) What do models learn? 3) How do models predict? The first question is the most complicated to answer. The research literature addressing this area is meager, but includes studies that use information theory to explain the behavior of models during optimization [229, 241]. The second and third questions are more pragmatic, relevant for practical applications, well-studied, and more connected to each other [247, 85, 212]. The features learned in vision models for decision-making are mainly evaluated using feature visualization techniques. These techniques compute the region of input images that the models look at to make a decision based on reverting the forward path or treating the models as a black box using iterative optimization. Moreover, researchers investigated the behavior of models as black boxes via post-hoc analysis to identify why the models predict a specific class. An alternative to black-box analysis is using methods such as Bayesian inference, which are more transparent by design. The contribution of this thesis to explainable AI research is revisiting radial basis function networks (RBFs) and adapting them as classifiers for CNNs by solving a few architectural hindrances. The proposed models compute a similarity metric between test and training images and derive visual clues about the decision-making process of the vision models. This research is the first to use RBFs on top of the traditional computer vision backbones. Evaluating the robustness of models using RBF classifiers against anomalies and adversarial attacks is an open question for future research.

Robustness: Researchers very quickly discovered robustness issues in computer vision models. CNN performance shows a decline in the presence of different lighting conditions and variability in the pose of the input images. The robustness problem had even more impact in the medical domain because of manual changes in image acquisition parameters, different image acquisition vendors, and frequent imaging software and hardware updates. Domain adaptation and lifelong learning in the presence of concept drift are the offsprings of the robustness and generalization issues and have tremendous exciting research potential. This thesis presented a method for data homogenization that enhances merging data from different datasets and it is practical for domain adaptation. One of the hot topics threatening the validity of CNN's for vision problems is adversarial attacks. Researchers have found that images which appear identical to the human eye can be optimized to fool vision models into making an incorrect decision. This thesis offered a method based on reverting the CNNs to visualize the models' feature response and detect adversarial attacks with very high accuracy. This research can be extended to use black-box feature visualization to detect attacks on any model and optimize the input to reduce the adversarial effects in future work. Moreover, this thesis presented a novel embedding technique for rotation invariant vision transformers to improve model robustness against input rotation.

Applying rotation invariant transformers to small datasets, especially aerial images and histology datasets, to leverage the rotation invariance as inductive bias is another promising research offspring of this thesis.

Fairness: Neural networks became very popular because of their strength in approximating arbitrary functions for classification or regression solely from data without any knowledge of the task. Although neural networks provide the opportunity to learn with minimal inductive biases, the optimization process instead follows the most efficient direction in parameters space to minimize the optimization objective (loss function) based on the existing biases in the datasets. Using these biases helped to solve the problems that researchers had not found any analytical solution to before, such as motion artifact reduction presented in Chapter 5. However, social activists rapidly discovered the drawbacks of learning from data in the social fairness aspect of face recognition (FR) systems for surveillance. The collected datasets were biased, in that the majority of the images were of white male celebrities, which was reflected in the trained models when they returned a higher accuracy for the majority race in the datasets. Studies showed that the models produce a lower accuracy for racial minorities, and that this inequality was even visible when comparing the models' accuracy for recognizing females and children with males. This thesis offered relevant research and findings about a standard method of measuring biases in FR systems and showed that racial awareness and bias are not necessarily correlated. The research concerning fairness is also quite an exciting and simultaneously challenging area. Data-driven techniques are an option for reducing biases by collecting datasets with equal populations from all sensitive features, such as race and gender. A balanced dataset is a solution to the problem faced by FR systems. However, problems such as recruitment and job application processing confront more challenges due to biases in ground truth labels based on previous hiring decisions, which opens a lot of fascinating topics for future research.

7.3 Practical Discussions

Alongside all the debates about the trustworthiness of AI models for applications where human safety and privacy are involved [123, 2, 107], AI-based models have also found their way into less critical applications [250]. However, AI projects still suffer from a very high failure rate in development and post-deployment due to problems such as concept drift. This section briefly discusses the content of this thesis related to applications and optimization.

Applications: Despite the early challenges in deploying ML and DL techniques, this thesis has shown several successful examples of ML and DL in real-world applications. Data preprocessing and cleaning before training a model is one of the most critical components of any ML pipeline. Face alignment for FR systems

or facial expression estimation is an example of data preparation before training. Although DL-based models are unrivaled for vision problems, their performance is highly dependent on the quality of the data. Determining the mutual information of the data samples and target pattern requires further research; a visual review of the datasets before model development is the key to success in applied projects [122]. Neural networks are applicable and highly recommended to approximate classical methods that are computationally expensive (such as the iterative reconstruction of computed tomography scans) or enhance their performance where analytical solutions do not exist (for example, in motion artifact reduction). This thesis offered an application of three-dimensional CNNs in reducing motion artifacts in volumetric cone-beam computed tomography (CBCT) scans with great success. This research path was extraordinarily successful and gained positive feedback and attention from clinical experts. The particular area of research is novel and ripe for further research in similar applications, such as sparseness artifact reduction, auto-segmentation, and dose calculation from CBCT scans for cancer therapy.

Optimization: ML and DL present the opportunity to explore and search among a family of neural networks to model all possible problems in computer vision. However, this large degree of freedom appears at the expense of the vast search space of parameters and potential models. Optimization concentrates on techniques that are key to neural architecture search, hyper-parameter (HP) tuning, and finding the shortest path to a stable minimum for a given dataset and model. This thesis presented the observed patterns for model and HP optimization based on ML and DL algorithms for small datasets and proposed combining supervised and unsupervised learning to enable the optimization of RBFs as classifiers for conventional CNN architectures. So far, neural architecture search has been aimed at minimizing the number of flops and latency in inference regardless of sample efficiency. Sample efficiency is another challenge in practical applications where data or labels are scarce. Hence, architectures searched for the highest sample efficiency are critical for practical applications. Other directions for future research include limiting search space and constraining optimization techniques to more explainable and robust methods that serve the purposes of trustworthy AI. The current top-performing computer vision models are derived from automatically searched architectures that target latency optimization and disregard models' explainability. There is a belief in a trade-off between accuracy and explainability in the scientific community [310]. The drop in accuracy occurs when predictive complexities are removed to make the models more explainable. However, another critical research piece refers to this trade-off as a myth [222] and encourages researchers to optimize intrinsically interpretable models to the same level of performance as black box models. Neural architecture search in the space of intrinsically interpretable and explainable models clarifies this controversial research area to show the correctness of these contradicting opinions.

