Generative and Adversarial Learning for Object Recognition

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

ASTHA VERMA

(PhD18101)

(Department of Electronics and Communication Engineering)

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY
NEW DELHI- 110020

September, 2023

THESIS CERTIFICATE

The work contained in this thesis entitled, **Generative and Adversarial Learning for Object Recognition**, has also been submitted to Indraprastha Institute of Information Technology (IIIT), Delhi PhD program. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Advisors' Name

1) Dr. A V Subramanyam, Associate Professor, Depts. of ECE & CSE, IIIT Delhi

Rajiv Ratn Shah

2) Dr. Rajiv Ratn Shah, Assistant Professor, Depts. of HCD & CSE, IIIT Delhi

Place: New Delhi
Date: September 2023

ACKNOWLEDGEMENTS

I am deeply grateful to my advisors, Dr. A V Subramanyam and Dr. Rajiv Ratn Shah, for providing invaluable feedback on my analysis and framing, which significantly enhanced the quality of this thesis. Their kind words and guidance have been a constant source of motivation during my entire PhD journey.

I extend my gratitude to the Indraprastha Institute of Information Technology for their financial support during the development of this thesis. I would like to express my appreciation to the support personnel of IIIT-D, particularly to Mr. Adarsh, Mr. Bhawani, Ms. Sanjana, Ms. Binu, Ms. Preeti Patel, and Mr. Raju Biswas for their exceptional assistance in swiftly resolving all the technical and administrative issues. Special thanks to Mr. Rajendra from IIIT-D Library for his instrumental assistance in accessing the necessary materials, contributing to the success of my thesis.

I'm deeply thankful to Professor Shin'ichi Satoh for the invaluable internship at NII Tokyo, which has significantly enhanced my PhD journey. Special thanks to Dr. Zheng Wang and Dr. Hong Liu for their consistent support and guidance throughout.

I would like to thank my colleagues, Dr. Kajal Kansal, Dr. Monika Jain, Dr. Ambuj Mehrish, Mohit Sharma, Hemant Yadav, and Saumya Chaturvedi for guiding me during my PhD. I am also grateful to Siddhesh Bangar for his invaluable help during the completion of my thesis. Special thanks go to Naman Lal, Divij Gera, Himanshu Singh, Rameshwar Mishra, Nilanjana Chatterjee, Ritwik Mishra, and Mohammad Jauhar for their unwavering support and motivation throughout my PhD journey.

Above all, I express my gratefulness and love to my husband and colleague, Avinash Anand, whose continuous support made this work possible. I would also like to thank my parents, siblings, and other family members for their regular encouragement and unconditional support.

Astha Verma

(Astha Verma)

ABSTRACT

Generative modeling and adversarial learning have significantly advanced the field of computer vision, particularly in object recognition and synthesis, unsupervised domain adaptation, and adversarial attacks and defenses. These techniques have enabled the creation of more accurate and robust models for critical applications. In particular, we develop algorithms for fine-grained object recognition (Re-ID) and classification tasks. Re-ID involves matching objects across non-overlapping cameras, which is challenging due to visual recognition hurdles like pose change, occlusion, illumination variation, low resolution, and modality differences. On the other hand, object classification is another aim to categorize input data into pre-defined classes, using patterns learned from training data.

In this context, our thesis is motivated by the potential of generative modelling to synthesize novel human views, which can be used for unsupervised learning of Re-ID models. Unsupervised Re-ID suffers from domain discrepancies between labeled source and unlabeled target domains. Existing methods adapt the model using augmented samples, either by translating source samples or assigning pseudo labels to the target. However, translation methods may lose identity details, while label assignment may give noisy labels. Our approach is distinct from other methods in that it decouples the ID and non-ID features in a cyclic manner, which promotes better adaptation to pose and background, thereby resulting in richer novel views. This approach could improve the accuracy of Re-ID models for the unlabeled target domain, thus enhancing their robustness in real-world settings.

Furthermore, we aim to analyze the robustness of Re-ID and classification models and propose adversarial attack and defense methods to enhance their reliability. Adversarial attacks are a malicious technique that manipulates input data to cause machine learning models to make incorrect predictions or classifications. Adversarial defense methods, including adversarial training, certified defense, and detection mechanisms, are used to protect models from such attacks. By integrating adversarial attack and defense methods into model development and deployment, the risk of incorrect Re-ID and misclassification can be minimized, leading to robust models. This is especially important in critical applications such as surveillance and security systems. Our thesis aims to propose adversarial attack and defense mechanisms for Re-ID models and certify the robustness of classification models in both white-box and black-box settings.

Specifically, we address the limitations of conventional adversaries that consider Euclidean space and ignore the geometry of the pixels. We propose a stronger attack by incorporating geometry using the Wasserstein metric attack. To defend against such adversarial attacks, we propose a stochastic neural network that uses isotropic and anisotropic Gaussian noise to parameterize stochasticity. These parameters are learned under a meta-learning framework to make our defense more effective and scalable.

Finally, in order to provide a provable guarantee of a black-box model robustness, we propose a certified black-box defense via zeroth-order (ZO) optimization for image classification tasks. Previous works suffer from high model variance and low performance on high-dimensional datasets due to inadequate denoiser design and limited utilization of ZO techniques. To address these limitations, we introduce a robust UNet denoiser (RDUNet). RDUNet enables the model to learn intricate details while maintaining low reconstruction error, surpassing the performance of previously developed custom-trained denoisers.

We extensively evaluate our proposed generative and adversarial techniques using publicly available Re-ID and classification datasets - Market-1501, DukeMTMC-ReID, MSMT17, CUHK03, Veri-776, CIFAR-10, CIFAR-100, STL-10, Tiny Imagenet, and MNIST.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **i**

ABSTRACT **ii**

LIST OF TABLES **ix**

LIST OF FIGURES **xiv**

LIST OF SYMBOLS **xiv**

LIST OF ABBREVIATIONS **xv**

1 Introduction **1**

1.1 Motivation 1

1.2 Research Aims 3

1.3 Dissertation Organization 9

1.3.1 Contributions of this research 10

1.3.2 Applications and Future Work 11

2 Related Work **12**

2.1 Introduction 12

2.1.1 Evolution of fine grained object recognition (Re-ID) and Classification 12

2.2 Generative Learning 16

2.2.1 Domain Adaptation for Re-ID task 19

2.2.2 Unsupervised Domain Adaptation 20

2.3 Adversarial Learning 21

2.3.1 Adversarial Attacks 22

2.3.2 Adversarial Defense 24

2.4 Conclusion 27

3 Unsupervised Domain Adaptation for Person Re-Identification via Cyclic Generation **29**

3.1 Introduction 29

3.2 Proposed Methodology 33

3.2.1 Generative Learning 33

3.2.2 Discriminative Learning 41

3.3 Experiments 45
3.3.1 Datasets and Metrics 45
3.3.2 Implementation Details 46
3.3.3 Comparison with State-of-the-arts 46
3.3.4 Ablation Study 49
3.3.5 Evaluation of the Quality and Diversity of Generated Images 53
3.3.6 Visualizations 56
3.3.7 Background Adaptation of target domain 59
3.4 Chapter Summary 59

4 Adversarial Attack on Re-ID 61
4.1 Introduction 61
4.2 Preliminaries 63
4.3 Methodology 65
4.3.1 Wasserstein Metric Attack 65
4.3.2 Iterative Wasserstein Space Projection 65
4.4 Experiment 67
4.4.1 Datasets and Settings 67
4.4.2 White-Box Attack 68
4.4.3 Black-Box Attack 70
4.4.4 Effect of Regularization Parameter 71
4.4.5 Effect of size of local transport plan 71
4.5 Image Quality 72
4.6 Chapter Summary 73

5 Re-ID Defense 75
5.1 Introduction 75
5.2 Methodology 77
5.2.1 Proposed Stochastic Neural Network (SNN) 77
5.2.2 Meta Training 81
5.3 Experimental Settings 83
5.3.1 Performance against adversarial attacks 83
5.3.2 Comparison with State-of-the-art 86
5.3.3 Ablation Study 87
5.3.4 Challenges 88
5.4 Chapter Summary 89

6 Certified Zeroth-order Black-Box Defense 90
6.1 Introduction 90
6.1.1 ZO Optimization for adversarial learning. 94

6.2 Preliminaries 94
6.3 Methodology 96
6.3.1 Proposed Robust Architecture 97
6.3.2 Proposed Objective Function 98
6.3.3 Proposed Black-Box Defense Methods 100
6.4 Experimental Settings 103
6.4.1 Comparison with SOTA 104
6.4.2 Performance on Image Classification 105
6.4.3 Performance on Image Reconstruction 107
6.4.4 Ablation Study 108
6.5 Chapter Summary 112

7 Limitations 113
7.1 GAN-based Unsupervised Domain Adaptation 113
7.2 Wasserstein Metric Attack 113
7.3 Meta Perturbed Defense 114
7.4 Certified Defense 114

8 Thesis Summary and Future Directions 116
8.1 Introduction 116
8.2 Summary 116
8.3 Future Work 118
8.3.1 Unsupervised Domain Adaptation for Video Re-ID 118
8.3.2 Future of Adversarial Attacks in Wasserstein Space 119
8.3.3 Robustness to Adaptive Attacks 120

LIST OF TABLES

3.1 The comparison with other state-of-the-art methods. The mAP, CMC-1, CMC-5, and CMC-10 (%) results are reported. source → target represents the setting of training on the source dataset and testing on the target dataset. **Red** and blue fonts indicate the best and the second best results. $\surd$ indicates that the corresponding method uses the generation technique to augment training samples. 47

3.2 Comparison with SOTA hybrid methods. source → target represents the setting of training on the source dataset and testing on the target dataset. 'Duke' refers to DukeMTMC-ReID and 'Market' refers to Market-1501 datasets. 48

3.3 Quantitative evaluation of effectiveness of pose change in unsupervised domain. 'M', 'D' and 'MSMT' denote the Market-1501 (Zheng *et al.*, 2015), DukeMTMC-ReID (Zheng *et al.*, 2017) and MSMT17 (Wei *et al.*, 2018) datasets respectively. 51

3.4 Ablation study on different types of training samples. "Supervised" stands for the labeled source domain. "Direct Testing" denotes the unlabeled target domain by the model trained in supervised setting. "Single Encoding" refers to generation of intermediate image. "Cyclic Encoding" indicates our proposed IPES-GAN network. source → target represents the setting of training on the source dataset and testing on the target dataset. 'M' and 'D' denote the Market-1501 and DukeMTMC-ReID datasets, respectively. The mAP and CMC-1 (%) results are reported. 52

3.5 Ablation study on loss functions. source → target represents the setting of training on the source dataset and testing on the target dataset. 'M' and 'D' denote the Market-1501 and DukeMTMC-ReID datasets respectively. The mAP and CMC-1 (%) results are reported. 52

3.6 Quantitative evaluation to check quality of image. LPIPS: Learned perceptual image similarity, higher values indicate more similarity. FID: Fréchet Inception distance, lower the better. 53

3.7 The Mean Opinion Score (MOS) test results on the generated images from DukeMTMC-ReID to Market-1501. 55

3.8 Maximum Mean Discrepancy (MMD). $\mathcal{S}$ and $\mathcal{T}$ represents source and target dataset, respectively. $\mathcal{S} \rightarrow \mathcal{T}$ (source → target) represents the setting of training on the source dataset and testing on the target dataset. 'M', 'D' and 'MS' denote the Market-1501, DukeMTMC-ReID and MSMT17 datasets, respectively. **Red** and blue fonts indicates MMD after training and before training. 55

4.1 White-Box Attack comparison with ODFA (Zheng *et al.*, 2018), TCIAA (Wang *et al.*, 2020*a*), UAP (Li *et al.*, 2019*b*) and Meta-Attack (Yang *et al.*, 2021*a*) 68

4.2 White-Box Attack comparison with TCIAA (Wang *et al.*, 2020*a*) for perturbation radius ($\epsilon = 8$). Lower the rank accuracy, better is the performance of the threat model. 69

4.3 White-Box Attack comparison with TCIAA (Wang *et al.*, 2020*a*) on models trained on MSMT17 (Wei *et al.*, 2018). 70

4.4 Attacking Re-ID Backbones. IDE(ResNet-50) (Zheng *et al.*, 2016*b*) and Inception-v3 (Qian *et al.*, 2017) trained on CUHK03 dataset. We compare our proposed approach with GAP (Poursaeed *et al.*, 2018),U-PGD (Madry *et al.*, 2017), TCIAA (Wang *et al.*, 2020*a*) ($\epsilon = 16$). . 71

4.5 **Cross-Model Attack:** Known N/W → Unknown N/W denotes the noise is learned by a known model and tested on an unknown model. All experiments carried out on Market-1501 (Zheng *et al.*, 2015). . . 72

4.6 SSIM scores and MMD to compare the quality of adversarial samples ($\epsilon = 8$) on Market-1501 (Zheng *et al.*, 2015) and DukeMTMC-ReID (Ristani *et al.*, 2016) datasets. Larger SSIM scores coupled with larger MMD values indicates better performance. 73

5.1 Performance of Re-ID models against adversarial attacks after applying our defense method. 'R-50' is the baseline model of ResNet-50. 'MP-SNN (R-50)' is our proposed approach. 84

5.2 Performance of Re-ID models against adversarial attacks after applying our defense method. 'OS' is the baseline model of OSNet. 'MP-SNN (OS)' is our proposed approach MP-SNN with OSNet as backbone. 84

5.3 Performance against our proposed Wasserstein Metric Attack (WMA). 'R-50' is SOTA ResNet-50 He *et al.* (2016) and 'OS' is OSNet Zhou *et al.* (2019). ($\epsilon = 8/255$). 85

5.4 Generalization of MP-SNN against various attacks. PGD-AT is adversarial training using adversarial samples generated using PGD attack. 'R-50' is the baseline SOTA ResNet-50 (He *et al.*, 2016). CMC-1 in %. Dataset is Market-1501 ($\epsilon = 8/255$). 86

5.5 Comparison with SOTA on Market-1501 dataset. CMC-1 in %. * denotes re-ranking. $\epsilon = 5/255$. 86

5.6 Effect of FCA on MP-SNN. $\epsilon = 8/255$ 87

5.7 Effect of loss functions and meta-training on our proposed MP-SNN. 'ML' is our proposed meta-learning strategy. Dataset is Market-1501. $\epsilon = 8/255$. 87

6.1 Comparison with SOTA certified defense techniques in white-box (W) and black-box (B) settings on CIFAR-10 dataset. 'R' and 'C' are RGE and CGE optimization techniques. 'q'=192. 105

6.2 Comparison of our defense with previous ZO defense approaches. 'q' is the number of queries. 106

6.3 Performance comparison with SOTA image reconstruction tasks which are Vanilla Model (Raj *et al.*, 2020), FO-DS (Salman *et al.*, 2020), FO-AE-DS (Zhang *et al.*, 2022*d*), ZO-DS (R) (Zhang *et al.*, 2022*d*) and ZO-AE-DS (C) (Zhang *et al.*, 2022*d*) on MNIST dataset. ($q = 192$) 108

6.4 Effect of loss functions on our proposed (ZO-AE-RUDS) for $q = 192$. Dataset is CIFAR-10. 109

6.5 Effect of RDUNet training strategies on ZO-AE-RUDS for $q = 192$. Dataset is CIFAR-10. 109

6.6 Comparison of our proposed approach with previous ZO optimization approaches on the CIFAR-10 dataset for different noise levels. Query ‘q’=192. 110

6.7 Comparison of our proposed approach with previous ZO optimization approaches on STL-10 dataset for different noise levels. Query ‘q’=192. 111

6.8 Certified Accuracy for different classifiers on CIFAR-10 dataset for noise level $\sigma = \{0.25\}$. Query ‘q’=192. 112

LIST OF FIGURES

1.1 Generative and adversarial methodologies have demonstrated notable advantages in the realm of computer vision applications. The central objective of this inquiry is directed towards Re-ID and Image classification undertakings. Through the use of a range of generative and adversarial techniques, we have attained exceptional levels of performance on these tasks, surpassing previous state-of-the-art benchmarks. 3

1.2 The objective of our research is to address the constraints of prior works in the field of object recognition. Specifically, our thesis emphasizes on generative and adversarial approaches, as illustrated by the green box. Meanwhile, the red box highlights the limitations of various generative and adversarial techniques. 4

1.3 Person Re-identification through Unsupervised Domain Adaptation. The input comprises of a set of labeled source images and a set of unlabeled target images. A shared encoder is utilized to obtain a common embedding, which is then leveraged by a generator to produce generated images as output. **(Best Viewed in Colors.)** 5

1.4 An illustration of adversarial attack on person Re-ID. By introducing imperceptible noise to the query images, the similarity score between the query image and the true positive decreases from 0.845 to 0.208, while the similarity score with the true negative from the gallery image increases from 0.179 to 0.732. The adversarial noise has been resized to the range [0, 1] for visualization purposes. 6

1.5 The illustration of adversarial defense. Vanilla adversarial defense is training of DNN with clean and adversarial images. We aim to increase the robustness by introducing model and data optimization. 7

1.6 We examine a situation in which defense against adversarial attacks is illustrated for a black-box model, with the interaction between the defender and the model limited to input-output function queries. . . 8

2.1 It illustrates the development from early works to deep learning-based methods for Re-ID task and highlights their susceptibility to adversarial attacks. In order to address the robustness and safeguarding of target models, researchers have proposed adversarial defense techniques. . 13

2.2 The thesis focuses on the progression from neural networks to deep learning techniques and ultimately, to generative and adversarial learning approaches for Object Recognition. The central emphasis is on unsupervised domain adaptation, as well as adversarial attack and defense within the realm of generative and adversarial learning methods. . . 18

3.1 (a) We make a comparison with four representative generation based methods, including SPGAN (Deng *et al.*, 2018), SBSGAN (Huang *et al.*, 2019*a*), FD-GAN (Ge *et al.*, 2018), and CR-GAN (Chen *et al.*, 2019), by translating the image $I^{\mathcal{S}}$ in the source domain according to the image $I^{\mathcal{T}}$ in the target domain. Different colors of squares indicate different backgrounds, and different shapes in the squares indicate different poses. (b) Illustration of images generated by our model IPES-GAN in the source domain. The source (concatenated with target pose) and the target images are input to our framework. The output is the generated images with the superimposed pose, showing that these images have adapted to the pose of the target images (**Best viewed in colors**). . . 31

3.2 The overview of our Individual-preserving Environmental-switching Network (IPES-GAN). Pose encoders help to achieve pose guidance. Style encoders learn the camera-style features. Identity encoder is shared between labeled source and unlabeled target domain. Image generators' (decoders') output are environmental switched images $(I^{\mathcal{S}\rightarrow\mathcal{G}}, I^{\mathcal{T}\rightarrow\mathcal{G}})$ which are passed again to the generator (encoder-decoder) network to generate cyclic generated images ($\widetilde{I^{\mathcal{S}}}$, $\widetilde{I^{\mathcal{T}}}$). Discriminator helps to achieve image-refinement. KL Divergence loss is calculated between the probability distribution predicted by our identity encoder and a standard CNN model E_{std} trained on source dataset. (**Best viewed in colors**). . . . 34

3.3 Our encoder network has pose encoders (E_p^S, E_p^T), style encoders (E_s^S, E_s^T) and shared identity encoder E_{id}. Source (target) images are concatenated with cross-domain 18-channel pose keypoints are input to the pose encoder to capture pose features. Similarly, we capture source (target) style features and shared identity features. (**Best viewed in colors**). 35