Bibliography

- [1] Defense Advanced Research Projects Agency. Broad agency announcement: Explainable artificial intelligence (XAI). <https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>, 2016. Online; accessed 3 December 2022.
- [2] Muneeb Ahmed, Sarfaraz Masood, Musheer Ahmad, and Ahmed A. Abd El-Latif. Intelligent driver drowsiness detection for traffic safety based on multi CNN deep model and facial subsampling. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):19743–19752, 2021.
- [3] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.
- [4] Timur R. Almaev and Michel F. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 356–361. IEEE, 2013. ISSN: 2156-8111.
- [5] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):1–74, 2021.
- [6] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I Madai. Explainability for artificial intelligence in healthcare: a multi-disciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):310, 2020.
- [7] Adam Tan Mohd Amin, Siti Salasiah Mokri, Rozilawati Ahmad, Fuad Ismail, and Ashrani Aizzuddin Abd Rahni. Evaluation methodology for respiratory signal extraction from clinical cone-beam CT (CBCT) using data-

- driven methods. *International Journal of Integrated Engineering*, 13(5):1–8, 2021.
- [8] Mohammadreza Amirian, Markus Kächele, Günther Palm, and Friedhelm Schwenker. Support vector regression of sparse dictionary-based features for view-independent action unit intensity estimation. In *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 854–859. IEEE, IEEE, 2017.
- [9] Mohammadreza Amirian, Markus Kächele, and Friedhelm Schwenker. Using radial basis function neural networks for continuous and discrete pain estimation from bio-physiological signals. In Friedhelm Schwenker, Hazem M. Abbas, Neamat El Gayar, and Edmondo Trentin, editors, *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, volume 9896, pages 269–284. Springer, Springer, 2016.
- [10] Mohammadreza Amirian, Markus Kächele, Patrick Thiam, Viktor Kessler, and Friedhelm Schwenker. Continuous multimodal human affect estimation using echo state networks. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16*, pages 67–74. Association for Computing Machinery, 2016.
- [11] Mohammadreza Amirian, Javier A. Montoya-Zegarra, Jonathan Gruss, Yves D. Stebler, Ahmet Selman Bozkir, Marco Calandri, Friedhelm Schwenker, and Thilo Stadelmann. PrepNet: A convolutional auto-encoder to homogenize CT scans for cross-dataset medical image analysis. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–7. IEEE, 2021.
- [12] Mohammadreza Amirian, Javier A Montoya-Zegarra, Ivo Herzig, Peter Eggenberger Hotz, Lukas Lichtensteiger, Marco Morf, Alexander Züst, Pascal Paysan, Igor Peterlik, Stefan Scheib, et al. Mitigation of motion-induced artifacts in cone beam computed tomography using deep convolutional neural networks. *Medical Physics*, 50(10):6228–6242, 2023.
- [13] Mohammadreza Amirian, Katharina Rombach, Lukas Tuggener, Frank-Peter Schilling, and Thilo Stadelmann. Efficient deep cnns for cross-modal automated computer vision under time and space constraints. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*. ZHAW Zürcher Hochschule für Angewandte Wissenschaften, September 2019.
- [14] Mohammadreza Amirian and Friedhelm Schwenker. Radial basis function networks for convolutional neural networks to learn similarity distance metric and improve interpretability. *IEEE Access*, 8:123087–123097, 2020.

- [15] Mohammadreza Amirian, Friedhelm Schwenker, and Thilo Stadelmann. Trace and detect adversarial attacks on CNNs using feature response maps. In Luca Pancioni, Friedhelm Schwenker, and Edmondo Trentin, editors, *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, Lecture Notes in Computer Science, pages 346–358. Springer, Springer, 2018.
- [16] Mohammadreza Amirian, Lukas Tuggener, Ricardo Chavarriaga, Yvan Putra Satyawan, Frank-Peter Schilling, Friedhelm Schwenker, and Thilo Stadelmann. Two to trust: Automl for safe modelling and interpretable deep learning for robustness. In *Trustworthy AI - Integrating Learning, Optimization and Reasoning*, pages 268–275, Cham, 2021. Springer International Publishing.
- [17] Michael R Anderberg. *Cluster Analysis for Applications*, volume 19. Elsevier, 2014.
- [18] Anders H. Andersen and Avinash C. Kak. Simultaneous algebraic reconstruction technique (SART): A superior implementation of the ART algorithm. *Ultrasonic Imaging*, 6(1):81–94, 1984.
- [19] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 584–592. JMLR.org, 2014.
- [20] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *Computing Research Repository (CoRR)*, abs/2104.10972, 2021.
- [21] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR*, 2015.
- [22] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than CNNs? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 26831–26843. Curran Associates, Inc., 2021.
- [23] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI fairness 360: An extensible toolkit for detecting and

- mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4):1–15, 2019.
- [24] Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, 6:64270–64277, 2018.
- [25] Franz Boas and Dominik Fleischmann. CT artifacts: causes and reduction techniques. *Imaging in Medicine*, 4(2):229–240, 2012.
- [26] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry J Ackel, Urs Muller, Phil Yeres, and Karol Zieba. VisualBackProp: Efficient visualization of CNNs for autonomous driving. In *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4701–4708. IEEE, IEEE, 2018.
- [27] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *Computing Research Repository (CoRR)*, abs/1604.07316, 2016.
- [28] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 29. Curran Associates, Inc., 2016.
- [29] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [30] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *6th International Conference on Learning Representations, ICLR*. OpenReview.net, 2018.
- [31] David S. Broomhead and David Lowe. Multivariable functional interpolation and adaptive networks, complex systems, vol. 2. *Complex Systems*, 2, 1988.
- [32] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

- [33] Vanessa Buhrmester, David Münch, and Michael Arens. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, 3(4):966–989, 2021.
- [34] Thorsten M. Buzug. *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*. Springer, 2008.
- [35] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, IEEE, 2018.
- [36] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuvver M. Rao, Troy D. Kelley, Dave Braines, Murat Sensoy, Christopher J. Willis, and Prudhvi Gurram. Interpretability of deep learning models: A survey of results. In *Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*, pages 1–6. IEEE, IEEE, 2017.
- [37] Thitiporn Chanwimaluang and Guoliang Fan. An efficient blood vessel detection algorithm for retinal images using local entropy thresholding. In *Proceedings of the International Symposium on Circuits and Systems IS-CAS.*, volume 5, pages V–21–V–24. IEEE, 2003.
- [38] Gaoyu Chen, Xiang Hong, Qiaoqiao Ding, Yi Zhang, Hu Chen, Shujun Fu, Yunsong Zhao, Xiaoqun Zhang, Hui Ji, Ge Wang, Qiu Huang, and Hao Gao. AirNet: Fused analytical and iterative reconstruction with deep neural network regularization for sparse-data CT. *Medical Physics*, 47(7):2916–2930, 2020.
- [39] Gaoyu Chen, Yunsong Zhao, Qiu Huang, and Hao Gao. 4d-AirNet: a temporally-resolved CBCT slice reconstruction method synergizing analytical and iterative method with deep learning. *Physics in Medicine & Biology*, 65(17):175020, 2020.
- [40] Sheng Chen, Steve A Billings, Colin FN Cowan, and Peter M Grant. Practical identification of NARMAX models using radial basis functions. *International Journal of Control*, 52(6):1327–1350, 1990.
- [41] Gong Cheng, Ceyuan Yang, Xiwen Yao, Lei Guo, and Junwei Han. When deep learning meets metric learning: Remote sensing image scene classifica-

- tion via learning discriminative CNNs. *IEEE Transactions on Geoscience and Remote Sensing*, 56(5):2811–2821, 2018.
- [42] Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3642–3649. IEEE, 2012.
- [43] Dan Cireşan, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. A committee of neural networks for traffic sign classification. In *The 2011 International Joint Conference on Neural Networks*, pages 1918–1921. IEEE, 2011. ISSN: 2161-4407.
- [44] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: fooling deep structured visual and speech recognition models with adversarial examples. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, pages 6980–6990. Curran Associates Inc., 2017.
- [45] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2990–2999. PMLR, 2016.
- [46] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- [47] Jonathan Crabbé and Mihaela van der Schaar. Label-free explainability for unsupervised models. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 4391–4420, 2022.
- [48] Kate Crawford. Artificial intelligence’s white guy problem. <https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html>, 2022. Online; accessed 7 December 2022.
- [49] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pages 1–35. Springer, 2017.
- [50] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123. IEEE, 2019.