3.4 The pose and style (environment) features obtained from pose encoders (E_p^S, E_p^T) and style encoders (E_s^S, E_s^T) respectively, are concatenated and exchanged between the source and target domain. We also term this exchange as environmental swap. (Refer Figure 3.2 for legend of different modules). (**Best viewed in colors**). 36

3.5 We calculate the **feature consistency loss** between the features of real images and generated images for feature reconstruction with minimum noise. (**Best viewed in colors**). 38

3.6 We calculate **cyclic reconstruction loss** between original images and their cyclically generated counterparts to provide a supervision for unpaired source and target domain images. (**Best viewed in colors**). . 39

3.7 Our approach to incorporating cross-camera style variation into the generated images involves the formulation of a **camera loss** function. (**Best viewed in colors**). 40

3.8 Pipeline of our approach. Our IPES-GAN network is used to generate images (Translated Images) with environment style (pose and background) of target image while preserving the identity-related features of the source images. The cross-entropy loss $\mathcal{L}_{CE}$ performs classification on labeled source images. The Soft cross-entropy loss $\mathcal{L}_{Soft\_CE}$ and divergence loss $\mathcal{L}_{KL}$ classify generated images that are pseudo-labeled with source labels. We employ these losses on generated images to fine-tune the identity encoder, which in turn improves generation. (**Best viewed in colors**). 42

3.9 KL Divergence between the probability predicted by our trained identity encoder E_{id} and standard CNN model E_{std}. 44

3.10 (a) Qualitative comparison of our method with FD-GAN (Ge *et al.*, 2018). (b) our method with SPGAN (Deng *et al.*, 2018). (**Best viewed in colors**). 50

3.11 Visualization of translated images from the source (Market-1501, the first column) to the target (DukeMTMC-ReID, the first row). (**Best viewed in colors**). 54

3.12 Visualization of translated images from the source (DukeMTMC-ReID, the first column) to the target (Market-1501, the first row). (**Best viewed in colors**). 54

3.13 Visualization of translated images from the source (DukeMTMC-ReID, the first column) to the target (MSMT17, the first row). (**Best viewed in colors**). 56

3.14 Visualization of translated images from the source (Market-1501, the first column) to the target (MSMT17, the first row). (**Best viewed in colors**). 57

3.15 The top-10 predictions made by our proposed IPES-GAN model with a majorly occluded image as query. The green boxes represent correct predictions of query image from the gallery set and red represents the incorrect predictions. Red Circles in the query images represent the occlusion. 58

3.16 The figure shows the *Generated Image ($M \rightarrow D$)* adapts to pose and background of *Target (D)* image while maintaining identity-related features of the *Source (M)* image. M and D denotes Market-1501 and DukeMTMC-ReID. (**Best viewed in colors**). 59

4.1 We show images with four pixels to demonstrate the difference between Wasserstein perturbations and Δ_∞ perturbations. A small perturbation Δ_W shifts the image one pixel to the right, which is minimal in Wasserstein distance but maximal in Δ_∞ distance. In contrast, a small perturbation Δ_∞ changes all pixels to be grayer, which is minimal in Δ_∞ distance but maximal in Wasserstein distance. 62

4.2 The top-10 predictions before and after our (a) non-targeted (b) targeted Wasserstein attack. As is visible in (a) after attack the top-10 predictions can be any other sample except the same identity (ID) as query whereas in (b) same IDs as target adversary sample are mainly included in the ranking list, degrading the performance of Re-ID model. The green boxes represent correct predictions of query image from the gallery set and red represents the incorrect predictions. 63

4.3 Wasserstein Ball. 64

4.4 The figure illustrates the adversarial sample obtained by adding noise to clean images. 66

4.5 The adversarial samples obtained after attack. Above row shows images for perturbation radius $\epsilon = 8$ which are better than below row with radius $\epsilon = 16$. 70

4.6 The adversarial samples generated for different values of regularization parameter λ. 71

4.7 Attack performance for different transport sizes. 73

5.1 Meta Perturbed Defense: We propose a SNN by adding noise modules to a SOTA backbone. We train this SNN using a novel meta-learning strategy with tasks as ‘V’, ‘P’, and ‘PA’ to obtain our proposed framework MP-SNN. 76

5.2 Top-10 predictions for Re-ID task on Market-1501 dataset after attack on ResNet-50 and after applying our defense. Green box - correct prediction; Red box - incorrect prediction. ‘AQ’ is attacked query image. 76

5.3 **Overview of MP-SNN.** Our SNN injected with anisotropic noise in the penultimate layer and isotropic noise in the inner layers is learnt via meta-learning. During training, we have n tasks over training data D, which we divide into $n-1$ meta-train tasks and a single meta-test task. In meta-train stage, L_{FCA} and L_{Iso} are calculated as meta-train loss $\mathcal{L}_{mtr}(\Theta)$. In meta-test stage, we copy the original model and update it using $\mathcal{L}_{mtr}(\Theta)$. We compute meta-test losses $\mathcal{L}_{mte}(\Theta)$ and $\mathcal{L}_{mte}(\Theta^{'})$ on original and updated model respectively. Finally, we update our original model with $L_{meta}(\Theta)$. 77

6.1 We make a comparison with four previous certified defense methods, including RS (Cohen *et al.*, 2019), DS(W) (Salman *et al.*, 2020), DS(B) (Salman *et al.*, 2020), ZO-DS (Zhang *et al.*, 2022*d*) and ZO-AE-DS (Zhang *et al.*, 2022*d*) (ZO-optimization approaches). ‘W’ and ‘B’ refer to white-box (defense technique can utilize weights of target model f) and black-box settings. ‘x’ - input sample, ‘η’ - noise, ‘E’ - Encoder, ‘D’ - Decoder, ‘f’ - target model, f_s - surrogate model (proxy of f) and ‘z’ - latent feature vector. 93

6.2 An overview of proposed certified defense mechanism via robust UNet denoiser RDUNet. Noise is added to input sample x which is given as input to the robust denoiser. The output of denoiser is the residual map which when added to noisy image x^* gives denoised output $\hat{x}$. The denoised output is input of the autoencoder architecture which is then send as input to black-box model f. 97

6.3 Architecture of our defense technique ZO-AE-RUDS. 102

6.4 Comparison of Certified Accuracy on low-dimension (CIFAR-10, CIFAR-100) and high-dimension (STL-10 and Tiny Imagenet) datasets for different l_2 radius at query $q = 192$. 106

6.5 Effect of different denoisers on our RGE and CGE ZO optimization-based defense approaches for different l_2-radii at q=192. Dataset is CIFAr-10. 108

6.6 Training progress of our certified black-box defense ZO-RUDS (RGE) and ZO-AE-RUDS (CGE) over number of epochs. Dataset used is CIFAR-10. Query ‘q’ = 192, $\sigma = 0.25$. 111

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ML	Machine Learning
Re-ID	Re-Identification
IDs	Identities
DNN	Deep Neural Network
UDA	Unsupervised Domain Adaptation
CNN	Convolution Neural Network
GAN	Generative Adversarial Network
AE	Autoencoder
VAE	Variational Autoencoder
DANN	Domain-Adversarial Neural Network
KL	Kullback-Leibler
CE	Cross-Entropy Loss
Soft_CE	Soft Cross-Entropy Loss
CMC	Cumulative Matching Characteristics
mAP	Mean Average Precision
ASPP	Atrous Spatial Pyramid Pooling
FID	Fréchet Inception Distance
LPIPS	Learned Perceptual Image Patch Similarity
MOS	Mean Opinion Score
BPFC	Bit Plane Feature Consistency
ReLU	Rectified Linear Unit
MMD	Maximum Mean Discrepancy
FGSM	Fast Gradient Sign Method
I-FGSM	Iterative Fast Gradient Sign Method
MI-FGSM	Momentum Iterative Fast Gradient Sign Method
PGD	Projected Gradient Descent
AT	Adversarial Training
N\W	Network
SSIM	Structural Similarity Index
SSN	Stochastic neural network
FO	First Order
ZO	Zeroth Order
RS	Randomized Smoothing
DS	Denoised Smoothing
RGE	Randomized Gradient Estimation
CGE	Co-ordinate Gradient Estimation
Conv	Convolution
RKHS	Reproducing Kernel Hilbert Space
SGD	Stochastic Gradient Descent
ADAM	Adaptive Moment Estimation

SCA	Standard Certified Accuracy
RCA	Randomized Certified Accuracy
RMSE	Root Mean Square Error
ViT	Vision Transformer

CHAPTER 1

Introduction

1.1 Motivation

Rapid advancements in the field of machine learning (ML) and computer vision have revolutionized the way we process and analyze data. These techniques involve the use of algorithms to automatically learn and improve from experience, making it possible to extract meaningful insights and patterns from vast amounts of data with incredible speed and accuracy. By leveraging the power of machine learning techniques, our research interests are focused on two areas that aim to explore the potential of data-driven solutions for addressing complex real-world problems. The first is Re-ID, which involves the re-identification of objects, people, or events in different settings, and is the primary focus of our thesis. The second area is classification task, which involves categorizing objects into various classes. The development of effective Re-ID and classification techniques is crucial for a variety of applications, including surveillance, security, and healthcare.

Re-ID is a problem of re-identifying a given query image from a gallery set, captured from non-overlapping camera views. It is a popular field of computer vision, widely employed in a batch of valuable multimedia applications, such as public security, traffic surveillance, human behavior modeling, and smart city (Wang *et al.*, 2015, 2016; Ye *et al.*, 2016; Li *et al.*, 2018*a*; Yao *et al.*, 2019; Huang *et al.*, 2019*b*; Li *et al.*, 2019*d*; Wang *et al.*, 2019*b*; Zhao *et al.*, 2019*a*; Zeng *et al.*, 2020; Zhao *et al.*, 2021*a*; Gu *et al.*, 2022; Wu *et al.*, 2023).

Re-ID can be formally defined as follows: the training data $D \in R^{d\times N}$ consists of N samples, each represented by a column $x_i \in R^{d\times 1}$ in the data matrix D. For each sample x_i, its label $y_i \in R^{K\times 1}$ is a one-hot encoded vector, where K is the number of identities. The distance between any two different samples x_i and x_j can be computed using a distance metric such as Euclidean distance, denoted by $\mathcal{D}$. Specifically, it is given as,

$$\mathcal{D}(x_i, x_j) = \|x_i - x_j\|_2^2 \tag{1.1}$$

Our objective is to train a deep feature representation model for Re-ID that can produce an optimal feature embedding such that the sample x_j, which belongs to the same identity as x_i, has the lowest distance to x_i, i.e., it is the closest match. Specifically, we aim to learn a feature embedding that satisfies the condition $\mathcal{D}(x_i, x_j) < \mathcal{D}(x_i, x'_j)$, where x'_j represents any sample that belongs to a different identity than x_i.

Generative learning has emerged as a powerful tool for addressing the problem of Re-ID by generating synthetic images similar to real data (Deng *et al.*, 2018; Huang *et al.*, 2019*a*; Qu *et al.*, 2023; Liu *et al.*, 2023). However, due to domain shift between existing source data and an unknown target data, the effectiveness of deep neural networks (DNNs) is limited in real-world applications, where the model fails to generalize to unknown domains (Ge *et al.*, 2018; Chen *et al.*, 2019; Feng *et al.*, 2021*a*). Unsupervised domain adaptation (UDA) is a promising solution to this problem, with aim to adapt the DNN to target domain without labeled data (Sun and Saenko, 2016; Bousmalis *et al.*, 2017; Ding *et al.*, 2020; Ge *et al.*, 2020; Li *et al.*, 2021*a*; Rami *et al.*, 2022; Peng *et al.*, 2023; Qi *et al.*, 2023).

DNNs have made remarkable advancements in Re-ID tasks, but they remain vulnerable to adversarial attacks. Prior research has explored attack strategies such as constrained optimization, gradient optimization, generative models, and estimated decision boundaries in both white-box and black-box settings, among others, as documented in literature references (Zheng *et al.*, 2018; Li *et al.*, 2019*b*; Zhao *et al.*, 2019*b*; Bai *et al.*, 2020*b*; Wang *et al.*, 2020*a*; Ding *et al.*, 2021; Gong *et al.*, 2022; Wang *et al.*, 2022*b*,*a*; Yang *et al.*, 2022; Subramanyam, 2023). l_p-based perturbation norms are the primary approach in these studies, but they ignore the data's intrinsic geometry. The Wasserstein metric, which incorporates a cost matrix for pixel mass movement during attacks, is a more robust alternative. Our thesis is the first to intorduce Wasserstein metric for white-box and black-box adversarial attacks on Re-ID systems by projecting adversarial samples in the Wasserstein ball.

In recent times, defending against adversarial attacks has emerged as a critical research domain. To enhance the robustness of DNNs against adversarial attacks in both white-box and black-box settings while preserving high performance on clean data, adversarial defense techniques such as model optimization, data optimization, and additional network utilization are being explored (Madry *et al.*, 2017; Wong and Kolter, 2018; Bai *et al.*, 2020*b*; Bouniot *et al.*, 2020; Wang *et al.*, 2022*a*; Salman *et al.*, 2022;

Zhang *et al.*, 2022*d*; Seo *et al.*, 2023). Our objective is to offer robust protection against adversarial attacks for Re-ID tasks through a combination of model optimization (stochastic network) and data optimization (adversarial training). Furthermore, our goal is not only to provide empirical defense (adversarial training) in a white-box setting for Re-ID tasks but also to ensure the effectiveness of our defense in popular computer vision tasks like classification through certification (Akhtar *et al.*, 2021; Liang *et al.*, 2022).

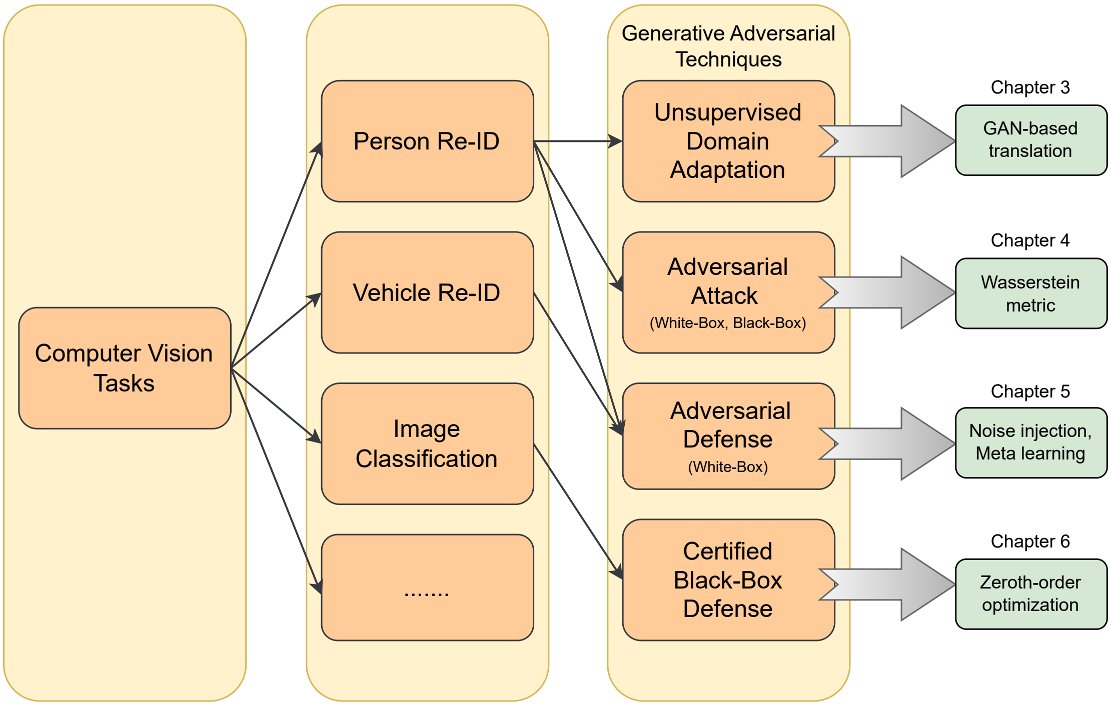


Figure 1.1: Generative and adversarial methodologies have demonstrated notable advantages in the realm of computer vision applications. The central objective of this inquiry is directed towards Re-ID and Image classification undertakings. Through the use of a range of generative and adversarial techniques, we have attained exceptional levels of performance on these tasks, surpassing previous state-of-the-art benchmarks.

1.2 Research Aims

The focus of this thesis, as depicted in Figure 1.1, is on the application of generative and adversarial techniques in computer vision tasks. Generative and adversarial learning are both active areas of research in machine learning with various potential applications. The research objectives of this thesis can be broadly categorized into two areas: improving generative models and enhancing the robustness of existing DNNs. Recent studies in generative and adversarial learning have concentrated on developing

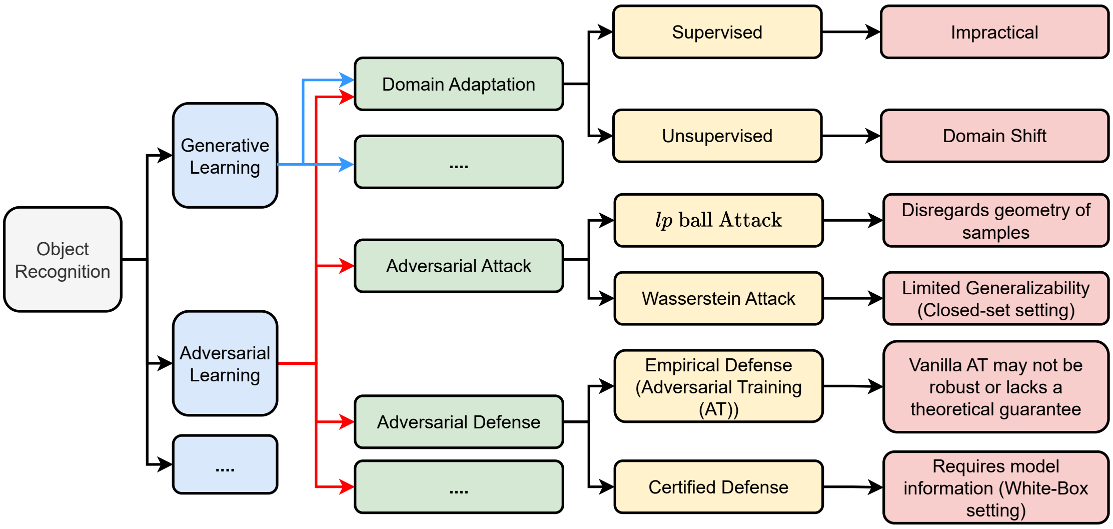

Object Recognition
Generative Learning
Adversarial Learning
....
Domain Adaptation
....
Adversarial Attack
Adversarial Defense
....
Supervised
Unsupervised
lp ball Attack
Wasserstein Attack
Empirical Defense (Adversarial Training (AT))
Certified Defense
Impractical
Domain Shift
Disregards geometry of samples
Limited Generalizability (Closed-set setting)
Vanilla AT may not be robust or lacks a theoretical guarantee
Requires model information (White-Box setting)

Figure 1.2: The objective of our research is to address the constraints of prior works in the field of object recognition. Specifically, our thesis emphasizes on generative and adversarial approaches, as illustrated by the green box. Meanwhile, the red box highlights the limitations of various generative and adversarial techniques.

models capable of generating high-quality samples with diverse features (Karras *et al.*, 2017; Song and Ermon, 2019; Ho *et al.*, 2020; Esser *et al.*, 2021; Zhang *et al.*, 2022*b*,*c*; Ge *et al.*, 2023). Moreover, in adversarial learning, researchers have worked on developing more efficient adversarial attack and defense methods. Adversarial attacks aim to expose and exploit machine learning models' vulnerabilities and identify potential weaknesses for attackers (Bai *et al.*, 2020*b*; Yuan *et al.*, 2021; Feng *et al.*, 2021*b*; Subramanyam, 2023). Empirical defense (adversarial training) and certified defense algorithms are among the adversarial defense techniques that can improve the robustness of machine learning models to adversarial attacks (Wong *et al.*, 2020; Bai *et al.*, 2020*b*; Wang *et al.*, 2022*b*; Gong *et al.*, 2022). The objective of this thesis is to create novel algorithms for generative and adversarial learning in Re-ID and classification tasks. To achieve this goal, the following four research objectives (aims) have been identified.