- [51] Mairead Daly, Alan McWilliam, Ganesh Radhakrishna, Ananya Choudhury, and Cynthia L. Eccles. Radiotherapy respiratory motion management in hepatobiliary and pancreatic malignancies: a systematic review of patient factors influencing effectiveness of motion reduction with abdominal compression. *Acta Oncologica*, 61(7):833–841, 2022. PMID: 35611555.
- [52] Michael E. Dawson, Anne M. Schell, and Diane L. Filion. The electrodermal system. In John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson, editors, *Handbook of Psychophysiology*, pages 217–243. Cambridge University Press, 4 edition, 2017.
- [53] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.
- [54] Taye Girma Debelee, Mohammadreza Amirian, Achim Ibenthal, Günther Palm, and Friedhelm Schwenker. Classification of mammograms using convolutional neural network based feature extraction. In *International Conference on Information and Communication Technology for Development for Africa*, pages 89–98. Springer, September 2017.
- [55] Taye Girma Debelee, Abrham Gebreselasie, Friedhelm Schwenker, Mohammadreza Amirian, and Dereje Yohannes. Classification of mammograms using texture and cnn based extracted features. *Journal of Biomimetics, Biomaterials and Biomedical Engineering*, 42:79–97, 8 2019.
- [56] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. COVAREP — a collaborative voice analysis repository for speech technologies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964. IEEE, 2014. ISSN: 2379-190X.
- [57] Mehdi Dehghan and Vahid Mohammadi. The numerical solution of fokker–planck equation with radial basis functions (RBFs) based on the meshless technique of kansa’s approach and galerkin method. *Engineering Analysis with Boundary Elements*, 47:38–63, 2014.
- [58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.
- [59] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694. IEEE, 2019.

- [60] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.
- [61] Owen Dillon, Paul J Keall, Chun-Chien Shieh, and Ricky T O’Brien. Evaluating reconstruction algorithms for respiratory motion guided acquisition. *Physics in Medicine & Biology*, 65(17):175009, 2020.
- [62] Simone Disabato and Manuel Roveri. Learning convolutional neural networks in presence of concept drift. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, IEEE, 2019.
- [63] Weisheng Dong, Peiyao Wang, Wotao Yin, Guangming Shi, Fangfang Wu, and Xiaotong Lu. Denoising prior driven deep neural network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2305–2318, 2018.
- [64] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215. IEEE, IEEE, 2018.
- [65] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.
- [66] Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. A systematic review of robustness in deep learning for computer vision: Mind the gap? *Computing Research Repository (CoRR)*, abs/2112.00639, 2021.
- [67] K.-L. Du, K.K.M. Cheng, and M.N.S. Swamy. A fast neural beamformer for antenna arrays. In *IEEE International Conference on Communications. Conference Proceedings. ICC 2002 (Cat. No.02CH37333)*, volume 1, pages 139–144. IEEE, IEEE, 2002.
- [68] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *Computing Research Repository (CoRR)*, abs/1605.07277, 2016.

- [69] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.
- [70] Ulrik V Elström, Ludvig P Muren, Jörgen BB Petersen, and Cai Grau. Evaluation of image quality for different kV cone-beam CT acquisition and reconstruction methods in the head and neck region. *Acta Oncologica*, 50(6):908–917, 2011.
- [71] Julien Erath, Tim Vöth, Joscha Maier, and Marc Kachelrieß. Forward and cross-scatter estimation in dual source CT using the deep scatter estimation (DSE). In Hilde Bosmans, Guang-Hong Chen, and Taly Gilat Schmidt, editors, *Medical Imaging 2019: Physics of Medical Imaging*, volume 10948, page 24. International Society for Optics and Photonics, SPIE, 2019.
- [72] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical Report 3, University of Montreal, 2009. Online; accessed 7 December 2022.
- [73] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning $SO(3)$ equivariant representations with spherical CNNs. *International Journal of Computer Vision*, 128(3):588–600, 2019.
- [74] Reuben Feinman, Ryan R. Curtin, Saurabh Shintre, and Andrew B. Gardner. Detecting adversarial samples from artifacts. *Computing Research Repository (CoRR)*, abs/1703.00410, 2017.
- [75] Lee A. Feldkamp, Lloyd C. Davis, and James W. Kress. Practical cone-beam algorithm. *Journal of the Optical Society of America A*, 1(6):612–619, 1984.
- [76] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 28. Curran Associates, Inc., 2015.
- [77] Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. Using meta-learning to initialize bayesian optimization of hyperparameters. In *Proceedings of the International Conference on Meta-learning and Algorithm Selection - Volume 1201*, MLAS’14, pages 3–10. Citeseer, CEUR-WS.org, 2014.
- [78] Roger Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, Ltd, 2000.

- [79] Richard Franke. A critical comparison of some methods for interpolation of scattered data. Technical report, NAVAL POSTGRADUATE SCHOOL MONTEREY CA, 1979.
- [80] Dennis Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.
- [81] Yarín Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org, 2016.
- [82] Stephen J. Gardner, Weihua Mao, Chang Liu, Ibrahim Aref, Mohamed Elshaikh, Joon K. Lee, Deepak Pradhan, Benjamin Movsas, Indrin J. Chetty, and Farzan Siddiqui. Improvements in CBCT image quality using a novel iterative reconstruction algorithm: A clinical evaluation. *Advances in Radiation Oncology*, 4(2):390–400, 2019.
- [83] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings fo the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423. IEEE, 2016.
- [84] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *Computing Research Repository (CoRR)*, abs/1811.12231, 2018.
- [85] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89. IEEE, 2018.
- [86] Lars Gjestebj, Bruno De Man, Yannan Jin, Harald Paganetti, Joost Verburg, Drosoula Giantsoudi, and Ge Wang. Metal artifact reduction in CT: Where are we after four decades? *IEEE Access*, 4:5826–5849, 2016.
- [87] Stefan Glüge, Mohammadreza Amirian, Dandolo Flumini, and Thilo Stadelmann. How (not) to measure bias in face recognition networks. In Frank-Peter Schilling and Thilo Stadelmann, editors, *Proceedings of the IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, Lecture Notes in Computer Science, pages 125–137. Springer, Springer International Publishing, 2020.

- [88] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [89] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceeding of International Conference on Learning Representations*. arXiv, 2015.
- [90] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine*, 38(3):50–57, 2016.
- [91] Katharine Grant and Rainer Raupach. Safire: Sinogram affirmed iterative reconstruction. https://cdn0.scrvt.com/39b415fb07de4d9656c7b516d8e2d907/1800000000306520/d80046026fd1/ct_SAFIRE_White_Paper_1800000000306520.pdf, 2012. Online; accessed 6 December 2022.
- [92] Mara Graziani, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, John O. Prior, Lode Lauwaert, Wessel Reijers, Adrien Depeursinge, Vincent Andrearczyk, and Henning Müller. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artificial Intelligence Review*, pages 1–32, 2022.
- [93] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *Computing Research Repository (CoRR)*, abs/1702.06280, 2017.
- [94] Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boullé, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michèle Sebag, Alexander Statnikov, Wei-Wei Tu, and Evelyne Viegas. Analysis of the AutoML challenge series 2015–2018, 2017.
- [95] Yo Seob Han, Jaejun Yoo, and Jong Chul Ye. Deep residual learning for compressed sensing CT reconstruction via persistent homology analysis. *Computing Research Repository (CoRR)*, abs/1611.06391, 2016.
- [96] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.
- [97] Mattias P. Heinrich, Mark Jenkinson, Sir Michael Brady, and Julia A. Schnabel. MRF-based deformable registration and ventilation estimation of lung CT. *IEEE Transactions on Medical Imaging*, 32(7):1239–1248, 2013.

- [98] Ivo Herzig, Pascal Paysan, Stefan Scheib, Alexander Züst, Frank-Peter Schilling, Javier Montoya, Mohammadreza Amirian, Thilo Stadelmann, Peter Eggenberger Hotz, Rudolf Marcel Fuchsli, et al. Deep learning-based simultaneous multi-phase deformable image registration of sparse 4d-cbct. *Medical Physics*, 49(6):e325–e326, 2022.
- [99] Kashmir Hill. Wrongfully accused by an algorithm. <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html>, 2020. Online; accessed 6 December 2022.
- [100] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [101] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In Aasa Feragen, Marcello Pelillo, and Marco Loog, editors, *International Workshop on Similarity-Based Pattern Recognition*, volume 9370, pages 84–92. Springer, Springer, 2015.
- [102] Lukas Hollenstein, Lukas Lichtensteiger, Thilo Stadelmann, Mohammadreza Amirian, Lukas Budde, Jürg Meierhofer, Rudolf M Fuchsli, and Thomas Friedli. Unsupervised learning and simulation for complexity management in business operations. *Applied data science: lessons learned for the data-driven business*, pages 313–331, 2019.
- [103] Godfrey Hounsfield. Method of and apparatus for examining a body by radiation such as x or gamma radiation. Technical report, Originating Research Org. not identified, 1975.
- [104] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *Computing Research Repository (CoRR)*, abs/1704.04861, 2017.
- [105] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141. IEEE, 2018.
- [106] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Deep transfer metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 325–333. IEEE, 2015.
- [107] Kai Huang, Ximeng Liu, Shaojing Fu, Deke Guo, and Ming Xu. A lightweight privacy-preserving CNN feature extraction framework for mobile sensing. *IEEE Transactions on Dependable and Secure Computing*, 18(3):1441–1455, 2019.

- [108] Xiaojie Huang, Junjie Shan, and Vivek Vaidya. Lung nodule detection in CT using 3d convolutional neural networks. In *Proceedings of the IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 379–383. IEEE, 2017.
- [109] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020.
- [110] Yukun Huang, Yongcai Guo, and Chao Gao. Efficient parallel inflated 3d convolution architecture for action recognition. *IEEE Access*, 8:45753–45765, 2020.
- [111] Michael J Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 4533–4543. PMLR, 2021.
- [112] Herbert Jaeger. The “echo state” approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148(34):13, 2001.
- [113] David A Jaffray, Jeffrey H Siewerdsen, John W Wong, and Alvaro A Martinez. Flat-panel cone-beam computed tomography for image-guided radiation therapy. *International Journal of Radiation Oncology*Biography*Physics*, 53(5):1337–1349, 2002.
- [114] Talia Jarema and Trent Aland. Using the iterative kV CBCT reconstruction on the varian halcyon linear accelerator for radiation therapy-planning CT datasets: A feasibility study. *International Journal of Radiation Oncology*Biography*Physics*, 105(1):E719–E720, 2019.
- [115] Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.
- [116] Markus Kächele, Mohammadreza Amirian, Patrick Thiam, Philipp Werner, Steffen Walter, Günther Palm, and Friedhelm Schwenker. Adaptive confidence learning for the personalization of pain intensity estimation systems. *Evolving Systems*, 8(1):71–83, 2017.

- [117] Markus Kächele, Patrick Thiam, Mohammadreza Amirian, Friedhelm Schwenker, and Günther Palm. Methods for person-centered continuous pain intensity assessment from bio-physiological channels. *IEEE Journal of Selected Topics in Signal Processing*, 10(5):854–864, 2016.
- [118] Markus Kächele, Patrick Thiam, Mohammadreza Amirian, Philipp Werner, Steffen Walter, Friedhelm Schwenker, and Günther Palm. Multimodal data fusion for person-independent, continuous estimation of pain intensity. In Lazaros Iliadis and Chrisina Jayne, editors, *Proceedings of the International Conference on Engineering Applications of Neural Networks (EANN)*, Communications in Computer and Information Science, pages 275–285. Springer, 2015.
- [119] Stefan Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bulletin International de l' Académie Polonaise des Sciences et des Lettres*, 35:355–357, 1937.
- [120] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: Analysis and improvement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4441–4449. IEEE, 2018.
- [121] N. Karimi, S. Kazem, D. Ahmadian, H. Adibi, and L.V. Ballestra. On a generalized gaussian radial basis function: Analysis and applications. *Engineering Analysis with Boundary Elements*, 112:46–57, 2020.
- [122] Andrej Karpathy. A recipe for training neural networks. <http://karpathy.github.io/2019/04/25/recipe/>, 2019. Online; accessed 7 December 2022.
- [123] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durrezi. Trustworthy artificial intelligence: A review. *ACM Computing Surveys (CSUR)*, 55(2):39:1–39:38, 2022.
- [124] Kamran Kazemi, Mohammadreza Amirian, and Mohammad Javad Dehghani. The s-transform using a new window to improve frequency and time resolutions. *Signal, image and Video processing*, 8(3):533–541, 2014.
- [125] Benjamin Keck, Hannes G. Hofmann, Holger Scherl, Markus Kowarschik, and Joachim Hornegger. High resolution iterative CT reconstruction using graphics hardware. In *Proceedings of the IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, pages 4035–4040, Oct 2009. ISSN: 1082-3654.
- [126] Roland Kehrein. The prosody of authentic emotions. In *Proceedings of the International Conference on Speech Prosody*, pages 423–426, 2002.