Aim 1: To Investigate Unsupervised Domain Adaptation in Person Re-ID

Significant progress has been made in solving the problem of person Re-ID in a supervised setting through recent advances such as (Zheng *et al.*, 2015; Xiao *et al.*, 2016; Hermans *et al.*, 2017; Bai *et al.*, 2017; Zheng *et al.*, 2017; Bai *et al.*, 2017; Zhong *et al.*, 2018*b*; Sun *et al.*, 2018; Tang *et al.*, 2019; Zhao *et al.*, 2019*a*, 2020*a*,*b*,*a*, 2021*a*; Gu *et al.*, 2022; Wu *et al.*, 2023). However, the supervised setting faces challenges in the open-set scenario due to the impracticality of manually annotating hundreds and

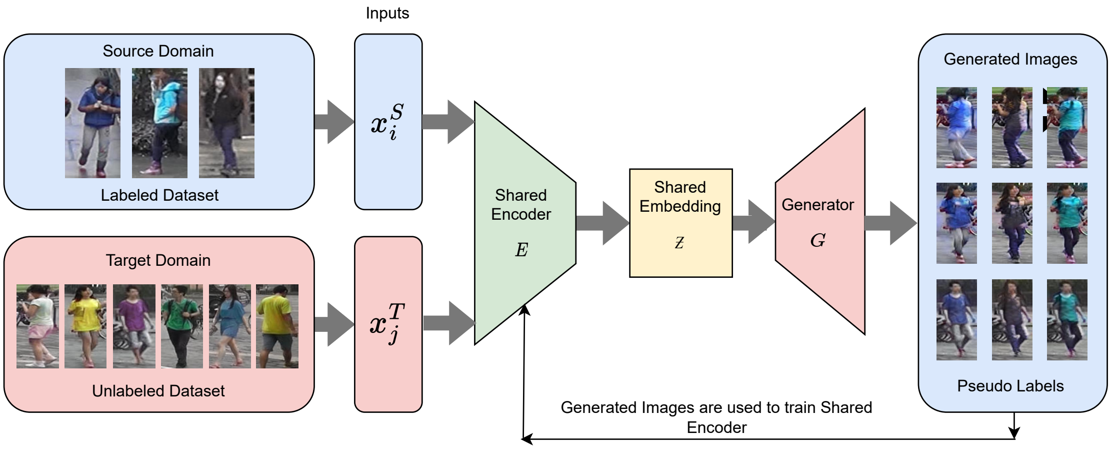


Figure 1.3: Person Re-identification through Unsupervised Domain Adaptation. The input comprises of a set of labeled source images and a set of unlabeled target images. A shared encoder is utilized to obtain a common embedding, which is then leveraged by a generator to produce generated images as output. **(Best Viewed in Colors.)**

thousands of images in unseen scenes. Additionally, directly applying a Re-ID model trained on an existing dataset (source) to a new dataset (target) results in significant performance degradation due to domain shift. Existing methods in this field aim to reduce domain shift caused by changes in context, camera style, or viewpoint by fine-tuning and adapting the Re-ID model with augmented samples. These samples can be obtained by either translating source samples to the target style or assigning pseudo labels to the target as shown in Figure 1.3. However, the former method may result in the loss of identity details while preserving redundant source background during translation, while the latter technique may introduce noisy labels when the model encounters previously unseen backgrounds and person poses (refer to Domain Adaptation block in Figure 1.2). In order to overcome the limitation of previous works, we aim to solve the unsupervised person Re-ID problem by designing a model that generates cross-domain images using a cyclic generation network. Our model separates the representation into individual and environmental parts to maintain identity-related features while adapting to the background and pose-related information.

Aim 2: To Investigate Adversarial Attack on Person Re-ID

The investigation in *Aim 1* led to the design of a reliable technique to improve DNNs for person Re-ID, however these models can be vulnerable to adversarial attacks as shown in Figure 1.4. Adversarial attacks have been studied extensively in closed-set settings like classification, object detection, and segmentation (Goodfellow *et al.*,

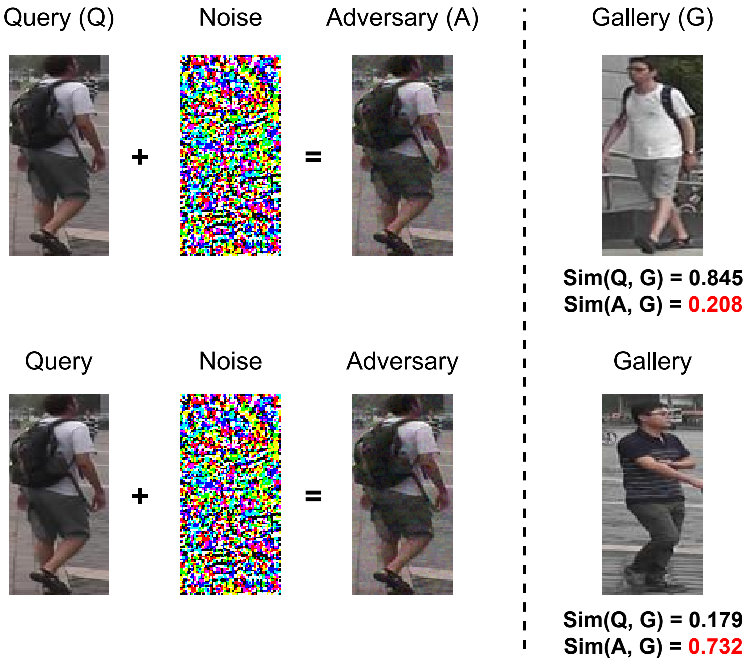


Figure 1.4: An illustration of adversarial attack on person Re-ID. By introducing imperceptible noise to the query images, the similarity score between the query image and the true positive decreases from 0.845 to 0.208, while the similarity score with the true negative from the gallery image increases from 0.179 to 0.732. The adversarial noise has been resized to the range [0, 1] for visualization purposes.

2014*b*; Kurakin *et al.*, 2016; Madry *et al.*, 2017; Xie *et al.*, 2017; Dong *et al.*, 2018*b*; Zhang and Wang, 2019; Liu *et al.*, 2019*a*; Qiu *et al.*, 2019; Tu *et al.*, 2020; Xu *et al.*, 2020; Saha *et al.*, 2020; Li *et al.*, 2021*d*; Ren *et al.*, 2020; Kim *et al.*, 2021; Liu *et al.*, 2022; Liu and Hu, 2022; Li *et al.*, 2023; Deb *et al.*, 2023). However, there have been few attempts to investigate adversarial attacks in open-set retrieval problems like person Re-ID, where source and target datasets have non-overlapping labels (Zheng *et al.*, 2018; Li *et al.*, 2019*b*; Zhao *et al.*, 2019*b*; Bai *et al.*, 2020*b*; Wang *et al.*, 2020*a*; Ding *et al.*, 2021; Yang *et al.*, 2021*a*). Previous works on adversarial attacks in person Re-ID mainly focus on l_∞ perturbations and its corresponding l_p generalization, using metric-based attacks (Zheng *et al.*, 2018; Bai *et al.*, 2020*b*), GAN-based attacks (Zhao *et al.*, 2019*b*; Wang *et al.*, 2020*a*), or meta-learning based attacks (Yang *et al.*, 2021*a*) (refer to Adversarial Attack block in Figure 1.2). We aim to use Wasserstein ball perturbations, which provide more generalized image perturbations in the form of pixel mass movement. Unlike previous methods, our approach does not require training to learn perturbations. To the best of our knowledge, this is the first proposal to use Wasserstein ball perturbations in open-set ranking problems such as Re-ID. Wasserstein metric

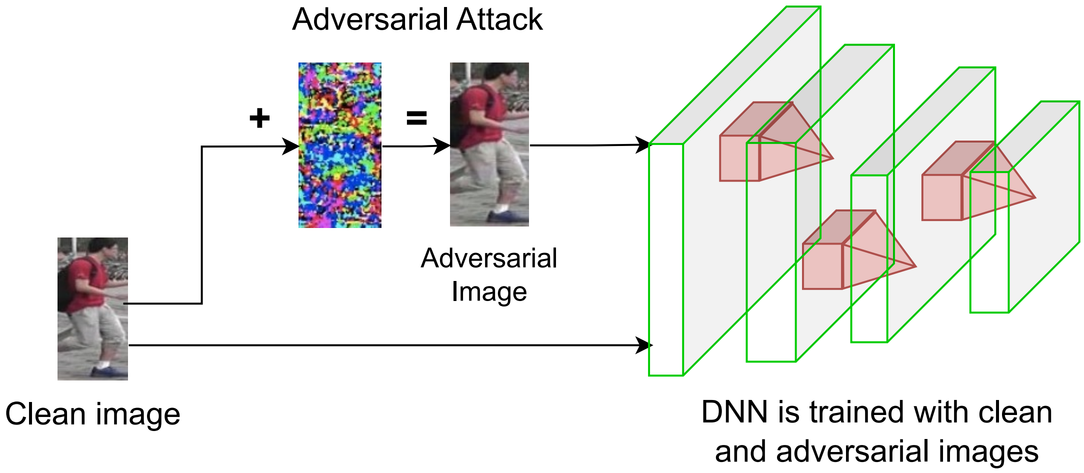


Figure 1.5: The illustration of adversarial defense. Vanilla adversarial defense is training of DNN with clean and adversarial images. We aim to increase the robustness by introducing model and data optimization.

based perturbations provide more generalized image perturbations than l_p ball attacks in Re-ID systems, but have not been proposed for open-set ranking problems. Previous works have used Wasserstein metric for classification tasks (Wong *et al.*, 2019; Wu *et al.*, 2020*a*; Hu *et al.*, 2020), but it cannot be directly applied to Re-ID. The approach proposed here is to project clean images into a Wasserstein ball to generate adversarial samples, attacking an entire ranking model in an open-set setting.

Aim 3: To Investigate Adversarial Defense

Adversarial defense refers to the techniques and methods used to enhance the robustness of machine learning models against adversarial attacks as shown in Figure 1.5. Our aim is to enhance DNNs' robustness against adversarial attacks by introducing model and data optimization techniques, including stochastic network, meta-learning, and adversarial training. Numerous works have emphasized the dependence on adversarial training as a leading approach for protecting against adversarial attacks, as observed in prior research (Madry *et al.*, 2017; Bai *et al.*, 2020*b*; Bouniot *et al.*, 2020; Bai *et al.*, 2021*a*; Xu *et al.*, 2021*a*; Wang *et al.*, 2022*a*; Kinfu and Vidal, 2022; Addepalli *et al.*, 2022; Cheng *et al.*, 2023). However, employing vanilla adversarial training fails to provide robustness against strong adversarial attacks and is prone to overfitting to a particular attack leading to poor model generalization (Song *et al.*, 2018; Tsipras *et al.*, 2019; Lin *et al.*, 2020; Devaguptapu *et al.*, 2021; de Jorge Aranda *et al.*, 2022; Baytaş and Deb, 2023) (refer to Adversarial Defense block in Figure 1.2). We aim to develop a defense mechanism that defends against strong adversarial attacks while maintaining

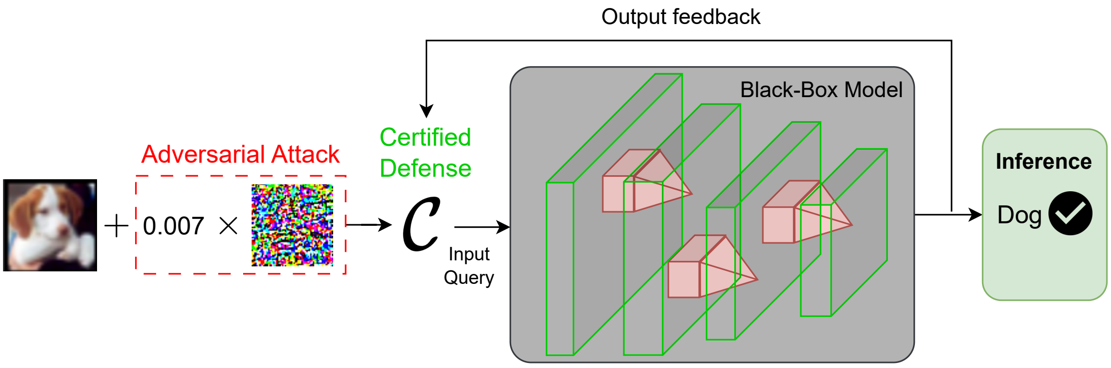


Figure 1.6: We examine a situation in which defense against adversarial attacks is illustrated for a black-box model, with the interaction between the defender and the model limited to input-output function queries.

high generalizability. Our approach involves a meta perturbed stochastic neural network that learns anisotropic and isotropic noise distribution in a novel meta-learning defense algorithm. Previous studies have shown that isotropic or anisotropic noise injection improves generalizability, but using only one type of noise has limitations (Jeddi *et al.*, 2020; Eustratiadis *et al.*, 2021). Our approach combines both types of noise to provide a richer noise distribution for a more challenging Re-ID task.

Aim 4: To Investigate Certified Black-box Defense

Empirical defenses refer to those that have been demonstrated to be resilient against known adversarial attacks through empirical evidence. Among all empirical defenses tested so far, adversarial training (Szegedy *et al.*, 2013; Kurakin *et al.*, 2016; Madry *et al.*, 2017; Zhang *et al.*, 2019*b*; Kinfu and Vidal, 2022; Cheng *et al.*, 2023) has been proven to be the most effective. Our robustification algorithm, as described in *Aim 3*, falls under the category of empirical defense. However, these methods may not always be certifiably robust (Uesato *et al.*, 2018; Croce and Hein, 2020). Another line of research is certified defense, where an off-the-shelf model's prediction is certified within the neighborhood of the input. These methods are called certified defense techniques (Katz *et al.*, 2017; Wong and Kolter, 2018; Raghunathan *et al.*, 2018; Salman *et al.*, 2019, 2020, 2022). There are two primary scenarios in which defense methods are implemented: white-box, where the model parameters and architectures are known, and black-box, where only input queries and output feedback are available. Prior certified defense methods have primarily been executed in a white-box setting (Cohen *et al.*, 2019; Raghunathan *et al.*, 2018; Salman *et al.*, 2022; Rumezhak *et al.*, 2023) or have employed surrogate models as a proxy for the target model (Salman *et al.*, 2020; Nayak *et al.*, 2023) (refer to Adversarial Defense block in Figure 1.2). However, a recent proposal by (Zhang *et al.*, 2022*d*) suggested a black-box defense for DNN models using a zeroth-order optimization approach, which relied solely on model queries and output feedback and did not involve surrogate models. Nonetheless, this approach's efficacy was constrained to low-dimensional datasets. We aim to implement a certified black-box defense using ZO optimization techniques. This defense method only relies on model queries and output feedback for obtaining information, enabling it to achieve high performance even with high-dimensional datasets (Figure 1.6). Our objective is to design a robust UNet denoiser (RDUNet) for the target model and then prepend it with our custom-trained autoencoder to suggest the ZO-AE-RUDS defense mechanism. Unlike the state-of-the-art (SOTA) method proposed by (Zhang *et al.*, 2022*d*), which resulted in high model variance on direct application of ZO optimization to the custom-trained denoiser, our RDUNet reduces model variance due to its architectural advantages over previous denoisers and performs better with direct application of ZO optimization.

1.3 Dissertation Organization

The subsequent chapters of this thesis are organized as follows.

Chapter 2 discusses the literature survey for generative and adversarial learning techniques for Re-ID and classification tasks.

Chapter 3 presents a novel Individual-preserving and Environmental-switching cyclic generation network (IPES-GAN) for unsupervised domain adaptation person Re-ID to disentangle environment and identity-related feature space between the source and target domains so as to preserve the identity-related cues of source domain image while adapting to the cross-domain environment.

Chapter 4 introduces Wasserstein metric for adversarial attack on Re-ID. We iteratively perturb the query images by performing l_∞ perturbation as the first step and then projecting the adversarial sample in the Wasserstein ball of radius ϵ followed by clamping so that perturbation lies in $[0, 1]$ pixel space.

Chapter 5 proposes a robust meta perturbed stochastic neural network (MP-SNN) for defense against adversarial attacks in object Re-ID task. Our MP-SNN learns both anisotropic and isotropic noise distributions in a meta-learning framework.

Chapter 6 presents a certified black-box defense mechanism based on the preprocessing technique of pre-pending a robust denoiser to the predictor to remove adversarial noise using only the input queries and the feedback obtained from the model. We design a novel robust UNet denoiser RDUNet which defends a black-box model with ZO optimization approaches.

Chapter 7 concludes the thesis by summarizing the key contributions in the field of generative and adversarial learning, and discusses the future research work.

1.3.1 Contributions of this research

Generative learning and adversarial attack and defense have emerged as powerful techniques for improving the accuracy and robustness of re-identification and classification tasks in computer vision. In recent years, researchers have developed a range of generative models that can create synthetic data to augment existing datasets, thereby improving the accuracy of models trained on limited data. At the same time, adversarial attack and defense techniques have been developed to identify and defend against attacks on these models, making them more robust and reliable. This thesis provides a comprehensive review of recent research on generative learning and adversarial attack and defense in the context of re-identification and classification tasks, highlighting their contributions and potential applications. It also discusses open challenges and future directions for research in this area. The detailed investigation of each of the techniques suggested above lead to the following contributions of this thesis:

1. We propose a novel deep neural network (DNN) for unsupervised domain adaptation person Re-ID to disentangle environment and identity-related feature space between the source and target domains so as to preserve the identity-related cues of source domain image while adapting to the cross-domain environment. We jointly optimize our generative (disentangling) and discriminative (adaptation) modules. We introduce a cross-domain cyclic generation framework to achieve effective disentanglement and adaptation of appearance and environment features.

2. The DNNs learned in the above work are vulnerable to adversarial attacks. This motivated us to propose a new approach for adversarial attacks on Re-ID using the Wasserstein metric. Our method involves iteratively altering the query images through l_∞ perturbation as the initial step. The perturbed sample is then projected within the Wasserstein ball of radius ϵ and subsequently clamped to ensure that the perturbation remains within the [0,1] pixel space.

3. In order to further improve the robustness of DNN, we introduce a robust defense algorithm against adversarial attacks in re-identification task. Our defense learns both anisotropic and isotropic noise distributions in a meta-learning framework. We derive a novel feature covariance alignment loss which ensures high clean performance while providing robustness against wide variety of adversarial attacks.