- [127] Viktor Kessler, Patrick Thiam, Mohammadreza Amirian, and Friedhelm Schwenker. Multimodal fusion including camera photoplethysmography for pain recognition. In *2017 International Conference on Companion Technology (ICCT)*, pages 1–4, September 2017.
- [128] Viktor Kessler, Patrick Thiam, Mohammadreza Amirian, and Friedhelm Schwenker. Pain recognition with camera photoplethysmography. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–5, November 2017.
- [129] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.
- [130] Buomsoo Kim, Jinsoo Park, and Jihae Suh. Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, 134:113302, 2020.
- [131] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9004–9012. IEEE, 2019.
- [132] Donghwan Kim, Sathish Ramani, and Jeffrey A. Fessler. Combining ordered subsets and momentum for accelerated X-ray CT image reconstruction. *IEEE Transactions on Medical Imaging*, 34(1):167–178, Jan 2015.
- [133] Hayeon Kim, M. Saiful Huq, Ron Lalonde, Christopher J. Houser, Sushil Beriwal, and Dwight E. Heron. Early clinical experience with varian halcyon v2 linear accelerator: Dual-isocenter IMRT planning and delivery with portal dosimetry for gynecological cancer treatments. *Journal of Applied Clinical Medical Physics*, 20(11):111–120, 2019.
- [134] Daniel Kindsvater, Sascha Meudt, and Friedhelm Schwenker. Fusion architectures for multimodal cognitive load recognition. In Friedhelm Schwenker and Stefan Scherer, editors, *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction (MPRSS)*, volume 10183, pages 36–47. Springer, 2016.
- [135] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.
- [136] Andreas Kofler, Markus Haltmeier, Christoph Kolbitsch, Marc Kachelrieß, and Marc Dewey. A U-Nets cascade for sparse view computed tomography. In Florian Knoll, Andreas Maier, and Daniel Rueckert, editors, *Proceedings*

- of the First International Workshop on Machine Learning for Medical Image Reconstruction (MLMIR) Held in Conjunction with MICCAI*, Lecture Notes in Computer Science, pages 91–99. Springer, 2018.
- [137] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *ICCV Workshops*, pages 554–561, 2013.
- [138] Arvind Krishna. IBM CEO’s letter to congress on racial justice reform. <https://www.ibm.com/policy/facial-recognition-sunset-racial-justice-reforms/>, 2020. Online; accessed 6 December 2022.
- [139] Alex Krizhevsky. Learning multiple layers of features from tiny images. <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>, 2009. Online; accessed 3 December 2022.
- [140] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, volume 25, pages 1097–1105, 2012.
- [141] Miroslav Kubat. Decision trees can initialize radial-basis function networks. *IEEE Transactions on Neural Networks*, 9(5):813–821, 1998.
- [142] Gaurav Kumar and Pradeep Kumar Bhatia. A detailed review of feature extraction in image processing systems. In *Proceedings of the Fourth International Conference on Advanced Computing & Communication Technologies*, pages 5–12. IEEE, 2014. ISSN: 2327-0659.
- [143] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7):1550–1560, 2017.
- [144] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Proceeding of International Conference on Learning Representations*. arXiv, 2016.
- [145] Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321–348, 2019.
- [146] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

- [147] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural computation*, 1(4):541–551, December 1989.
- [148] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [149] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-CNN for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.
- [150] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(185):1–52, 2018.
- [151] Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5775–5783. IEEE, 2017.
- [152] Yang Li, Wei-Gang Cui, Hui Huang, Yu-Zhu Guo, Ke Li, and Tao Tan. Epileptic seizure detection in EEG signals using sparse multiscale radial basis function networks and the fisher vector approach. *Knowledge-Based Systems*, 164:96–106, 2019.
- [153] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, and Xiaofeng Wang. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Transactions on Dependable and Secure Computing*, 18(1):72–85, 2018.
- [154] Haofu Liao, Wei-An Lin, S. Kevin Zhou, and Jiebo Luo. ADN: Artifact disentanglement network for unsupervised metal artifact reduction. *IEEE Transactions on Medical Imaging*, 39(3):634–643, 2020.
- [155] Yi Liao, Shu-Cherng Fang, and Henry L.W. Nuttle. Relaxed conditions for radial-basis function networks to be universal approximators. *Neural Networks*, 16(7):1019–1028, 2003.
- [156] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 32. Curran Associates, Inc., 2019.

- [157] Wei-An Lin, Haofu Liao, Cheng Peng, Xiaohang Sun, Jingdan Zhang, Jiebo Luo, Rama Chellappa, and Shaohua Kevin Zhou. DuDoNet: Dual domain network for CT metal artifact reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10504–10513. IEEE, 2019.
- [158] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *Computing Research Repository (CoRR)*, abs/1907.11692, 2019.
- [159] Zhengying Liu, Zhen Xu, Shangeth Rajaa, Meysam Madadi, Julio C. S. Jacques Junior, Sergio Escalera, Adrien Pavao, Sebastien Treguer, Wei-Wei Tu, and Isabelle Guyon. CNN architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017.
- [160] Zhengying Liu, Zhen Xu, Shangeth Rajaa, Meysam Madadi, Julio C. S. Jacques Junior, Sergio Escalera, Adrien Pavao, Sebastien Treguer, Wei-Wei Tu, and Isabelle Guyon. Towards automated deep learning: Analysis of the AutoDL challenge series 2019. In *Proceedings of the NeurIPS Competition and Demonstration Track*, pages 242–252. PMLR, PMLR, 2020.
- [161] Steve Lohr. Facial recognition is accurate, if you’re a white guy. <https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html>, 2018. Online; accessed 6 December 2022.
- [162] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, pages 1–18, 2019.
- [163] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE, IEEE, 1999.
- [164] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [165] Jiajun Lu, Theerasit Issaranon, and David Forsyth. SafetyNet: Detecting and rejecting adversarial examples robustly. In *IEEE International Conference on Computer Vision (ICCV)*, pages 446–454. IEEE, 2017.
- [166] Qing Lyu, Hongming Shan, Yibin Xie, Alan C. Kwan, Yuka Otaki, Keiichiro Kuronuma, Debiao Li, and Ge Wang. Cine cardiac MRI motion artifact

- reduction using a recurrent neural network. *IEEE Transactions on Medical Imaging*, 40(8):2170–2181, 2021.
- [167] Yuanyuan Lyu, Jiajun Fu, Cheng Peng, and S. Kevin Zhou. U-DuDoNet: Unpaired dual-domain network for CT metal artifact reduction. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science, pages 296–306. Springer International Publishing, 2021.
- [168] Yuanyuan Lyu, Wei-An Lin, Haofu Liao, Jingjing Lu, and S. Kevin Zhou. Encoding metal mask projection for metal artifact reduction in computed tomography. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 147–157. Springer, Springer-Verlag, 2020.
- [169] Ryan Mac. Facebook apologizes after a.i. puts ‘primates’ label on video of black men. <https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html>, 2021. Online; accessed 6 December 2022.
- [170] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018.
- [171] Andreas K. Maier, Christopher Syben, Bernhard Stimpel, Tobias Würfl, Mathis Hoffmann, Frank Schebesch, Weilin Fu, Leonid Mill, Lasse Kling, and Silke Christiansen. Learning with known operators reduces maximum error bounds. *Nature Machine Intelligence*, 1(8):373–380, 2019.
- [172] Joscha Maier, Stefan Sawall, Marc Kachelrieß, and Yannick Berker. Deep scatter estimation (DSE): feasibility of using a deep convolutional neural network for real-time x-ray scatter prediction in cone-beam CT. *SPIE Medical Imaging*, 10573:56, 2018.
- [173] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *Computing Research Repository (CoRR)*, abs/1306.5151, 2013.
- [174] Weihua Mao, Chang Liu, Stephen J. Gardner, Farzan Siddiqui, Karen C. Snyder, Akila Kumarasiri, Bo Zhao, Joshua Kim, Ning Winston Wen, Benjamin Movsas, and Indrin J. Chetty. Evaluation and clinical application of a commercially available iterative reconstruction algorithm for CBCT-based IGRT. *Technology in Cancer Research & Treatment*, 18, 2019. PMID: 30803367.

- [175] Ričards Marcinkevičs and Julia E Vogt. Interpretable models for granger causality using self-explaining neural networks. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.
- [176] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5058–5067. IEEE, 2017.
- [177] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 471–478. IEEE, 2018.
- [178] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *Proceedings of the 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 471–478. IEEE, IEEE, 2018.
- [179] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.
- [180] Benjamin Bruno Meier, Ismail Elezi, Mohammadreza Amirian, Oliver Dürr, and Thilo Stadelmann. Learning neural models for end-to-end clustering. In *Artificial Neural Networks in Pattern Recognition*, pages 126–138, Cham, 2018. Springer International Publishing.
- [181] Dongyu Meng and Hao Chen. MagNet: A two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 135–147. ACM, 2017.
- [182] Meng Joo Er, Shiqian Wu, Juwei Lu, and Hock Lye Toh. Face recognition with radial basis function (RBF) neural networks. *IEEE Transactions on Neural Networks*, 13(3):697–710, 2002.
- [183] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *5th International Conference on Learning Representations, ICLR*. OpenReview.net, 2017.
- [184] Jan Hendrik Metzen, Mumtaz Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2774–2783. IEEE, 2017.
- [185] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

- [186] John Moody and Christian J Darken. Fast learning in networks of locally-tuned processing units. *Neural computation*, 1(2):281–294, 1989.
- [187] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 86–94. IEEE, 2017.
- [188] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582. IEEE, 2016.
- [189] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. *Comput. Speech Lang.*, 60:2616–2620, 2017.
- [190] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.
- [191] Vassilis C. Nicodemou, Iason Oikonomidis, and Antonis Argyros. Single-shot 3d hand pose estimation using radial basis function networks trained on synthetic data. *Pattern Analysis and Applications*, 23(1):415–428, 2020.
- [192] Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. In *Proceedings of the IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67, 1999.
- [193] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.
- [194] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.
- [195] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.
- [196] Randal S. Olson, Ryan J. Urbanowicz, Peter C. Andrews, Nicole A. Lavender, La Creis Kidd, and Jason H. Moore. Automating biomedical data science through tree-based pipeline optimization. In Giovanni Squillero and Paolo Burelli, editors, *Proceedings of the European Conference on the Applications of Evolutionary Computation*, Lecture Notes in Computer Science, pages 123–137. Springer, Springer, 2016.

- [197] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.
- [198] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *Computing Research Repository (CoRR)*, abs/1605.07277, 2016.
- [199] Hyoung Suk Park, Sung Min Lee, Hwa Pyung Kim, Jin Keun Seo, and Yong Eun Chung. CT sinogram-consistency learning for metal-induced beam hardening correction. *Medical Physics*, 45(12):5376–5384, 2018.
- [200] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012. ISSN: 1063-6919.
- [201] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: an imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 8024–8035. Curran Associates, Inc., 2019.
- [202] Pascal Paysan, Marcus Brehm, Adam Wang, Dieter Seghers, and Josh Star-Lack. Iterative image reconstruction in image-guided radiation therapy. <https://patentscope.wipo.int/search/en/detail.jsf?docId=W02017066248>, November 16 2021. US Patent 11,173,324.
- [203] Pascal Paysan, Igor Peterlík, Toon Roggen, Liangjia Zhu, Claas Wessels, Jan Schreier, Martin Buchacek, and Stefan Scheib. Deep learning methods for image guidance in radiation therapy. In Frank-Peter Schilling and Thilo Stadelmann, editors, *Proceedings of the 9th IAPR Artificial Neural Networks in Pattern Recognition Workshop (ANNPR)*, volume 12294 of *Lecture Notes in Computer Science*, pages 3–22. Springer, 2020.
- [204] Pascal Paysan, Adam Strzelecki, Felipe Arrate, Peter Munro, and Stefan G. Scheib. Convolutional network based motion artifact reduction in cone-beam CT. In *AAPM annual meeting 2019, e-Poster*, volume 46, pages E340–E341. WILEY , USA, 2019.
- [205] Stephanie T. H. Peeters, Femke Vaassen, Colien Hazelaar, Ana Vaniqui, Eva Rousch, Debby Tissen, Esther Van Enkevort, Michiel De Wolf, Michel C. Öllers, Wouter van Elmpt, Karolien Verhoeven, Judith G. M. Van Loon,

- Bettine A. Vosse, Dirk K. M. De Ruysscher, and Gloria Vilches-Freixas. Visually guided inspiration breath-hold facilitated with nasal high flow therapy in locally advanced lung cancer. *Acta Oncologica*, 60(5):567–574, 2021. PMID: 33295823.
- [206] Roger Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.
- [207] Igor Peterlik, Adam Strzelecki, Mathias Lehmann, Philippe Messmer, Peter Munro, Pascal Paysan, Mathieu Plamondon, and Dieter Seghers. Reducing residual-motion artifacts in iterative 3d CBCT reconstruction in image-guided radiation therapy. *Medical Physics*, 48(10):6497–6507, 2021.
- [208] Tomaso Poggio and Federico Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.
- [209] Jonathan Posner, James A. Russell, and Bradley S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3), 2005.
- [210] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *Computing Research Repository (CoRR)*, abs/1710.05941, 2017.
- [211] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop*. JMLR, 2017.
- [212] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology: Artificial Intelligence*, 2(3):e190043, 2020.
- [213] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. ImageNet-21K pretraining for the masses. *Computing Research Repository (CoRR)*, abs/2104.10972, 2021.
- [214] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30, 2015.
- [215] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and

- affective interactions. In *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013.
- [216] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.
- [217] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: Too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–10. IEEE, 2020.
- [218] David W. Romero and Jean-Baptiste Cordonnier. Group equivariant stand-alone self-attention for vision. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021.
- [219] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015.
- [220] Avi Rosenfeld and Ariella Richardson. Why, who, what, when and how about explainability in human-agent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20*, pages 2161–2164. International Foundation for Autonomous Agents and Multiagent Systems, 2020.
- [221] Ron Rubinstein, Alfred M Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.
- [222] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- [223] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [224] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

- [225] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520. IEEE, 2018.
- [226] Fernando Fernandes dos Santos, Pedro Foletto Pimenta, Caio Lunardi, Lucas Draghetti, Luigi Carro, David Kaeli, and Paolo Rech. Analyzing and increasing the reliability of convolutional neural networks on GPUs. *IEEE Transactions on Reliability*, 68(2):663–677, 2018.
- [227] D. R. Sarvamangala and Raghavendra V. Kulkarni. Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, 15(1):1–22, 2021.
- [228] Ikuro Sato, Hiroki Nishimura, and Kensuke Yokoi. APAC: Augmented Pattern classification with neural networks. *Computing Research Repository (CoRR)*, abs/1505.03229, 2015.
- [229] Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.
- [230] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.
- [231] Alena-Kathrin Schnurr, Khanlian Chung, Tom Russ, Lothar R. Schad, and Frank G. Zöllner. Simulation-based deep artifact correction with convolutional neural networks for limited angle artifacts. *Zeitschrift für Medizinische Physik*, 29(2):150–161, 2019.
- [232] Ralf Schulze, Ulrich Heil, D Groß, Dan Dominik Bruellmann, Egor Dranishnikov, Ulrich Schwanecke, and Elmar Schoemer. Artefacts in CBCT: a review. *Dentomaxillofacial Radiology*, 40(5):265–273, 2011.
- [233] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [234] Friedhelm Schwenker and Christian Dietrich. Initialisation of radial basis function networks using classification trees. *Neural Network World*, 10(3):473–482, 2000.
- [235] Friedhelm Schwenker, Hans A. Kestler, and Günther Palm. Three learning phases for radial-basis-function networks. *Neural Networks*, 14(4):439–458, 2001.