4. We further introduce a certified black-box defense mechanism based on the preprocessing technique of pre-pending a robust denoiser to the predictor to remove adversarial noise using only the input queries and the feedback obtained from the model. Our UNet-based robustification model gives high performance for both low-dimensional and high-dimensional datasets.

1.3.2 Applications and Future Work

The applications of generative learning, adversarial attack and defense for Re-ID and classification tasks are extensive and wide-ranging. Generative models can synthesize high-quality data, improving the accuracy and efficiency of classification tasks while also facilitating data augmentation for training models. Adversarial attacks are a significant challenge in these tasks, but developing effective defense mechanisms can improve the robustness and security of these models. Future work in this area could include the development of more advanced generative models, the exploration of different types of adversarial attacks, and the creation of more sophisticated defense mechanisms. Furthermore, the integration of these techniques with other machine learning approaches, such as deep learning and reinforcement learning, may lead to even more powerful and accurate models for Re-ID and classification tasks.

CHAPTER 2

Related Work

2.1 Introduction

The key contribution of this thesis is the development of generative and adversarial learning algorithms for Re-ID and classification tasks. In this chapter, we provide a summary of the overall development of Re-ID and classification tasks and then explain the classical techniques followed by an overview of the popular generative and adversarial learning techniques. We discuss generative learning which covers generation-based, feature learning-based, and unsupervised domain adaptation based methods for person Re-ID task in Section 2.2. Further, in Section 2.3, we discuss the works of adversarial learning which covers adversarial attack and defense for Re-ID and classification tasks.

2.1.1 Evolution of fine grained object recognition (Re-ID) and Classification

Re-ID and classification has undergone significant evolution in recent years due to advancements in computer vision and machine learning (Ballard, 1981; Lowe, 1987; Lamdan *et al.*, 1988; Crevier and Lepage, 1997; Huang and Russell, 1997; Wang, 2013; LeCun *et al.*, 1998*a*). In the past, object recognition was based on hand-crafted features and template matching, which limited its effectiveness in handling variations in object appearance, pose, and lighting conditions. However, with the rise of deep learning and convolutional neural networks (CNNs), object recognition has become more accurate and robust (Krizhevsky *et al.*, 2012; Girshick *et al.*, 2014; LeCun *et al.*, 2015; He *et al.*, 2016; Wang *et al.*, 2018*b*). This has led to the development of advanced algorithms that can recognize and re-identify objects across different environments and conditions, including those with occlusions or partial views. Additionally, the incorporation of transfer learning and unsupervised learning methods has enabled the creation of models that can generalize well to new and unseen data, making object Re-ID and classification more practical and applicable to real-world scenarios (Tzeng *et al.*, 2014;

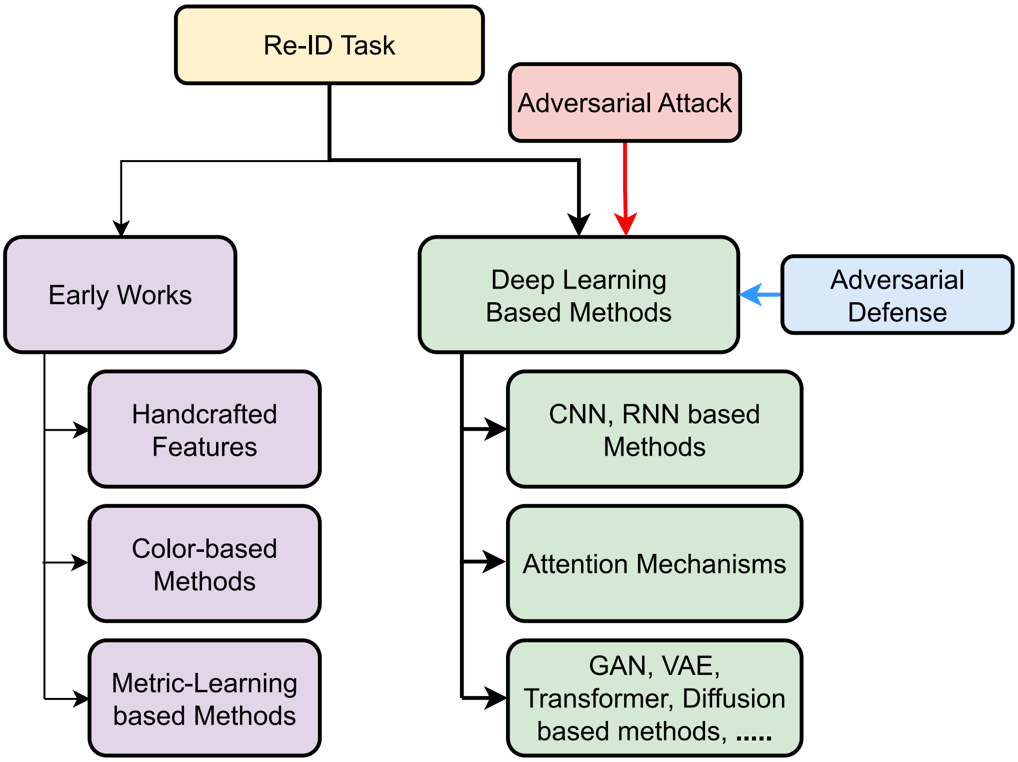


Figure 2.1: It illustrates the development from early works to deep learning-based methods for Re-ID task and highlights their susceptibility to adversarial attacks. In order to address the robustness and safeguarding of target models, researchers have proposed adversarial defense techniques.

Sun and Saenko, 2016; Bousmalis *et al.*, 2017; Hoffman *et al.*, 2017; Lin *et al.*, 2018; Lv *et al.*, 2018; Yang *et al.*, 2020). In the following, we discuss the advances for Re-ID and classification tasks in detail.

Advances for Re-ID

We discuss below the early works, deep learning based methods and adversarial attacks and defense methods for the Re-ID task in detail.

Early Works. One of the first definitions of Re-ID was given by (Plantinga, 1961). Since then it has been studied in many research works and documentation (Rorty, 1973; Cocchiarella, 1977). Re-ID is a challenging task in computer vision, which involves identifying individuals/vehicles across different cameras or in different scenarios. The early representative methods for object Re-ID are simple feature extraction techniques (Ballard, 1981; Lowe, 1987; Lamdan *et al.*, 1988; Crevier and Lepage, 1997). Motivated by these works many complex feature extraction methods were proposed, such as texture and color-based approaches (Lowe, 2004; Bay *et al.*, 2006; Dalal and Triggs, 2005; Soo, 2014), multi-camera tracking (Huang and Russell, 1997; Wang, 2013), hand-crafted features (Hirzer *et al.*, 2012; Zhao *et al.*, 2014; Liao *et al.*, 2015; He *et al.*, 2019*a*) (refer to Early Works block in Figure 2.1). (Gray and Tao, 2008) proposed viewpoint invariant pedestrian recognition that learns a similarity function from a set of training data and utilizes an ensemble of localized features. (Farenzena *et al.*, 2010) proposed a method for Re-ID using color and texture information from clothing. (Zheng *et al.*, 2011) presented a method for Re-ID based on a sparse representation of image features.

Deep Learning Based Methods. The main drawback of the earlier works on Re-ID is their limited capacity to extract high-level features from complex and diverse data sources, such as images and videos. These methods typically rely on handcrafted features, which require significant manual effort and may not be suitable for all types of data. Moreover, they may not be able to capture complex relationships among features, leading to reduced accuracy. In contrast, deep learning methods have shown significant promise in automatically learning complex features and relationships from large-scale datasets, leading to improved performance in Re-ID tasks. In recent years, several pioneering and effective generative and adversarial techniques are proposed for object Re-ID (Zheng *et al.*, 2015, 2016*c*; Liao *et al.*, 2015; Xiao *et al.*, 2016; Dong *et al.*, 2016; Zhou *et al.*, 2017; Yang *et al.*, 2017; Zhang *et al.*, 2018*b*; Li *et al.*, 2018*c*; Liu *et al.*, 2018*d*; Bai *et al.*, 2018; Yang *et al.*, 2018; Zhou *et al.*, 2019; Qi *et al.*, 2019; Wang *et al.*, 2019*a*; Wei *et al.*, 2019; Zhu *et al.*, 2019; Zhang *et al.*, 2020*a*; Zhao *et al.*, 2021*a*; Gu *et al.*, 2022; Wu *et al.*, 2023) (refer to Deep Learning Based Methods block in Figure 2.1).

(Zheng *et al.*, 2015) proposed a scalable method using a combination of local features and global features. (Zheng *et al.*, 2016*c*) introduced a discriminatively learned deep metric that is optimized for ranking accuracy. (Zhou *et al.*, 2017) proposed a deep learning method that uses a multi-scale feature aggregation approach to improve performance. (Liu *et al.*, 2018*d*) presented a large-scale dataset for Re-ID and evaluated several state-of-the-art methods on the dataset. (Yang *et al.*, 2017) presented a method for vehicle Re-ID that uses a Siamese network architecture and a re-ranking algorithm to improve accuracy. (Wei *et al.*, 2019) introduced a method for vehicle Re-ID using a hierarchical feature aggregation approach that captures both global and local features. (Wu *et al.*, 2023) proposed a camera-aware representation learning to address the issue of camera-imbalanced data distribution.

Adversarial Attacks and Defenses. Adversarial attack and defense are two opposing techniques in the field of artificial intelligence and machine learning (Szegedy *et al.*, 2013; Goodfellow *et al.*, 2014*b*; Kurakin *et al.*, 2016; Madry *et al.*, 2017; Ren *et al.*, 2020) as shown in Figure 2.1. Adversarial attack is when input data is manipulated to produce incorrect or unexpected results, while adversarial defense protects machine learning models against these attacks. Adversarial attacks can take many forms, from simple modifications to exploiting model vulnerabilities, with serious consequences in areas such as surveillance, self-driving cars and medical diagnosis (Xie *et al.*, 2017; Dong *et al.*, 2018*b*; Zhang and Wang, 2019; Liu *et al.*, 2019*a*; Qiu *et al.*, 2019; Tu *et al.*, 2020; Saha *et al.*, 2020; Li *et al.*, 2021*d*, 2023). Adversarial defense aims to develop robust models that can withstand these attacks (Xu *et al.*, 2020; Kim *et al.*, 2021; Liu *et al.*, 2022; Liu and Hu, 2022; Deb *et al.*, 2023).

Adversarial attacks involve modifying the input data to fool a Re-ID system into misidentifying a person. These attacks can be achieved by adding noise or perturbations to the input image, altering the illumination, or changing the pose or background of the person in the image (Zheng *et al.*, 2018; Oh *et al.*, 2018; Wang *et al.*, 2019*c*; Bai *et al.*, 2020*b*; Bouniot *et al.*, 2020).

Adversarial defenses, on the other hand, aim to improve the robustness of Re-ID models against such attacks. These defenses can include adding adversarial training data to the training set, using adversarial examples to fine-tune the model, or applying feature-level transformations to enhance the feature representation (Bai *et al.*, 2020*b*; Bouniot *et al.*, 2020; Rice *et al.*, 2020; Jin *et al.*, 2022; Li *et al.*, 2022; Dong *et al.*, 2022; Gong *et al.*, 2022). Overall, understanding the vulnerabilities of Re-ID models to adversarial attacks and developing effective defenses are essential for ensuring the reliability and security of Re-ID systems.

Advances for Classification

We discuss below the early works, deep learning based methods and adversarial attacks and defense methods for the classification task in detail.

In the early 1990s, neural networks were used for image classification, but they suffered from overfitting and required a large amount of labeled data (LeCun *et al.*, 1998*a*).

In 2012, (Krizhevsky *et al.*, 2012) introduced AlexNet, a deep convolutional neural network (CNN), which achieved state-of-the-art performance on the ImageNet dataset. Since then, deep learning-based methods have become the dominant approach for image classification (Goodfellow *et al.*, 2014*a*; Mirza and Osindero, 2014; Radford *et al.*, 2016*a*; Gatys *et al.*, 2016; Salimans *et al.*, 2016). With the dominance of deep learning-based methods for image classification, various adaptations such as the Conditional GAN (Mirza and Osindero, 2014) and Auxiliary Classifier GAN (ACGAN) (Odena *et al.*, 2017) were developed, with DCGANs (Radford *et al.*, 2016*a*) serving as the underlying architecture for these models.

Recent advances in generative and adversarial learning for image classification include the use of attention mechanisms (Li *et al.*, 2017; Zhang, 2018; Wu *et al.*, 2018*b*). Other recent work has focused on using generative models for unsupervised representation learning, which can help improve classification performance on limited labeled data (Donahue *et al.*, 2019; Kolesnikov *et al.*, 2019; Chen *et al.*, 2020*b*). Despite the significant progress made in the field of generative and adversarial learning for image classification, there are still several challenges and limitations that need to be addressed. These include issues with model interpretability and fairness, as well as difficulties in scaling these models to large datasets (Lipton *et al.*, 2018; Gong *et al.*, 2021).

Adversarial Attack and Defense. In the field of image classification, deep learning-based models have become the state-of-the-art for achieving high accuracy rates. However, these models are susceptible to adversarial attacks, which are intentional modifications to the input data designed to mislead the model's classification decision (Szegedy *et al.*, 2013; Goodfellow *et al.*, 2014*b*; Papernot *et al.*, 2016*b*; Carlini and Wagner, 2017; Dong *et al.*, 2018*a*; Eykholt *et al.*, 2018; Su *et al.*, 2019) which motivated adversarial defense works to provide robustness against these attacks (Szegedy *et al.*, 2013; Papernot and McDaniel, 2016; Kurakin *et al.*, 2016).

2.2 Generative Learning

Generative learning is a subfield of machine learning that focuses on modeling the distribution of input data in order to generate new samples that are similar to the original data. There are several generative learning techniques, including GANs (Brock *et al.*, 2018; Goodfellow *et al.*, 2020), VAEs (Kingma and Welling, 2013; Rezende and Mohamed, 2015), autoregressive (Gregor *et al.*, 2015; van den Oord *et al.*, 2016), flow-based (Dinh *et al.*, 2014; Kingma and Dhariwal, 2018) and diffusion-based models (Kingma *et al.*, 2021; Rombach *et al.*, 2022; Song *et al.*, 2023).

Generative Adversarial Network. Generative Adversarial Networks (GANs) were introduced as a framework for learning generative models (Goodfellow *et al.*, 2014*a*). However, early GAN models suffered from issues such as mode collapse, where the generator produces limited variations of the same image, and instability during training. The integration of GANs with deep neural networks (DNNs) has led to significant improvements in image classification performance. For example, the Deep Convolutional Generative Adversarial Network (DCGAN) model achieved state-of-the-art results in image generation tasks (Radford *et al.*, 2016*b*). Since then, GANs have been widely used for image synthesis tasks, such as image-to-image translation (Isola *et al.*, 2017) and style transfer (Radford *et al.*, 2016*a*; Gatys *et al.*, 2016). Conditional GANs (cGANs) were introduced, which generated images based on a given condition, such as class labels (Mirza and Osindero, 2014). (Salimans *et al.*, 2016) proposed a method called Improved GAN (iGAN), which improved the stability of GAN training.

Variational Autoencoder. Another generative model is the Variational Autoencoder (VAE), which is based on the idea of learning the latent representation of data by minimizing the reconstruction error and maximizing the divergence between the prior distribution and the latent distribution (Kingma and Welling, 2013; Pu *et al.*, 2020; Khan *et al.*, 2021). Recently, advances in diffusion based generative models have led to SOTA performance on geenration-based tasks. (Kingma *et al.*, 2021) developed a versatile range of generative models based on diffusion. It incorporates Fourier features into the diffusion model and employes a learnable specification of the diffusion process along with other novel modeling techniques. (Song *et al.*, 2023) proposed a novel inverse problem solver that utilizes unconditional diffusion models. It does not require expensive problem-specific training yet achieves competitive quality on various tasks.

Domain Adaptation. Domain adaptation in generative learning is a challenging problem that has received significant attention in the machine learning community. It refers to the ability of a generative model to adapt to new domains or data distributions, which is critical in real-world applications where the data distribution may change over time. Various domain adaptation techniques have been proposed, including domain adversarial training, cycle-consistent adversarial networks, variational autoencoders with

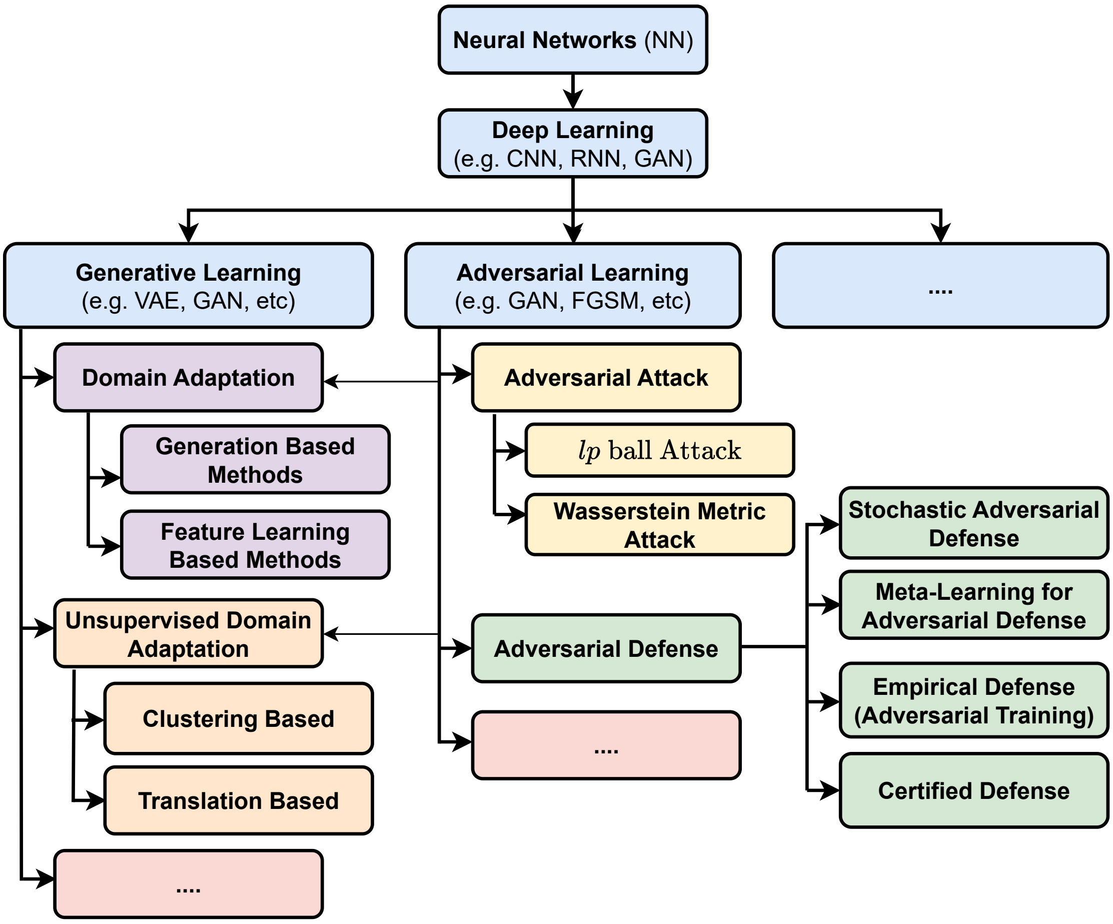


Figure 2.2: The thesis focuses on the progression from neural networks to deep learning techniques and ultimately, to generative and adversarial learning approaches for Object Recognition. The central emphasis is on unsupervised domain adaptation, as well as adversarial attack and defense within the realm of generative and adversarial learning methods.

domain-specific encoders, trasnformers and diffusion models (Ganin and Lempitsky, 2015; Tzeng *et al.*, 2017; Zhu *et al.*, 2017; Choi *et al.*, 2020; Chen *et al.*, 2021; Zhang *et al.*, 2023). These techniques aim to learn a mapping between the source and target domains by leveraging either the discriminative or the reconstruction capability of the generative model.