- [236] Friedhelm. Schwenker, Hans A. Kestler, Günther Palm, and M. Höher. Similarities of LVQ and RBF learning—a survey of learning rules and the application to the classification of signals from high-resolution electrocardiography. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 646–651. IEEE, IEEE, 1994.
- [237] Alex Serban, Koen van der Blom, Holger Hoos, and Joost Visser. Practices for engineering trustworthy machine learning applications. In *Proceedings of the IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN)*, pages 97–100. IEEE, IEEE, 2021.
- [238] Ignacio Serna, Aythami Morales, Julian Fierrez, and Nick Obradovich. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence*, 305:103682, 2022.
- [239] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, pages 464–468. Association for Computational Linguistics, 2018.
- [240] Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.
- [241] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *Computing Research Repository (CoRR)*, abs/1703.00810, 2017.
- [242] Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR)*., volume 1, pages 958–963. IEEE Comput. Soc, 2003.
- [243] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *5th International Conference on Learning Representations, ICLR*. OpenReview.net, 2015.
- [244] John R Smith. Ibm research releases ‘diversity in faces’ dataset to advance study of fairness in facial recognition systems. <https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/>, 2019. Online; accessed 7 December 2022.
- [245] Eduardo Soares and Plamen Angelov. A large dataset of real patients CT scans for COVID-19 identification, 2020. Type: dataset.

- [246] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012. IEEE, 2016.
- [247] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *Proceedings of the 3rd International Conference on Learning Representations, {ICLR}*. OpenReview.net, 2015.
- [248] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research (JMLR)*, 15(56):1929–1958, 2014.
- [249] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *Computing Research Repository (CoRR)*, abs/1505.00387, 2015.
- [250] Thilo Stadelmann, Mohammadreza Amirian, Ismail Arabaci, Marek Arnold, Gilbert François Duivesteijn, Ismail Elezi, Melanie Geiger, Stefan Lörwald, Benjamin Bruno Meier, Katharina Rombach, and Lukas Tuggener. Deep learning in the wild. In Luca Pancioni, Friedhelm Schwenker, and Edmondo Trentin, editors, *Proceedings of the IAPR Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR)*, Lecture Notes in Computer Science, pages 17–38. Springer, Springer, 2018.
- [251] Thilo Stadelmann, Vasily Tolkachev, Beate Sick, Jan Stampfli, and Oliver Dürr. Beyond ImageNet: Deep learning in industrial practice. In Martin Braschler, Thilo Stadelmann, and Kurt Stockinger, editors, *Applied Data Science*, pages 205–232. Springer, 2019.
- [252] Rainer Storn and Kenneth Price. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.
- [253] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7242–7252. IEEE, 2021.
- [254] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Computing Research Repository (CoRR)*, abs/2104.09864, 2021.

- [255] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.
- [256] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 769–777. IEEE, 2015.
- [257] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Proceedings of the Advances in Neural Information Processing Systems NIPS*, volume 27. Curran Associates, Inc., 2014.
- [258] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2892–2900. IEEE, 2015.
- [259] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Sparsifying neural network connections for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4856–4864. IEEE, 2016.
- [260] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. IEEE, 2015.
- [261] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. IEEE, 2016.
- [262] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceeding of International Conference on Learning Representations*. arXiv, 2014.
- [263] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. MnasNet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2815–2823. IEEE, 2019.

- [264] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.
- [265] Patrick Thiam, Viktor Kessler, Mohammadreza Amirian, Peter Bellmann, Georg Layher, Yan Zhang, Maria Velana, Sascha Gruss, Steffen Walter, Harald C. Traue, Daniel Schork, Jonghwa Kim, Elisabeth André, Heiko Neumann, and Friedhelm Schwenker. Multi-modal pain intensity recognition based on the senseemotion database. *IEEE Transactions on Affective Computing*, 12(3):743–760, 2021.
- [266] Patrick Thiam, Hans Kestler, and Friedhelm Schwenker. Multimodal deep denoising convolutional autoencoders for pain intensity classification based on physiological signals. In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, pages 289–296. SCITEPRESS - Science and Technology Publications, 2020.
- [267] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- [268] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2021.
- [269] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 10347–10357. PMLR, 2021.
- [270] Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 5866–5876. Curran Associates Inc., 2019.
- [271] Lukas Tuggener, Mohammadreza Amirian, Fernando Benites, Pius von Däniken, Prakhar Gupta, Frank-Peter Schilling, and Thilo Stadelmann. Design patterns for resource-constrained automated deep-learning methods. *AI*, 1(4):510–538, 2020.
- [272] Lukas Tuggener, Mohammadreza Amirian, Katharina Rombach, Stefan Lörwald, Anastasia Varlet, Christian Westermann, and Thilo Stadelmann.

- Automated machine learning in practice: State of the art and recent results. In *Proceedings of the 6th Swiss Conference on Data Science (SDS)*, pages 31–36. IEEE, IEEE, 2019.
- [273] Heang K. Tuy. An inversion formula for cone-beam reconstruction. *SIAM Journal on Applied Mathematics*, 43(3):546–552, 1983.
- [274] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *Computing Research Repository (CoRR)*, abs/1607.08022, 2016.
- [275] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2016.
- [276] Michel F. Valstar, Enrique Sanchez-Lozano, Jeffrey F. Cohn, Laszlo A. Jeni, Jeffrey M. Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. FERA 2017 - addressing head pose in the third facial expression recognition and analysis challenge. In *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 839–847. IEEE, IEEE, 2017.
- [277] Laurens Van der Maaten and Geoffrey Hinton. Visualizing non-metric similarities in multiple maps. *Machine Learning*, 87(1):33–55, 2012.
- [278] Joaquin Vanschoren. Meta-learning. In Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, editors, *Automated Machine Learning: Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, pages 35–61. Springer, 2019.
- [279] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 30. Curran Associates, Inc., 2017.
- [280] Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 163–172, 2012.