Domain adaptation for Re-ID task in generative learning is an important research area that has been gaining attention in recent years (Tzeng *et al.*, 2014; Long *et al.*, 2015; Ganin *et al.*, 2016; Bak *et al.*, 2018; Deng *et al.*, 2018; Wang *et al.*, 2018*a*; Tang *et al.*, 2019; Mekhazni *et al.*, 2020; Bai *et al.*, 2021*c*; Mohanty *et al.*, 2022; Mekhazni *et al.*, 2023). In this section, we discuss the generation based and feature learning based methods for cross-dataset person Re-ID task. Further, we discuss the generative learning applied in unsupervised domain adaptation for Re-ID.

2.2.1 Domain Adaptation for Re-ID task

We discuss below the generation-based and feature learning based domain adaptation methods for Re-ID task.

Generation-based Techniques. (Bak *et al.*, 2018) propose a domain adaptation technique taking advantage of the synthetic data. This dataset comprises diverse illumination conditions which efficiently guide the training of the model. (Wei *et al.*, 2018) propose a person transfer GAN with the style transfer and person identity preservation to bridge the domain gap. However, it results in mode collapse as multiple style transfer is performed using a single mapping function. (Deng *et al.*, 2018) defines learning via translation framework to maintain the identity of the person and incorporate domain differences between the two domains.

(Tang *et al.*, 2019) propose C^{2}GAN which applies key-points as weak supervision where key-points are not only used as an input such as in PG^2 (Ma *et al.*, 2017) and DSCF (Siarohin *et al.*, 2018), but also act as a generative object. The key-point generation via key-point cycles boosts the quality of generated images. (Ma *et al.*, 2018) proposed to disentangle input image into three different features, foreground, background, and pose. Further, a mapping function is learned to map the latent space to the feature embedding space, which is then decoded into real image space. In order to exploit the rich variations in an open surveillance setup, (Chen *et al.*, 2019) proposed to use a context rendering GAN (CR-GAN). The target instances guide the generation of a large number of source instances with diverse target domain contexts. (Huang *et al.*, 2022) proposed a new lifelong learning framework that combines meta-learning and continual learning to enhance the generalization and adaptability of deep neural networks.

Feature Learning Based Methods. (Wang *et al.*, 2018*a*) introduced a joint attribute and identity learning model for transferable representation. (Xiao *et al.*, 2016) propose a framework for learning deep feature representations from multiple domains. Some neurons learn representations shared across several domains, while others are effective only for a specific one.

(Long *et al.*, 2015) learn transferable features with statistical guarantees by using an optimal multi-kernel selection method to reduce the domain discrepancy. (Ganin *et al.*, 2016) exploit the idea that predictions of an effective domain transfer cannot discriminate between the source and target domains. (Xie *et al.*, 2018) learn semantic representations for unlabeled target samples by aligning labeled source centroid and pseudo-labeled target centroid. (Zheng *et al.*, 2022*b*) proposed a novel cross-domain alignment method in the homogeneous distance space, which is constructed by the newly designed stair-stepping alignment (SSA) matcher.

2.2.2 Unsupervised Domain Adaptation

Unsupervised domain adaptation is a popular approach in generative learning for Re-ID tasks. In an unsupervised domain adaptation scenario, labeled data is not available in the target domain, but abundant labeled data exists in a related source domain. To address this challenge, unsupervised domain adaptation methods attempt to learn a mapping between the source and target domains that can generalize well to the target domain (Tzeng *et al.*, 2014; Sun and Saenko, 2016; Bousmalis *et al.*, 2017; Hoffman *et al.*, 2017; Lin *et al.*, 2018; Lv *et al.*, 2018; Yang *et al.*, 2020). These methods involve training a model on the labeled source domain data and then adapting it to the target domain without any supervision.

(Sun and Saenko, 2016) propose to learn the non-linear transformation to correlate activations in source and target domains. In (Tzeng *et al.*, 2014), the domain gap between two datasets is reduced by optimizing domain invariance. It uses a soft label distribution matching loss to transfer information between tasks. (Bousmalis *et al.*, 2017) proposed an approach to transform the pixel space from source to target domain in an unsupervised way. (Hoffman *et al.*, 2017) proposed a model that adapts both to the feature level and pixel level. It enforces cycle consistency with task loss and does not require aligned pairs. (Lin *et al.*, 2018) propose an approach considering that the source and target datasets share the same mid-level attributes. They optimize their model using identity classification and attribute learning tasks. To reduce the domain gap between two datasets, it uses MMD distance for mid-level feature distributions.

Unsupervised Domain Adaptation for Re-ID Task. (Lv *et al.*, 2018) propose a model to learn Spatio-temporal patterns of the target data with the visual classifier trained on source data. It improves the classifier using both Spatio-temporal information and visual features. In (Zhong *et al.*, 2018*a*), a homogeneous and heterogeneous model is proposed. This model incorporates camera invariance by sampling an image and its camera style's positive pair from the same dataset. Similarly, it samples negative pairs from source and target datasets heterogeneously. In (Yang *et al.*, 2020), a multi-branch network is proposed, which learns the local and global features from labeled source data. Each network layer defines an unsupervised domain adaptation constraint, which aligns labeled source data's local and global features to unlabeled target data. (Ding *et al.*, 2020) propose a non-parametric classifier with a feature memory that maximizes the distance between all person images by treating each person as a separate class. It minimizes the distance between similar person images. In (Cheng *et al.*, 2022), a novel and robust network model named unsupervised domain adaptation hierarchical person re-identification network (H-Net) is proposed, which not only effectively reduces the impact of inaccurate identification of the hardest sample but also treats different positive samples differently by hierarchical feature collection.

In this thesis, we investigate the unsupervised person Re-ID, aiming to design a Re-ID model for the unlabeled target data by mitigating the gap between the source and target domains. In the next section, we discuss the state-of-the-art adversarial attack and defense algorithms for Re-ID and classification tasks.

2.3 Adversarial Learning

Adversarial learning is a rapidly growing sub-field of machine learning where a model is trained to generate samples that can fool another model called the discriminator (we discussed it in detail in Section 2.2). More recently, adversarial learning field has also been focused on studying adversarial attacks and developing techniques to protect against them. Adversarial attacks are a type of threat that involves introducing carefully crafted perturbations to input data, with the goal of causing a deep learning model to produce erroneous or unexpected outputs (Szegedy *et al.*, 2013; Goodfellow *et al.*, 2014*b*; Papernot *et al.*, 2016*b*; Carlini and Wagner, 2017; Dong *et al.*, 2018*a*; Eykholt *et al.*, 2018; Su *et al.*, 2019; Chen *et al.*, 2020*c*; Liu and Li, 2022; Zhou *et al.*, 2023).

Adversarial defense techniques, on the other hand, aim to enhance the robustness of these models against such attacks. A number of different adversarial defense techniques have been proposed, including adversarial training, where the model is trained on adversarial examples in addition to clean data, and defensive distillation, where a secondary model is trained to detect adversarial examples (Szegedy *et al.*, 2013; Papernot and McDaniel, 2016; Kurakin *et al.*, 2016; Tramèr *et al.*, 2017; Madry *et al.*, 2018; Dhillon *et al.*, 2018; Xie *et al.*, 2018; Yang *et al.*, 2022; Liu *et al.*, 2023). Adversarial pruning is another technique that has been shown to improve model robustness by removing the model's vulnerability to adversarial attacks (Liu *et al.*, 2018*a*, 2019*d*; Ding *et al.*, 2019; Wang *et al.*, 2020*b*; Jian *et al.*, 2022). The ongoing race between adversarial attacks and defense techniques has led to significant advances in the field of adversarial learning, and it remains an active area of research. In this section, we discuss in detail the adversarial attack and defense methods for Re-ID and classification tasks.

2.3.1 Adversarial Attacks

An adversarial attack is a type of attack on a machine learning model where an attacker intentionally introduces small perturbations to the input data in order to cause the model to make incorrect predictions. Let $\boldsymbol{\theta}$ be the parameters of the machine learning model, $\boldsymbol{x}$ be the input to the model, $\boldsymbol{y}$ are the labels corresponding to $\boldsymbol{x}$ and $L(\theta, x, y)$ is the loss function used to train the neural network. Then, adversarial sample $\tilde{\boldsymbol{x}}$ is obtained by adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the loss function with respect to $\boldsymbol{x}$. It is given as,

$$\tilde{x} = x + \epsilon \cdot sign(\nabla_x L(\theta, x, y)) \tag{2.1}$$

where ϵ is a small constant. Much adversarial attacks are performed in a closed-set setting to fool deep networks (Szegedy *et al.*, 2013; Kurakin *et al.*, 2016; Dong *et al.*, 2018*b*). The goal of such attacks is to undermine the reliability and trustworthiness of machine learning systems, and they can have serious consequences in a variety of applications, from image and speech recognition to autonomous vehicles and security systems (Moosavi-Dezfooli *et al.*, 2016; Carlini and Wagner, 2017; Athalye *et al.*, 2018*a*; Chen and Gu, 2020; Tashiro *et al.*, 2020; Li *et al.*, 2021*d*, 2023).

White-box Setting. (Szegedy *et al.*, 2013) proposed one-step gradient based method to generate adversarial samples by adding the sign of gradient of image pixels to original image. Many other works have been proposed since then, such as FGSM (Goodfellow *et al.*, 2014*b*), iterative FGSM (Kurakin *et al.*, 2018), PGD (Madry *et al.*, 2017). (Kurakin *et al.*, 2016) proposed an iterative method of step gradient method. (Dong *et al.*, 2018*b*) further proposed a method which uses momentum in iterative step gradient method. Later on, various methods were proposed for robust attack performance or efficiency (Moosavi-Dezfooli *et al.*, 2016; Carlini and Wagner, 2017; Athalye *et al.*, 2018*a*; Chen and Gu, 2020; Tashiro *et al.*, 2020).

Recently (Wong *et al.*, 2019; Hu *et al.*, 2020) proposed classification threat models based on Wasserstein distance which corresponds to mass pixel movement, more generalized to real world images and thus can be generalized to attack robust models, instead of l_∞ perturbations which depend on pixel independence.

Black-box Setting. Some other works are proposed in a more practical setting where no information is available about the target model called as black-box attack (Su *et al.*, 2019; Yuan *et al.*, 2021; Sun *et al.*, 2022). Recently many black-box attacks are based on transfer-based attacks (Wu *et al.*, 2020*b*; Li *et al.*, 2020; Zhang *et al.*, 2022*a*), (Wang *et al.*, 2021*c*) proposed generation-based substitute training strategy for improved data-free black-box attacking performance. (Sun *et al.*, 2022) proposed a triplet-player traditional data free framework for black-box adversarial attack. Recently many black-box attacks are based on transfer-based attacks (Wu *et al.*, 2020*b*; Li *et al.*, 2020; Zhang *et al.*, 2022*a*). Some previous works are query specific which update the perturbations based on the features of query images (Andriushchenko *et al.*, 2020; Chen *et al.*, 2020*a*). However, they are all based on l_p threat model which may not generalize well to real world images.

In a closed-set problem, a decision boundary can be formed within the feature space and if the threat model predicts a wrong class for an image, its considered a success. However, in the case of Re-ID the whole ranking order is to be changed to design a robust attack model. The open set problem of Re-ID is quite different as their performance is not only dependent on query image but also on gallery images, thus the classification threat models cannot be directly applied for these problems. Re-ID brings new challenges to design a robust attack system which will guide us to improve the performance and accuracy of our current robust Re-ID systems.

Adversarial Attack Against Re-ID

In recent times, there has been a significant amount of research put forth in the area of adversarial attacks on retrieval (Feng *et al.*, 2020; Bai *et al.*, 2020*a*; Li *et al.*, 2021*b*) and object Re-ID tasks (Zheng *et al.*, 2018; Oh *et al.*, 2018; Wang *et al.*, 2019*c*; Bai *et al.*, 2020*b*; Bouniot *et al.*, 2020; Yang *et al.*, 2021*a*; Ding *et al.*, 2021). (Oh *et al.*, 2018) showed that adding adversarial perturbations to images of people can significantly reduce the accuracy of person recognition systems. (Zheng *et al.*, 2018) proposed a method to push the features of adversarial sample away from an artificial guide in the feature space. (Wang *et al.*, 2019*c*) produced the adversarial samples by generating patterns of clothes of query images.

(Bai *et al.*, 2020*b*) proposed an adversarial metric attack to perturb the gallery images by extending FGSM, I-FGSM and MI-FGSM attacks used for classification. However, it does not generalize well to SOTA Re-ID models. (Wang *et al.*, 2020*a*) proposed a generator-based framework to produce perturbation and verify the universality of the generated perturbation. (Bouniot *et al.*, 2020) proposed metric attacks based on pushing guides, pulling guides, or a combination of both guides. (Wang *et al.*, 2020*c*) proposed a black-box adversarial attack on Re-ID models that can fool the model by generating adversarial examples using only its output probabilities. (Guo *et al.*, 2021) proposed an adversarial metric learning approach that can improve the robustness of Re-ID models against adversarial attacks. (Yang *et al.*, 2021*a*) uses a virtual-guided meta-learning scheme to learn universal adversarial perturbation. (Ding *et al.*, 2021) proposed a image agnostic and model insensitive approach for generating a universal adversarial perturbation. Many of the above approaches have poor transferability and visual quality of adversarial samples is poor due to perceptible noise. (Zhao *et al.*, 2022) proposed a novel adversarial patch generative adversarial network (AP-GAN) to generate adversarial patches instead of modifying the entire image.

In this thesis, we are the first to introduce Wasserstein threat model to attack Re-ID system. We have used an adversarial metric to generate an adversarial gallery sample by increasing the distance between clean query and perturbed gallery images.

2.3.2 Adversarial Defense

Adversarial defense refers to the set of techniques and strategies used to defend against adversarial attacks mentioned in Section 2.3.1. In this Section we will first discuss the defense methods on Re-ID and classification tasks in detail:

Adversarial Defense for Re-ID Task

Recently, some works focus on improving the training efficiency of adversarial training based methods (Rice *et al.*, 2020; Jin *et al.*, 2022; Li *et al.*, 2022; Dong *et al.*, 2022), providing defense against, deep metric learning (Zhou *et al.*, 2020; Zhou and Patel, 2022), and open-set Re-ID tasks (Bai *et al.*, 2020*b*; Bouniot *et al.*, 2020; Gong *et al.*, 2022). (Bai *et al.*, 2020*b*) proposed offline adversarial training where adversarial samples are generated with the frozen trained model. (Bouniot *et al.*, 2020) proposed an online adversarial training method where they use images as guides that are sampled during training. These guides are used to generate better adversarial samples to update the model's parameters. (Gong *et al.*, 2022) proposed joint adversarial defense by enhancing the contour and color feature of multi-modality images.

Stochastic Adversarial Defense. Recent works have shown that injecting noise; either fixed or learnable, into the target model improves adversarial robustness. (Liu *et al.*, 2018*c*) improves model robustness by injecting additive spherical Gaussian noise into various layers of the model. (He *et al.*, 2019*b*) proposed Parametric Noise Injection where a learnable intensity parameter controls a fixed spherical noise distribution. (Jeddi *et al.*, 2020) propose Learn2Perturb where a parametric isotropic noise perturbation-injection module is proposed. The parameters of the model and perturbation injection module are updated alternately. (Yu *et al.*, 2021) maximizes the entropy of the learned noise distribution; for this purpose, they introduce fully-trainable stochastic layers. (Eustratiadis *et al.*, 2021) proposed a defense by introducing anisotropic noise in a deep neural network. However, these works are primarily designed for closed-set tasks like classification, and a straightforward adoption in retrieval tasks may not be effective. Our work notably differs from these methods as we propose a learnable combination of anisotropic and isotropic noise modules in an open-set object Re-ID task.

Meta-Learning for Adversarial Defense. Meta-learning is a learning-to-learn concept to generalize to unknown tasks and distribution with limited training samples (Finn *et al.*, 2017; Nichol and Schulman, 2018). Meta-learning has various applications like few-shot learning (Finn *et al.*, 2017; Sun *et al.*, 2019), domain generalization (Li *et al.*, 2019*c*; Guo *et al.*, 2020; Chen *et al.*, 2022*a*), adversarial attack (Feng *et al.*, 2021*b*; Yuan *et al.*, 2021; Fang *et al.*, 2022), adversarial defense (Goldblum *et al.*, 2020), among others. Recently, meta-learning has been used in object Re-ID works (Zhao *et al.*, 2021*b*;

Yang *et al.*, 2021*b*; Bai *et al.*, 2021*b*; Choi *et al.*, 2021; Yang *et al.*, 2021*a*; Ni *et al.*, 2022). Recently, (Yang *et al.*, 2022) proposed meta-learning based defense, however it is computationally expensive as they use an additional dataset to capture variations in cross-domain. In this thesis, we propose a meta-learning scheme to learn a robust defense mechanism against various attacks on the Re-ID task. We chose meta-learning as we can naturally incorporate clean samples, noise parameters, and adversarial samples of a selected attack. And we do not need to train the system with samples from all attacks, which is extremely expensive.

Adversarial Defense for Classification Task

We broadly categorize the literature on robust defense into empirical and certified defense and briefly discuss them below in the white-box and black-box settings.

Empirical Defense (Adversarial Training). (Szegedy *et al.*, 2013) first proposed robust empirical defense in the form of adversarial training (AT). Due to AT, there has been a rapid increase in empirical defense methods (Chan *et al.*, 2019; Wang and Wang, 2022; Yan *et al.*, 2022; Cheng *et al.*, 2023; Wei *et al.*, 2023). (Zhang *et al.*, 2019*b*) proposed a tradeoff between robustness and accuracy that can optimize defense performance. In order to improve the scalability of AT, empirical robustness is provided by previous works which design computationally light alternatives of AT (Carmon *et al.*, 2019; Sehwag *et al.*, 2021). Some recent empirical defense works are based on the concept of distillation, initially proposed by (Hinton *et al.*, 2015). (Papernot *et al.*, 2016*b*) presented a defensive distillation strategy to counter adversarial attacks. (Folz *et al.*, 2020) gave a distillation model for the original model, which is trained using a distillation algorithm. It masks the model gradient in order to prevent adversarial perturbations from attacking the model's gradient information. (Addepalli *et al.*, 2020) proposed unique bit plane feature consistency (BPFC) regularizer to increase the model's resistance to adversarial attacks. (Cheng *et al.*, 2023) proposed to regularize the distributions of different classes to increase the difficulty of finding an attacking direction.

Certified Defense. Unlike empirical defense, the certified defense provides formal verification of robustness of the DNN model (Ma *et al.*, 2021; Gupta *et al.*, 2021; Elaalami *et al.*, 2022). Certified robustness is given by a 'safe' neighbourhood region around the input sample where the prediction of DNN reamins same. Previous works (Katz *et al.*, 2017; Dutta *et al.*, 2017; Tjeng *et al.*, 2017; Bunel *et al.*, 2018) in this field provide 'exact' certification, which is often compute-intensive and is not scalable to large architectures. (Katz *et al.*, 2017) proposed a robust simplex verification method to handle non-linear ReLU activation functions. Another line of work (Wong and Kolter, 2018; Zhang *et al.*, 2018*a*), which provides 'incomplete' verification, requires less computation; however, it gives faulty certification and can decline certification even in the absence of adversarial perturbation. Both 'exact' and 'incomplete' posthoc certification methods require customized architectures and hence are not suitable for DNNs (Zhang *et al.*, 2022*d*).

Another area of study focuses on in-process certification-aware training and prediction. For instance, a randomized smoothing (RS) involves perturbing the input samples with Gaussian noise. This process allows for the transformation of a given base classifier f into a new "smoothed classifier" g, using randomized smoothing. Importantly, this transformation ensures that g is certified to be robust in the L_2 norm. In (Cohen *et al.*, 2019), it was demonstrated that RS could offer formal assurances for adversarial robustness. As well as, there are several different RS-oriented verifiable defences that have been developed, including adversarial smoothing (Salman *et al.*, 2019), denoised smoothing (Salman *et al.*, 2020), smoothed ViT (Salman *et al.*, 2022), and feature smoothing (Addepalli *et al.*, 2021).

2.4 Conclusion

This chapter provides an overview of various generative and adversarial methods that can address the limitations of traditional feature extraction techniques in Re-ID and classification tasks. Generative models trained on a labeled source dataset often underperform when applied to a target dataset, which is typically unlabeled due to the impracticality of annotating large-scale datasets. To address this issue, we explore unsupervised domain adaptation techniques that preserve the appearance of the source domain while adapting to the pose and background of the target domain using a cyclic GAN-based approach. We introduce the use of the Wasserstein metric for adversarial attacks on Re-ID tasks and investigate the effectiveness of noise-perturbed adversarial defense trained using a meta-learning strategy for improving the robustness of these generative DNNs in Re-ID tasks. Additionally, we explore certified black-box defense optimized with zeroth-order optimization for classification tasks.

CHAPTER 3

Unsupervised Domain Adaptation for Person Re-Identification via Cyclic Generation

3.1 Introduction

Unsupervised domain adaptation in person Re-ID refers to the process of adapting a person Re-ID model trained on a source domain to perform well on a target domain, where the target domain has different characteristics, such as different lighting conditions, camera viewpoints, or environmental contexts, without requiring labeled data in the target domain (Fan *et al.*, 2018; Lin *et al.*, 2019; Huang *et al.*, 2018; Bak *et al.*, 2018; Ren *et al.*, 2019; Khatun *et al.*, 2020). The goal is to minimize the domain shift between the source and target domains, while retaining the discriminative power of the model in identifying individuals across domains. This is typically achieved through the following techniques.

Clustering-based Techniques. Clustering-based techniques have emerged as a popular approach for person re-identification (Re-ID) due to their ability to pseudo-label similar images together without requiring labeled data. These methods typically involve partitioning a set of feature vectors into clusters based on some distance metric, such as Euclidean or cosine distance. Once the clusters have been formed, they can be used to identify individuals by matching images from the same cluster (Cai *et al.*, 2018; Zhang *et al.*, 2019*a*; Cheng *et al.*, 2019; Zheng *et al.*, 2021; Zhang *et al.*, 2021; Zheng *et al.*, 2022*a*; Quan *et al.*, 2023).

The self-training scheme with clustering labels was originally proposed by PUL (Fan *et al.*, 2017) and UDAP (Song *et al.*, 2020). SSG (Fu *et al.*, 2019) and PAST (Zhang *et al.*, 2019*c*) subsequently built on this approach by incorporating human part features and implementing a progressive training strategy. More recently, MMT (Ge *et al.*, 2020) has been introduced, which utilizes coupledly trained networks and mean-teacher networks for mutual training, resulting in state-of-the-art performance. However, the automatic label assignment process in the clustering technique may bring in noise when the model trained on the source meets the uncertain style change of unseen target data (Fan *et al.*, 2018; Lin *et al.*, 2019). Furthermore, clustering-based methods typically prioritize utilizing solely the data in the target domain, and do not leverage the valuable labeled data available in the source domain (Ge *et al.*, 2022).

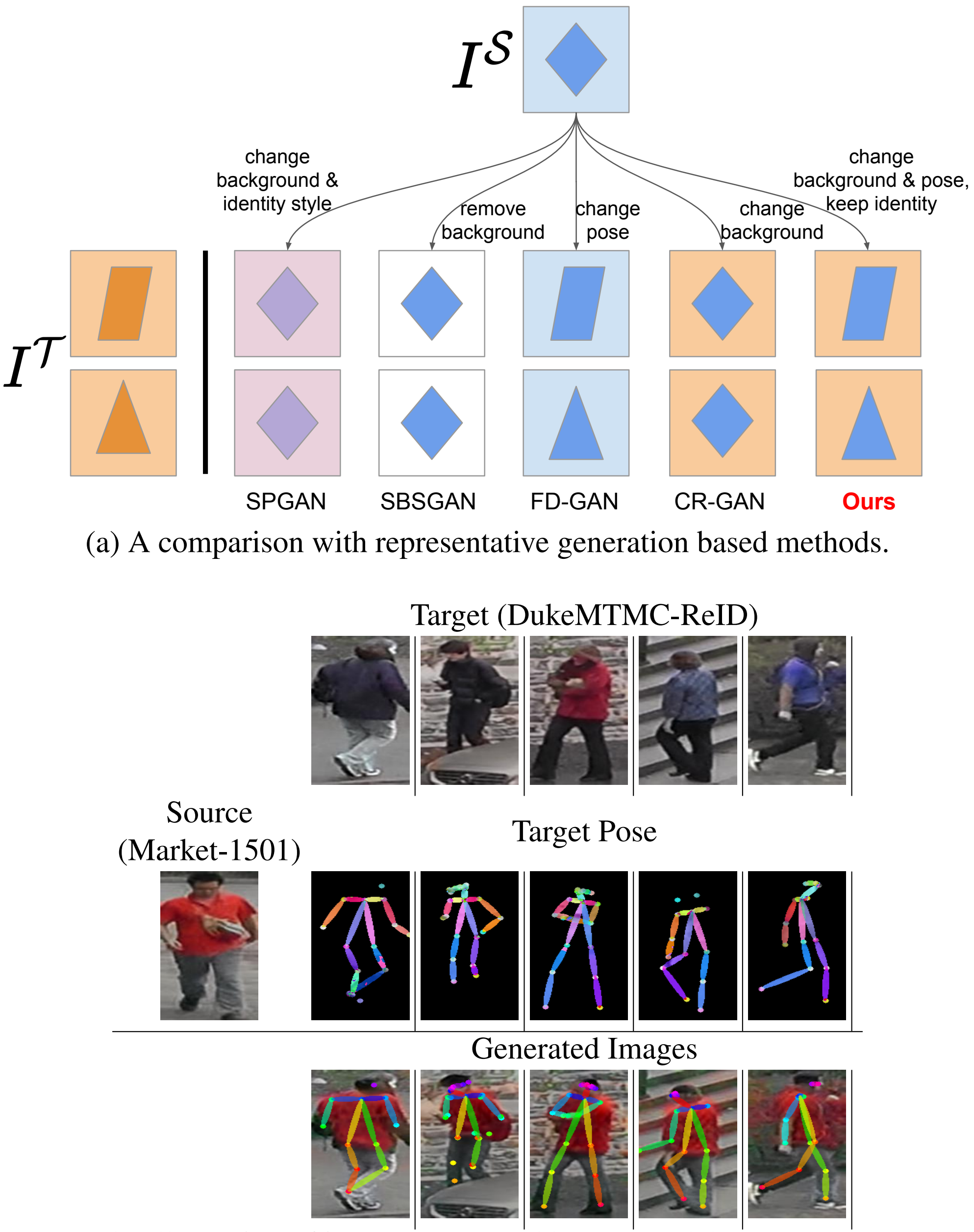
Image Translation. Recently, unsupervised image to image translation and extensive synthetic data using GANs have become popular. Several previous works use this technique to translate source samples to the target style and perform Re-ID by training a CNN with this synthetic data (Wei *et al.*, 2018; Deng *et al.*, 2018; Wang *et al.*, 2018*a*; Huang *et al.*, 2018; Bak *et al.*, 2018; Ren *et al.*, 2019; Khatun *et al.*, 2020). These methods are successful in mitigating the domain drift to some extent. Significant researches in this field are done by reducing the domain gap in terms of camera style change or pose change (Ma *et al.*, 2017; Siarohin *et al.*, 2018; Dong *et al.*, 2018*a*; Song *et al.*, 2019; Li *et al.*, 2019*e*). Some recent works on unsupervised domain adaptation in Re-ID proposed after the publication of this work are (Ge *et al.*, 2022; Ye *et al.*, 2022).

In this chapter, we delve into the topic of unsupervised domain adaptation in person Re-ID. Our goal is to create a Re-ID model that can effectively utilize unlabeled target data by bridging the gap between the source and target domains. To achieve this, we propose a novel cyclic generation network called the Individual-Preserving and Environmental-Switching GAN (IPES-GAN). Our framework utilizes data from both the source and target domains to train the model, and we separate the representation into two parts: the individual part, which retains identity-related features such as clothing color and style, and the environmental part, which includes identity non-related features such as pose and background.

Our network possesses two distinct features: firstly, we utilize decoupled features rather than fused features, which has been proven to be beneficial for generation and adaptation. Secondly, we use cyclic generation instead of one-step adaptive generation. We swap the source and target environment features to generate cross-domain images that maintain identity-related features conditioned with the source (or target) background features. We then change the environment features again to generate the input image, allowing the cyclic generation to run in a self-supervised manner.

Our contributions are summarized as follows:

- We propose a novel Individual-preserving and Environmental-switching cyclic generation network (IPES-GAN) for unsupervised domain adaptation person re-



(a) A comparison with representative generation based methods.

(b) Illustration of images generated by our model IPES-GAN.

Figure 3.1: (a) We make a comparison with four representative generation based methods, including SPGAN (Deng *et al.*, 2018), SBSGAN (Huang *et al.*, 2019*a*), FD-GAN (Ge *et al.*, 2018), and CR-GAN (Chen *et al.*, 2019), by translating the image $I^{\mathcal{S}}$ in the source domain according to the image $I^{\mathcal{T}}$ in the target domain. Different colors of squares indicate different backgrounds, and different shapes in the squares indicate different poses. (b) Illustration of images generated by our model IPES-GAN in the source domain. The source (concatenated with target pose) and the target images are input to our framework. The output is the generated images with the superimposed pose, showing that these images have adapted to the pose of the target images (**Best viewed in colors**).

identification to disentangle environment and identity-related feature space between the source and target domains so as to preserve the identity-related cues of source domain image while adapting to the cross-domain environment.

- We propose a joint optimization of our generative (disentangling) and discriminative (adaptation) modules. We introduce a cross-domain cyclic generation framework to achieve effective disentanglement and adaptation of appearance and environment features.
- We conduct experiments on three datasets - Market-1501 (Zheng *et al.*, 2015), DukeMTMC-ReID (Zheng *et al.*, 2017), and MSMT17 (Wei *et al.*, 2018). We demonstrate that our model adapts well to the target domain using different metrics - CMC, mAP, and MMD. We show that our method can generate high-quality cross-domain synthetic images with better fidelity and diversity with the help of two metrics: LPIPS (Zhang *et al.*, 2018*c*)and FID (Heusel *et al.*, 2017). Our method outperforms SOTA Re-ID in the unsupervised domain adaptation setting.

We make a comparison with four representative works, as shown in Figure 3.1(a). 1) SPGAN (Deng *et al.*, 2018) attempts to translate the source image to a target style, where person and background are treated as a whole. The translated image may obtain a similar target style, but the redundant source background still exists, and the person's appearance may lose its identity detail. 2) SBSGAN (Huang *et al.*, 2019*a*) removes the source background, but the target background is ignored. It will make the Re-ID model less generalizable. On the other hand, persons in different camera views are in various poses. The model using the source pose only in training may not adapt to the pose change in the target domain. 3) FD-GAN (Ge *et al.*, 2018) disentangles the pose information and generates more images with target poses. However, it needs paired image for training and cannot be extended for an unsupervised setting. 4) CR-GAN (Chen *et al.*, 2019) changes the background of the source image to that of the target background. However, similar to SBSGAN, the model can not adapt to the pose change in the target. Compared to these methods, our method has superiority in generating images with target background and pose (environmental) and maintaining relevant Re-ID information. In Figure 3.1(b), we show the augmentation of the target environment features with the source identity-related features. We depict that our network completely adapts to the background, camera style, and pose-related information while preserving the appearance of the source image as much as possible.

3.2 Proposed Methodology

Problem Statement. We study the problem of unsupervised domain adaptation in Re-ID. The input to our model is labeled source domain and unlabeled target domain data. We aim to learn two different mappings from source to target domain - cyclic-image generation and environment-switching generation.

Notation. Let us define some basic notations of our problem. The source dataset is defined as $\mathcal{S} = \{I^{\mathcal{S}}\}^{N_{\mathcal{S}}}$ with $N_{\mathcal{S}}$ number of images and $C_{\mathcal{S}}$ number of classes. $I^{\mathcal{S}} \in \mathbb{R}^{C \times H \times W}$ where C is the number of channels in the image, H is the height and W is the width of the image. The images are labeled as $\mathcal{Y}^{\mathcal{S}} = \{y^{\mathcal{S}}_{id}, y^{\mathcal{S}}_{cam}\}^{N_{\mathcal{S}}}$, where $y^{\mathcal{S}}_{id}$ and $y^{\mathcal{S}}_{cam}$ are the ID and camera ID of source dataset. The target dataset is defined as $\mathcal{T} = \{I^{\mathcal{T}}\}^{N_{\mathcal{T}}}$ with $N_{\mathcal{T}}$ number of total images. The target dataset do not have identity label. Since the camera ID label can be generated automatically, the images are labeled as $\mathcal{Y}^{\mathcal{T}} = \{y^{\mathcal{T}}_{cam}\}^{N_{\mathcal{T}}}$. We have additional information in form of the pose of the image. Pose of source and target images are given the annotation $\mathcal{P}^{\mathcal{S}} = \{P^{\mathcal{S}}\}^{N_{\mathcal{S}}}$ and $\mathcal{P}^{\mathcal{T}} = \{P^{\mathcal{T}}\}^{N_{\mathcal{T}}}$ respectively.

Overview of Our Approach. We show our IPES-GAN framework in Figure 3.2. It illustrates the decoupling of environment and shared appearance feature space. The environment-style adapted cross-domain images are generated with the switching of environment features while preserving the identity-related features. In Section 3.2.1, we perform generative learning to supervise the generation of our cross-domain images by swapping the environment components of the source and target domain in a cyclic manner. We describe the concatenation of the source (target) image with the target (source) pose. Our pose and style encoders use these concatenated images to encode environment features. The identity encoder extracts identity-related features which are shared between source and target domains. In Section 3.2.2, we show the use of these synthetic cross-domain images to train our CNN-based identity encoder, which later performs Re-ID on unseen data.

3.2.1 Generative Learning

Generative learning (generation) is performed by encoder-decoder architecture and a discriminator. We show in Figure 3.3, the encoder is comprised of three distinct en-

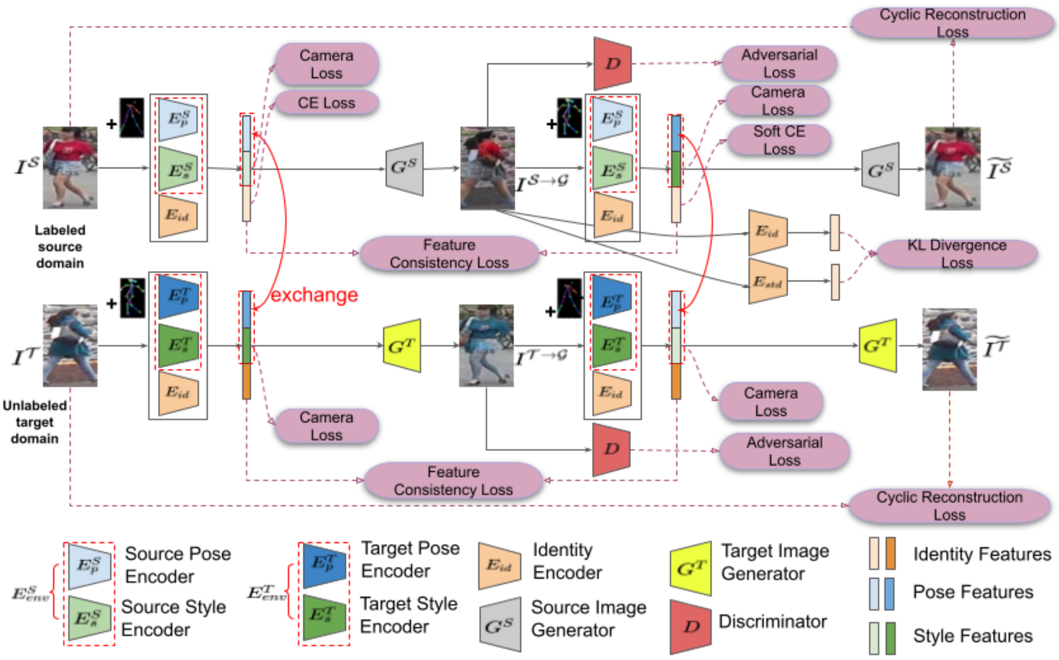


Figure 3.2: The overview of our Individual-preserving Environmental-switching Network (IPES-GAN). Pose encoders help to achieve pose guidance. Style encoders learn the camera-style features. Identity encoder is shared between labeled source and unlabeled target domain. Image generators' (decoders') output are environmental switched images ($I^{\mathcal{S}\to\mathcal{G}}, I^{\mathcal{T}\to\mathcal{G}}$) which are passed again to the generator (encoder-decoder) network to generate cyclic generated images ($\widetilde{I^{\mathcal{S}}}, \widetilde{I^{\mathcal{T}}}$). Discriminator helps to achieve image-refinement. KL Divergence loss is calculated between the probability distribution predicted by our identity encoder and a standard CNN model E_{std} trained on source dataset. (**Best viewed in colors**).

coders: the pose encoders ($E_p^{\mathcal{S}}, E_p^{\mathcal{T}}$), the style encoders ($E_s^{\mathcal{S}}, E_s^{\mathcal{T}}$), and the shared identity encoder E_{id}. Each of these encoders serves the purpose of capturing unique features: pose, style, and identity features, respectively. Pose and camera-style features together form environment features in the source and target domain. Environment features are swapped between the two domains and concatenated with shared appearance features. These cross-domain features are sent to the source (target) decoder to perform generation. Discriminator is used to identify whether the generated image is real or fake. We explain our encoders in detail below.

1. **Pose Encoder.** The input to the source (target) pose encoder is the source (target) image and the target (source) pose in order to capture cross-domain pose features.