- [281] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the IEEE/ACM International Workshop on Software Fairness*, pages 1–7. IEEE, ACM, 2018.
- [282] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C. Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O. Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *Proceedings of the IEEE International Conference on Cybernetics (CYBCO)*, pages 128–131. IEEE, 2013.
- [283] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using DropConnect. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1058–1066. PMLR, 2013.
- [284] Hai Wang, Bongnam Kang, and Daijin Kim. PFW: A face database in the wild for studying face identification and verification in uncontrolled environment. In *Proceedings of the 2nd IAPR Asian Conference on Pattern Recognition*, pages 356–360, 2008. ISSN: 0730-6512.
- [285] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274. IEEE, 2018.
- [286] Hong Wang, Yuexiang Li, Haimiao Zhang, Jiawei Chen, Kai Ma, Deyu Meng, and Yefeng Zheng. InDuDoNet: An interpretable dual domain network for CT metal artifact reduction. In *Proceedings of the 24th International Conference Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 107–118. Springer, Springer-Verlag, 2021.
- [287] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2612–2620. IEEE, 2017.
- [288] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.
- [289] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 692–702. IEEE, 2019.

- [290] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K. Nandi. Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16(5):1243–1267, 2020.
- [291] Shui-Hua Wang, Chaosheng Tang, Junding Sun, Jingyuan Yang, Chenxi Huang, Preetha Phillips, and Yu-Dong Zhang. Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling. *Frontiers in Neuroscience*, 12:818, 2018.
- [292] Tao Wang, Zexin Lu, Ziyuan Yang, Wenjun Xia, Mingzheng Hou, Huaiqiang Sun, Yan Liu, Hu Chen, Jiliu Zhou, and Yi Zhang. IDOL-net: An interactive dual-domain parallel network for CT metal artifact reduction. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 6(8):874–885, 2022.
- [293] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [294] Hayate Washio, Shingo Ohira, Yoshinori Funama, Masahiro Morimoto, Kentaro Wada, Masashi Yagi, Hiroaki Shimamoto, Yuhei Koike, Yoshihiro Ueda, Tsukasa Karino, Shoki Inui, Yuya Nitta, Masayoshi Miyazaki, and Teruki Teshima. Metal artifact reduction using iterative CBCT reconstruction algorithm for head and neck radiation therapy: A phantom and clinical study. *European Journal of Radiology*, 132, 2020.
- [295] Samuel Wehrli, Corinna Hertweck, Mohammadreza Amirian, Stefan Glüge, and Thilo Stadelmann. Bias, awareness, and ignorance in deep-learning-based face recognition. *AI and Ethics*, 2(3):509–522, 2022.
- [296] Maurice Weiler, Fred A. Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant CNNs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 849–858. IEEE, 2018.
- [297] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD birds 200. Report or Paper CNS-TR-2010-001, California Institute of Technology, 2010.
- [298] Ruben Wiersma, Elmar Eisemann, and Klaus Hildebrandt. CNNs on surfaces using rotation-equivariant features. *ACM Transactions on Graphics*, 39(4):92–1, 2020.
- [299] Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

- [300] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 748–756. IEEE, IEEE, 2018.
- [301] Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. Why we should have seen that coming: comments on microsoft’s tay “experiment,” and wider implications. *The ORBIT Journal*, 1(2):1–12, 2017.
- [302] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Proceedings of the European conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, pages 3–19. Springer, 2018.
- [303] Tobias Würfl, Mathis Hoffmann, Vincent Christlein, Katharina Breininger, Yixin Huang, Mathias Unberath, and Andreas K. Maier. Deep learning computed tomography: Learning projection-domain weights from image domain in limited angle problems. *IEEE Transactions on Medical Imaging*, 37(6):1454–1463, 2018.
- [304] Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, volume 15, pages 521–528. MIT Press, 2003.
- [305] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013. ISSN: 1063-6919.
- [306] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. FairGAN: Fairness-aware generative adversarial networks. In *Proceedings of the IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, IEEE, 2018.
- [307] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society, 2018.
- [308] Jin Yamanaka, Shigesumi Kuwashima, and Takio Kurita. Fast and accurate image super resolution by deep CNN with skip connection and network in network. In Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, and El-Sayed M. El-Alfy, editors, *Proceedings of the International Conference on Neural Information Processing (NIPS)*, volume 10635, pages 217–225. Springer, Springer, 2017.

- [309] Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam P. Harrison, Mohammadhadi Bagheri, and Ronald M. Summers. Deep lesion graphs in the wild: Relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9261–9270. IEEE, 2018.
- [310] Guang Yang, Qinghao Ye, and Jun Xia. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, 77:29–52, 2022.
- [311] Jiannan Yang, Fan Zhang, Bike Chen, and Samee U. Khan. Facial expression recognition based on facial action unit. In *Proceedings of the Tenth International Green and Sustainable Computing Conference (IGSC)*, pages 1–6. IEEE, 2019.
- [312] Xingyi Yang, Xuehai He, Jinyu Zhao, Yichen Zhang, Shanghang Zhang, and Pengtao Xie. COVID-CT-dataset: A CT scan dataset about COVID-19. *Computing Research Repository (CoRR)*, abs/2003.13865, 2020.
- [313] Suk Whan Yoon, Hui Lin, Michelle Alonso-Basanta, Nate Anderson, Ontida Apinorasethkul, Karima Cooper, Lei Dong, Brian Kempsey, Jaclyn Marcel, James Metz, Ryan Scheuermann, and Taoran Li. Initial evaluation of a novel cone-beam CT-based semi-automated online adaptive radiotherapy system for head and neck cancer treatment - a timing and automation quality study. *Cureus*, 12(8):e9660, 2020.
- [314] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P. Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 83–92. IEEE, 2020.
- [315] Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision*, 130(10):2425–2452, 2022.
- [316] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, pages 818–833. Springer International Publishing, 2014.

- [317] Chris Zhang, Wenjie Luo, and Raquel Urtasun. Efficient convolutions for real-time semantic segmentation of 3d point clouds. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 399–408. IEEE, 2018.
- [318] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.
- [319] Yanbo Zhang and Hengyong Yu. Convolutional neural network based metal artifact reduction in x-ray computed tomography. *IEEE Transactions on Medical Imaging*, 37(6):1370–1381, 2018.
- [320] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.
- [321] Zhicheng Zhang, Xiaokun Liang, Xu Dong, Yaoqin Xie, and Guohua Cao. A sparse-view CT reconstruction method based on combination of DenseNet and deconvolution. *IEEE Transactions on Medical Imaging*, 37(6):1407–1417, 2018.
- [322] Zhitao Zhao, Yang Lou, Yifeng Chen, Hongjun Lin, Renjie Li, and Genying Yu. Prediction of interfacial interactions related with membrane fouling in a membrane bioreactor based on radial basis function artificial neural network (ANN). *Bioresource Technology*, 282:262–268, 2019.
- [323] Alice Zheng, Michael Jordan, Ben Liblit, and Alex Aiken. Statistical debugging of sampled programs. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 16. MIT Press, 2003.
- [324] Yuqian Zhou, Jimin Pi, and Bertram E. Shi. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. In *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 872–877. IEEE, IEEE, 2017.
- [325] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z. Li. Multi-label CNN based pedestrian attribute learning for soft biometrics. In *International Conference on Biometrics (ICB)*, pages 535–540. IEEE, 2015.
- [326] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251. IEEE, 2017.

- [327] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8697–8710. IEEE, 2018.