   Pose Guided Generation. In order to learn cross-domain pose features, we concatenate the source (target) image $\in R^{C\times H\times W}$ with the target (source) pose $\in R^{C_p\times H\times W}$ at the input of pose encoder, where $p = 18$. The concatenated input $\in R^{(C+C_p)\times H\times W}$ is then fed to the pose encoder. The pose encoder gives us feature vector f_{pose} which represents the cross-domain pose features. We obtain the pose of the images using the state-of-art pose estimator model (Cao *et al.*,

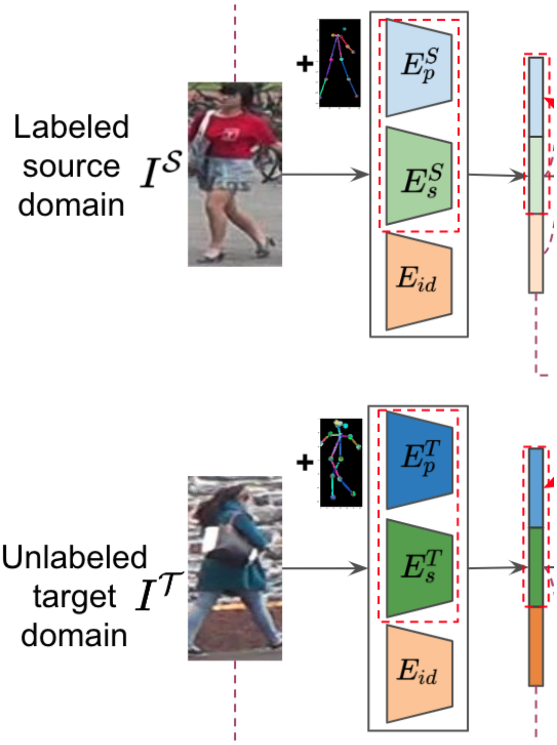


Figure 3.3: Our encoder network has pose encoders (E_p^S, E_p^T), style encoders (E_s^S, E_s^T) and shared identity encoder E_{id}. Source (target) images are concatenated with cross-domain 18-channel pose keypoints are input to the pose encoder to capture pose features. Similarly, we capture source (target) style features and shared identity features. (**Best viewed in colors**).

2018). Even if the pose information is unavailable, our network can generate cross-domain images by swapping the environment features of the source and target domain. However, the pose information allows our model to learn the global structure of the human body to generate high-quality cross-domain images with the pose of source and target images. The pose information increases the flexibility of our model to better adapt to unseen poses and thus can generalize well to unseen person Re-ID images. We show some of the examples generated by our model IPES-GAN in Figure 3.1(b). We superimpose the pose of generated images in the source domain and observe that the poses of generated and target images are very similar. It is visible that our generated images adapt to the target images' pose while preserving identity-related features of the source image.

2. **Style Encoder.** The input is source (target) image. It gives camera-style (background) features of source (target) as output.

3. **Identity Encoder.** The input is the source and target image. It is shared between source and target domains. It gives identity-related features as output. Since labels are present in the source domain, we enforce the identity encoder to capture identity-related information by using the classification loss.

Environmental Swap (exchange). As we show in Figure 3.4, the pose encoder gives us feature vector f_{pose} which represents cross-domain pose features. The style encoder provides the feature vector f_{style} which represents the background and style of the cam-

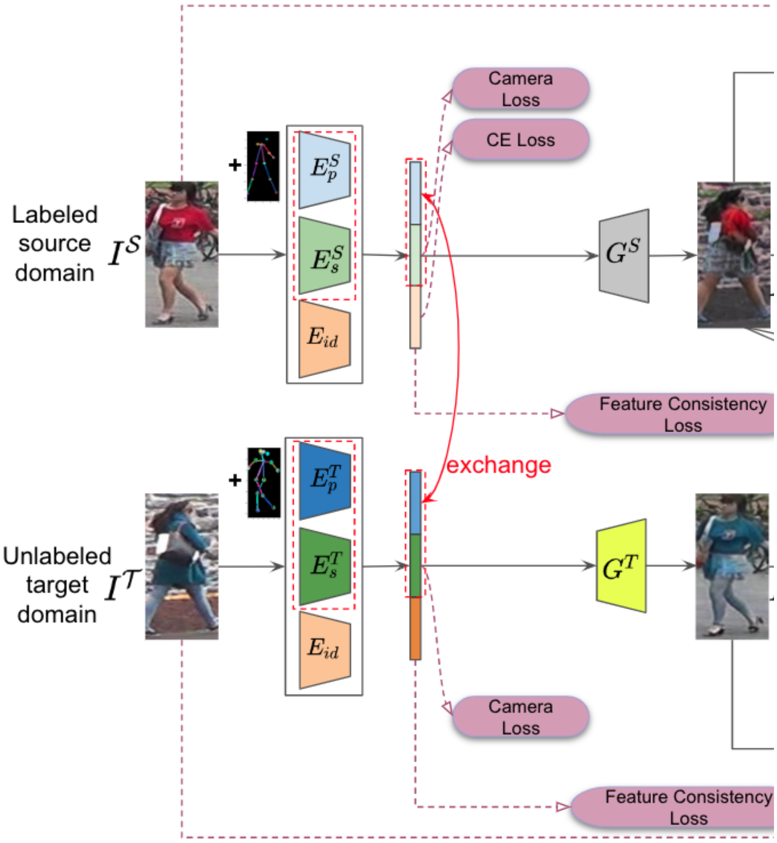


Figure 3.4: The pose and style (environment) features obtained from pose encoders (E_p^S, E_p^T) and style encoders (E_s^S, E_s^T) respectively, are concatenated and exchanged between the source and target domain. We also term this exchange as environmental swap. (Refer Figure 3.2 for legend of different modules). (**Best viewed in colors**).

era. This decoupling of foreground (pose features) and background (style features) forces our generative module to learn respective features separately and efficiently compared to learning features of the image as a whole.

In order to avoid confusion, in the rest of the chapter, pose encoder and style encoders are represented together as environment encoders $E_{env}^{\mathcal{S}}$ and $E_{env}^{\mathcal{T}}$, the environment features are denoted by f_{env}. f_{env} is the concatenation of pose and style features of the source and target image, represented as $f_{env}^{\mathcal{S}} = concat(f_{pose}^{\mathcal{S}}, f_{style}^{\mathcal{S}})$ and $f_{env}^{\mathcal{T}} = concat(f_{pose}^{\mathcal{T}}, f_{style}^{\mathcal{T}})$, where $f_{pose}^{\mathcal{S}}$ ($f_{pose}^{\mathcal{T}}$) and $f_{style}^{\mathcal{S}}$ ($f_{style}^{\mathcal{T}}$) are the private pose and style feature of source (target) image. $f_{env}^{\mathcal{S}}$ ($f_{env}^{\mathcal{T}}$) represents the private environment features of the source (target) image.

Let us consider a pair of images $I^{\mathcal{S}}$ and $I^{\mathcal{T}}$, respectively from the source and target domain. Our aim is to generate a new pedestrian image by swapping the environment features of the source and target images. The new feature distribution is given as

$f^{\mathcal{S}\to\mathcal{G}} = concat(f^{\mathcal{T}}_{env}, f_{ind})$ and $f^{\mathcal{T}\to\mathcal{G}} = concat(f^{\mathcal{S}}_{env}, f_{ind})$, where f_{ind} is the shared identity-related features of source and target image. $f^{\mathcal{S}\to\mathcal{G}}$ ($f^{\mathcal{T}\to\mathcal{G}}$) is the concatenated feature vector of environment features of target (source) images and shared identity-related features of source and target images. These features are fed to the decoders to generate cross-domain images. Taking $G^{\mathcal{S}}$ and $G^{\mathcal{T}}$ as the decoders, we generate images in the source and target domains as $I^{\mathcal{S}\to\mathcal{G}} = G^{\mathcal{S}}(f^{\mathcal{S}\to\mathcal{G}})$ and $I^{\mathcal{T}\to\mathcal{G}} = G^{\mathcal{T}}(f^{\mathcal{T}\to\mathcal{G}})$ respectively.

The intermediate generated image $I^{\mathcal{S}\to\mathcal{G}}$ (Figures 3.2 and 3.4) contains the environment features of the target image while aiming to preserve the identity-related features of the source image. Through this feature decoupling and generation, our model is forced to adapt to the target image's environment while maintaining the source image's individuality. $I^{\mathcal{T}\to\mathcal{G}}$ is the translated image with identity-related features of the target image. Our network generates cross-domain images by swapping the environment features of the source and target domain. The shared identity encoder learns the domain-invariant features and preserves the identity-related features of the two domains. Thus, in the target domain, the generated image $I^{\mathcal{T}\to\mathcal{G}}$ adapts to the pose and background of source image $I^{\mathcal{S}}$ while preserving appearance of target image $I^{\mathcal{T}}$.

Feature Consistency. In order to learn a mapping between two different domains with no information of paired data, we introduce feature reconstruction based on environment and identity-related features (Figure 3.5). These features help us to supervise and generate images in the source (target) domain with the pose and background information of the target (source) image.

As we show in Figure 3.5, the generated image $I^{\mathcal{S}\to\mathcal{G}}$ should maximize the environment information from the target image and appearance information from the source image so that we can generate images with pose and background from the target image while preserving the identity-related features of the source image. If we are successful in attaining this environment switching from target to source domain, we should be able to reconstruct our environment and identity-related features of the source and target domain again:

$$\mathcal{L}^{\mathcal{S}}_{Con\_Env} = \mathbb{E}[||f^{\mathcal{S}}_{env} - E^{\mathcal{S}}_{env}(I^{\mathcal{T}\to\mathcal{G}})||_1], \tag{3.1}$$

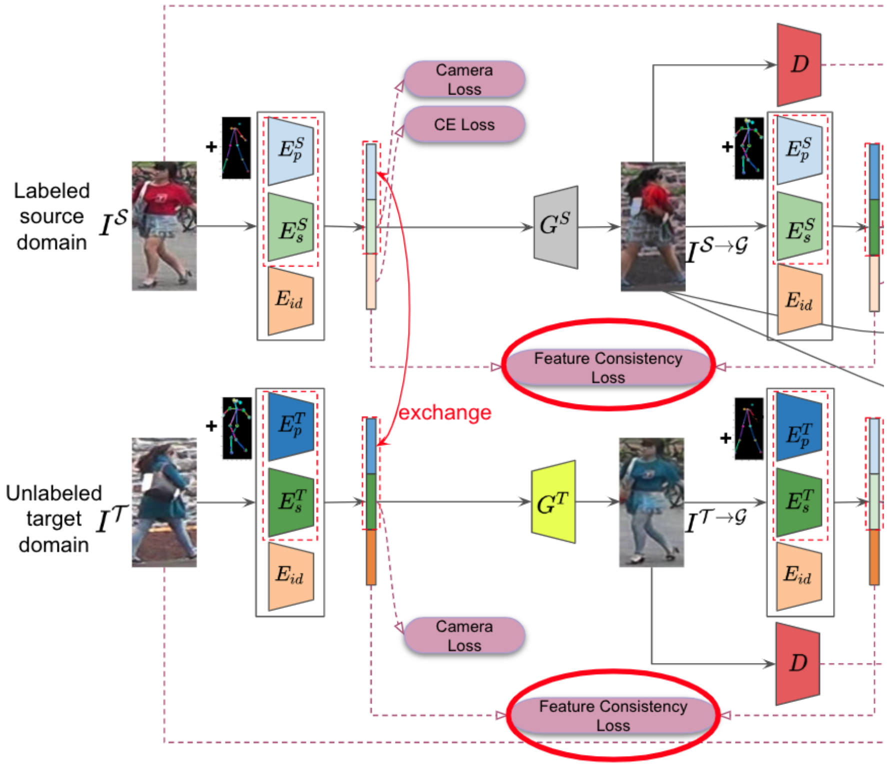


Figure 3.5: We calculate the **feature consistency loss** between the features of real images and generated images for feature reconstruction with minimum noise. (**Best viewed in colors**).

$$\mathcal{L}^{\mathcal{T}}_{Con\_Env} = \mathbb{E}[||f^{\mathcal{T}}_{env} - E^{\mathcal{T}}_{env}(I^{\mathcal{S}\to\mathcal{G}})||_1], \tag{3.2}$$

$$\mathcal{L}^{\mathcal{S}}_{Con\_Ind} = \mathbb{E}[||f_{ind} - E_{id}(I^{\mathcal{S}\to\mathcal{G}})||_1], \tag{3.3}$$

$$\mathcal{L}^{\mathcal{T}}_{Con\_Ind} = \mathbb{E}[||f_{ind} - E_{id}(I^{\mathcal{T}\to\mathcal{G}})||_1], \tag{3.4}$$

where $E^{\mathcal{S}}_{env}(I^{\mathcal{T}\to\mathcal{G}})$ $(E^{\mathcal{T}}_{env}(I^{\mathcal{S}\to\mathcal{G}}))$ gives the reconstructed environment features of the source (target) domain, and $E_{id}(I^{\mathcal{S}\to\mathcal{G}})$ and $E_{id}(I^{\mathcal{T}\to\mathcal{G}})$ give the shared reconstructed appearance features of the source and target domain. $\mathcal{L}^{\mathcal{S}}_{Con\_Env}$ $(\mathcal{L}^{\mathcal{T}}_{Con\_Env})$ is the environment feature consistency loss of source (target) domain, and $\mathcal{L}^{\mathcal{S}}_{Con\_Ind}$ $(\mathcal{L}^{\mathcal{T}}_{Con\_Ind})$ is the appearance feature consistency loss of source (target) domain.

The total feature consistency loss $(\mathcal{L}_{Con})$ can be formulated as:

$$\mathcal{L}_{Con} = \mathcal{L}^{\mathcal{S}}_{Con\_Ind} + \mathcal{L}^{\mathcal{T}}_{Con\_Ind} + \mathcal{L}^{\mathcal{S}}_{Con\_Env} + \mathcal{L}^{\mathcal{T}}_{Con\_Env}. \tag{3.5}$$

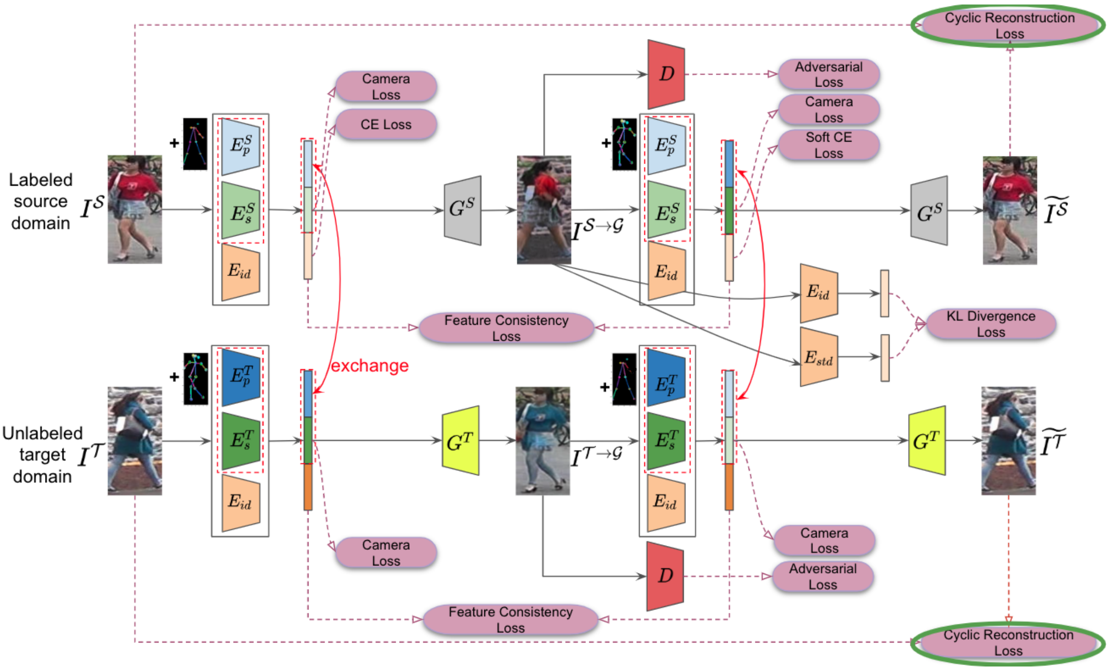


Figure 3.6: We calculate **cyclic reconstruction loss** between original images and their cyclically generated counterparts to provide a supervision for unpaired source and target domain images. (**Best viewed in colors**).

Cyclic Reconstruction. In addition to feature consistency loss, we provide supervision by performing cyclic reconstruction of image as shown in Figure 3.6. We feed our generated images $I^{\mathcal{S}\to\mathcal{G}}$ $(I^{\mathcal{T}\to\mathcal{G}})$ again in the environment encoders $E^{\mathcal{S}}_{env}(E^{\mathcal{T}}_{env})$ and shared identity encoder E_{id} to get the environment and identity-related features respectively. The decoders $G^{\mathcal{S}}$ and $G^{\mathcal{T}}$, take these features to reconstruct our original images in a cyclic manner. We use pixel-wise l_1 loss between the reconstructed $\widetilde{I^{\mathcal{S}}}$ $(\widetilde{I^{\mathcal{T}}})$ and original images $I^{\mathcal{S}}$ $(I^{\mathcal{T}})$ to achieve regularization. This cyclic reconstruction in the source and target domain can be represented as $\widetilde{I^{\mathcal{S}}} = G^{\mathcal{S}}(E^{\mathcal{S}}_{env}(I^{\mathcal{T}\to\mathcal{G}}), E_{id}(I^{\mathcal{S}\to\mathcal{G}}))$ and $\widetilde{I^{\mathcal{T}}} = G^{\mathcal{T}}(E^{\mathcal{T}}_{env}(I^{\mathcal{S}\to\mathcal{G}}), E_{id}(I^{\mathcal{T}\to\mathcal{G}}))$.

The total cyclic reconstruction loss $\mathcal{L}_{Cyc}$ can be represented in terms of the cyclic reconstruction loss $\mathcal{L}^{\mathcal{S}}_{Cyc}$ $(\mathcal{L}^{\mathcal{T}}_{Cyc})$ in the source (target) domain is given as follows:

$$\mathcal{L}_{Cyc} = \mathcal{L}^{\mathcal{S}}_{Cyc} + \mathcal{L}^{\mathcal{T}}_{Cyc} = \mathbb{E}[||I^{\mathcal{S}} - \widetilde{I^{\mathcal{S}}}||_1] + \mathbb{E}[||I^{\mathcal{T}} - \widetilde{I^{\mathcal{T}}}||_1]. \tag{3.6}$$

Camera-style Loss. To capture cross camera style variation in our generated images, we define camera loss (Figure 3.7) to constrain the camera style of generated image $I^{\mathcal{S}\to\mathcal{G}}$ to target camera style and $I^{\mathcal{T}\to\mathcal{G}}$ to source camera style. Style encoder gives the probability distribution of a translated image belonging to a camera. The camera loss

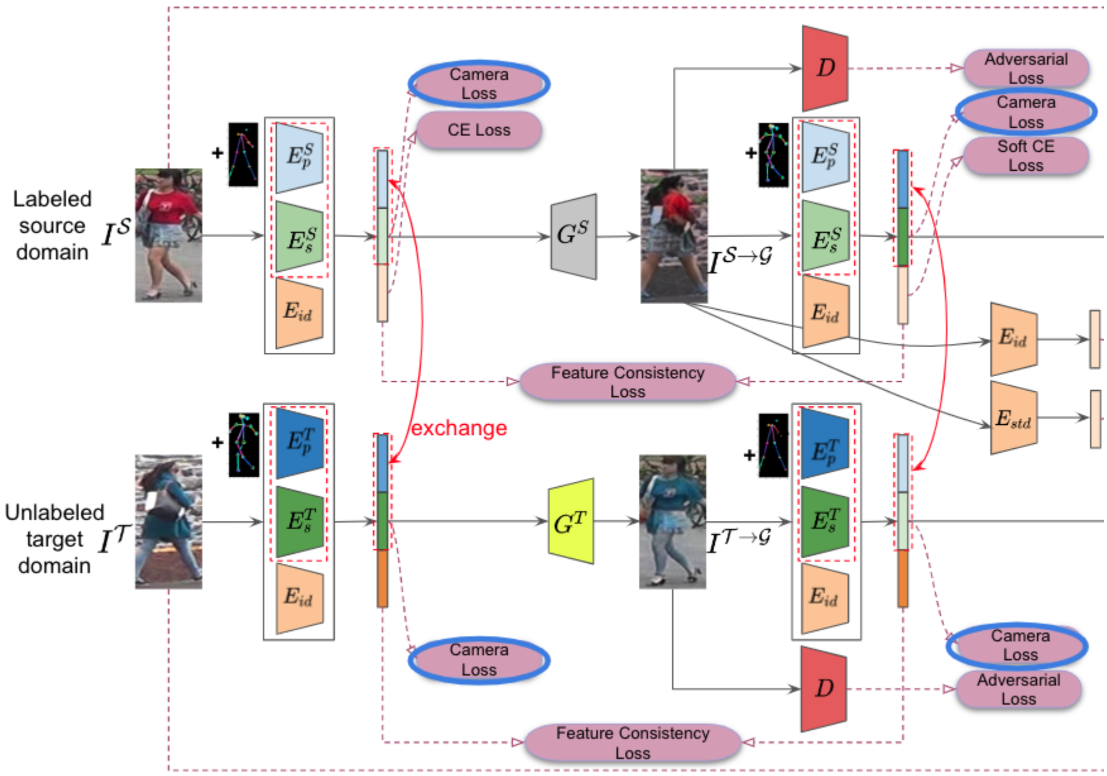


Figure 3.7: Our approach to incorporating cross-camera style variation into the generated images involves the formulation of a **camera loss** function. (**Best viewed in colors**).

for generated images in the source and target domain is given as follows:

$$\mathcal{L}_{Cam}^{\mathcal{S}\to\mathcal{G}} = \mathbb{E}[-\log(p(y_{cam}^{\mathcal{T}}/I^{\mathcal{S}\to\mathcal{G}}))] \tag{3.7}$$

$$\mathcal{L}_{Cam}^{\mathcal{T}\to\mathcal{G}} = \mathbb{E}[-\log(p(y_{cam}^{\mathcal{S}}/I^{\mathcal{T}\to\mathcal{G}}))], \tag{3.8}$$

where $p(y_{cam}^{\mathcal{T}}/I^{\mathcal{S}\to\mathcal{G}})$ and $p(y_{cam}^{\mathcal{S}}/I^{\mathcal{T}\to\mathcal{G}})$ are the predicted probabilities that $I^{\mathcal{S}\to\mathcal{G}}$ $(I^{\mathcal{T}\to\mathcal{G}})$ belong to its camera label $y_{cam}^{\mathcal{T}}$ $(y_{cam}^{\mathcal{S}})$.

We also classify the real images with corresponding camera labels in source and target domain. The camera loss for real source and target data is formulated as:

$$\mathcal{L}_{Cam}^{\mathcal{S}} = \mathbb{E}[-\log(p(y_{cam}^{\mathcal{S}}/I^{\mathcal{S}}))], \tag{3.9}$$

$$\mathcal{L}_{Cam}^{\mathcal{T}} = \mathbb{E}[-\log(p(y_{cam}^{\mathcal{T}}/I^{\mathcal{T}}))], \tag{3.10}$$

where $p(y_{cam}^{\mathcal{S}}/I^{\mathcal{S}})$ $(p(y_{cam}^{\mathcal{T}}/I^{\mathcal{T}}))$ is the predicted probability that $I^{\mathcal{S}}$ $(I^{\mathcal{T}})$ belongs to camera label $y_{cam}^{\mathcal{S}}$ $(y_{cam}^{\mathcal{T}})$.

The total camera-style loss can be formulated as:

$$\mathcal{L}_{Cam} = \mathcal{L}^{\mathcal{S}}_{Cam} + \mathcal{L}^{\mathcal{T}}_{Cam} + \mathcal{L}^{\mathcal{S}\to\mathcal{G}}_{Cam} + \mathcal{L}^{\mathcal{T}\to\mathcal{G}}_{Cam} \tag{3.11}$$

Adversarial Loss Function. The adversarial loss function is defined to decrease the gap between the real images and fake translated images (Goodfellow *et al.*, 2014*a*). It is formulated as:

$$\begin{aligned}\mathcal{L}_{Adv} = \mathbb{E}[\log D(I^{\mathcal{S}}) + \log(1 - D(I^{\mathcal{S}\to\mathcal{G}})]+ \\ \mathbb{E}[\log D(I^{\mathcal{T}}) + \log(1 - D(I^{\mathcal{T}\to\mathcal{G}})]\end{aligned} \tag{3.12}$$

3.2.2 Discriminative Learning

In our proposed approach IPES-GAN, we train both the generative and discriminative modules simultaneously. We train our network iteratively to generate high-quality synthetic data based on our generation module (which contains the shared identity encoder) and fine-tune the identity encoder with the high-quality synthetic data. After training, the learned identity encoder is used to perform discriminative (Re-ID) learning. It is used to match the query images to gallery images to evaluate the performance of our proposed approach. Thus, our identity encoder, which performs discriminative learning (Re-ID task) after training, is trained with the generated images along with generation (generative task). In contrast, the generative task in existing methods (Wei *et al.*, 2018), (Deng *et al.*, 2018), remains independent of the Re-ID task. They are often trained conventionally to generate data. Thus, we summarize our network as: (1) the generator synthesizes images which are used to fine-tune the identity encoder; (2) the learned identity encoder, in turn, influences the generator to generate better quality images; and (3) generator and identity encoder are jointly optimized.

We show the training pipeline for a single domain (source domain since it is labeled) in Figure 3.8. We generate the images and use these images to fine-tune the identity encoder. The learned identity encoder generates high-quality images, thus helping the generation process. We feed the generated images to the identity encoder during

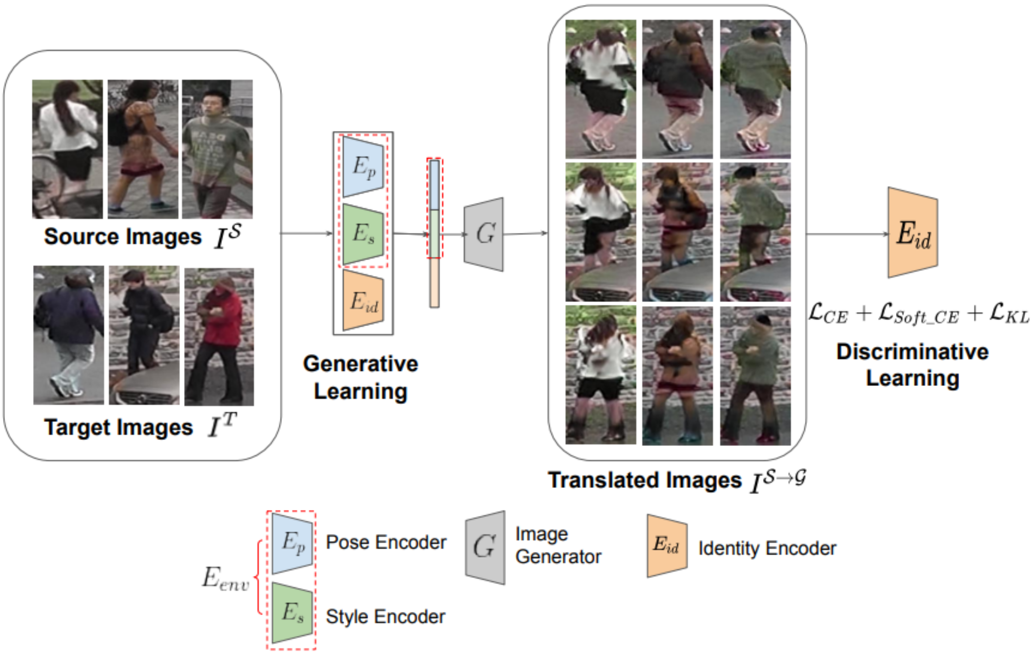


Figure 3.8: Pipeline of our approach. Our IPES-GAN network is used to generate images (Translated Images) with environment style (pose and background) of target image while preserving the identity-related features of the source images. The cross-entropy loss $\mathcal{L}_{CE}$ performs classification on labeled source images. The Soft cross-entropy loss $\mathcal{L}_{Soft\_CE}$ and divergence loss $\mathcal{L}_{KL}$ classify generated images that are pseudo-labeled with source labels. We employ these losses on generated images to fine-tune the identity encoder, which in turn improves generation. (**Best viewed in colors**).

the generation process, which reduces the operational complexity and saves the space needed to store the extensive synthetic data. Once identity encoder E_{id} is fine-tuned with the generated images, it is used to extract features for Re-ID matching in the target domain.

Soft Cross-Entropy Loss (Soft_CE). The vanilla cross-entropy loss is formulated as (Zhong *et al.*, 2018*b*):

$$\mathcal{L}_{CE} = \mathbb{E}[-\log(p(c))k(c)] \tag{3.13}$$

where the number of classes in the source domain is C_S and $c = \{1, ..., C_S\}$. An image belonging to label c is given the probability $p(c)$. The ground-truth distribution is $k(c)$. Since, a specific label y_{id}^S is given to each person in the training set, $k(c)$ can be given as,

$$k(c) = \begin{cases} 1 & c = y_{id}^{\mathcal{S}} \\ 0 & c \neq y_{id}^{\mathcal{S}} \end{cases} \tag{3.14}$$

Cross-entropy loss for a single person can be rewritten as $-\log p(y_{id}^{\mathcal{S}}/I^{\mathcal{S}})$. For the real images in the source domain we use the vanilla cross-entropy loss:

$$\mathcal{L}_{CE} = \mathbb{E}[-\log p(y_{id}^{\mathcal{S}}/I^{\mathcal{S}})] \tag{3.15}$$

We treat our style transferred images as the regular training samples. However, due to the environment-switching feature in our generated images, we define a smooth label regularization on the style transferred images to reduce noise and distribute the labels softly. In this loss function, we give non-zero weights to the ground truth label as well as other classes. The ground truth label is given less confidence, and other classes are given small weights. This weight assignment is given as:

$$k_{Soft\_CE}(c) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{C_{\mathcal{S}}} & c = y_{id}^{\mathcal{S}} \\ \frac{\epsilon}{C_{\mathcal{S}}} & c \neq y_{id}^{\mathcal{S}} \end{cases} \tag{3.16}$$

where $k_{Soft\_CE}(c)$ is the ground-truth distribution and $\epsilon \in (0, 1)$. With this new weight configuration, the cross-entropy loss is re-defined as:

$$\mathcal{L}_{Soft\_CE} = -(1 - \epsilon) \log p(y_{id}^{\mathcal{S}}/I^{\mathcal{S} \to \mathcal{G}}) - \frac{\epsilon}{C_{\mathcal{S}}} \sum_{c=1}^{C_{\mathcal{S}}} \log p(c) \tag{3.17}$$

KL Divergence. Apart from classifying the images using the soft cross-entropy loss, we also use KL divergence to supervise classification on generated images. In order to achieve this, we minimize the KL divergence between $p(I^{\mathcal{S} \to \mathcal{G}})$ and $q(I^{\mathcal{S} \to \mathcal{G}})$. $p(I^{\mathcal{S} \to \mathcal{G}})$ is the predicted probability of the generated images belonging to a particular class c in the source domain. Similarly, $q(I^{\mathcal{S} \to \mathcal{G}})$ is the probability predicted by the CNN model trained with identity loss on the original source training dataset as shown in Figure 3.9. KL divergence is given as:

$$\mathcal{L}_{KL} = \mathbb{E}[-\sum_{c=1}^{C_{\mathcal{S}}} q(c|I^{\mathcal{S} \to \mathcal{G}}) log(\frac{p(c|I^{\mathcal{S} \to \mathcal{G}})}{q(c|I^{\mathcal{S} \to \mathcal{G}})})] \tag{3.18}$$

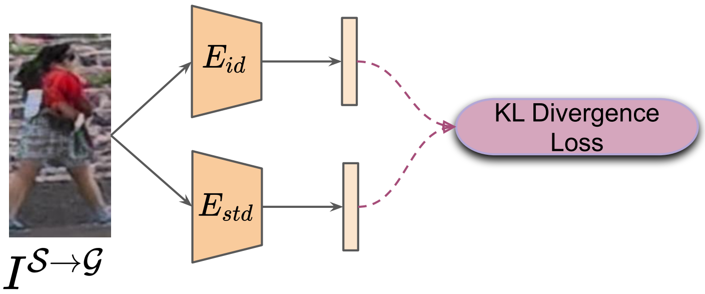


Figure 3.9: KL Divergence between the probability predicted by our trained identity encoder E_{id} and standard CNN model E_{std}.

Overall Objective Function. Our IPES-GAN is jointly optimized with (Eq. 3.6, 3.5, 3.11, 3.12, 3.15, 3.17, 3.18) losses,

$$\begin{aligned} \mathcal{L}_{IPES-GAN} = w_1\mathcal{L}_{Cyc} + w_2\mathcal{L}_{Con} + \mathcal{L}_{Cam} + \mathcal{L}_{Adv} \\ +\mathcal{L}_{CE} + w_3\mathcal{L}_{Soft\_CE} + w_4\mathcal{L}_{KL} \end{aligned} \tag{3.19}$$

where w_1, w_2, w_3 and w_4 are the weights assigned to the loss functions. $w_1 = 5$, $w_3 = 0.5$. By utilizing random search, we can acquire weights that promote stability throughout the entirety of the training process. The weight corresponding to cyclic reconstruction loss controls the cyclic generation of cross-domain data in self-supervised manner. We give a large weight of $w_1 = 5$ to cyclic-reconstruction loss L_{Cyc} which is in accordance to previous GAN-based image generation methods (Zhu *et al.*, 2017; Lee *et al.*, 2018; Huang *et al.*, 2018). The high value of w_1 helps us to generate better quality images. We add w_2 and w_4 after 20K iterations and linearly increase them from 0 to 3 in next 6K iterations and these are constant at 3 thereafter. w_6 corresponding to the soft CE loss is assigned a small value as the initial generated images contain a lot of noise and less appearance information. We did not initially add feature consistency and KL divergence losses to avoid training instability and introduced them after 20K iterations.

3.3 Experiments

We discuss the datasets used, implementation details, and experiments performed in this section.

3.3.1 Datasets and Metrics

We use three large-scale Re-ID datasets (Market-1501, DukeMTMC-ReID and MSMT17) to evaluate our proposed method.

Market-1501 (Zheng *et al.*, 2015). It is collected in front of a supermarket in Tsinghua University. It consists of images taken from 6 different cameras. It has 12,936 training images with 751 identities (IDs). The testing set consists of 19,732 gallery images with 750 IDs and 3,366 query images.

DukeMTMC-ReID (Zheng *et al.*, 2017). It is a dataset of surveillance video footage taken on Duke University's campus in 2014. This dataset has 16,522 training images with 1,404 IDs taken from 8 different cameras. The testing set comprises 17,661 gallery images with 702 IDs and 2,288 query images.

MSMT17 (Wei *et al.*, 2018). It utilizes a 15-camera network deployed on campus. It has 12 outdoor and three indoor cameras. This dataset has a total of 126,441 bounding boxes with 4,101 IDs. The training and testing ratio is 1:3. The training images are 32,621 bounding boxes of 1,041 IDs taken from 15 different cameras. The testing set comprises 82,161 gallery images, and query images are randomly selected to be 11,659.

Evaluation Metrics. We use several metrics to show our performance. (1) To evaluate the Re-ID performance, the standard Cumulative Matching Characteristics (CMC) values and mean Average Precision (mAP) (Zheng *et al.*, 2015) are adopted since one person has multiple ground truths in the gallery set. (2) We check the visual quality of generated images with two evaluation metrics: LPIPS (Zhang *et al.*, 2018*c*) uses deep features as perceptual similarity metric. We use AlexNet pre-trained on ImageNet for evaluation. Fréchet Inception Distance (FID) (Heusel *et al.*, 2017) measures the reality factor by checking the anomalies present in the generated images. The Mean Opinion Score (MOS) test is exploited to evaluate the quality and diversity of the generated images objectively. We use Maximum Mean Discrepancy (MMD) to quantify the domain gap between $\mathcal{T}$ (target) and $\mathcal{G}$ (generated) images, and $\mathcal{T}$ (target) and $\mathcal{S}$ (source) images.

3.3.2 Implementation Details

We use Adam optimizer for generator and discriminator with a learning rate set as 0.0002 and momentum 0.999. SGD optimizer is employed for the identity encoder with a learning rate of 0.001 and momentum of 0.9. We load source and target datasets simultaneously. We set a mini-batch size of 16. We randomly sample an equal number of images from both source and target datasets. Labels of source data are also loaded. While loading images from the source dataset, we chose a random ID and loaded these images from different cameras.

On the other hand, the labels of the target dataset are unknown. Thus, we randomly pick images from different cameras. We concatenate the image and the 18 channel pose data, which makes a 21 channel input to the pose encoder. The Encoder comprises the pose encoders, style encoders, and identity encoder. The pose encoders consist of two downsampling layers, four residual layers, and atrous spatial pyramid pooling (ASPP) (Chen *et al.*, 2018). The style encoders are a 6-layer convolution network. The identity encoder is based on a standard IBN-ResNet-50 pre-trained on ImageNet. We design decoders with similar architecture as the pose encoders. For style transfer, we use adaptive instance normalization at the time of decoding. We employ least-squares generative adversarial network (LSGAN) (Mao *et al.*, 2017). At the time of testing on the target dataset, we use our identity encoder to extract a 1024-dim feature vector after the global average pooling layer. We use 2 NVIDIA Tesla P100 PCIe 16 GB GPUs for performing all the experiments.

3.3.3 Comparison with State-of-the-arts

In this subsection, we compare the state-of-the-art methods to show the effectiveness of our generation network on the unsupervised Re-ID. We exploit the Market-1501 (DukeMTMC-ReID) as the source dataset and evaluate the methods on the DukeMTMC-ReID (Market-1501). We show the results in Table 3.1. We primarily included previous works that perform unsupervised Re-ID with person image generation for a fair comparison. PTGAN (Wei *et al.*, 2018), SPGAN+LMP (Deng *et al.*, 2018), CamStyle (Zhong *et al.*, 2018*b*), HHL (Zhong *et al.*, 2018*a*), SBSGAN (Huang *et al.*, 2019*a*), CR-GAN (Chen *et al.*, 2019), CSGLP (Ren *et al.*, 2019), CGAN-TM (Tang *et al.*, 2020) are recent methods based on image generation. Most of these methods adopt one backbone; for example, SPGAN+LMP, CamStyle, HHL, and CR-GAN use ResNet-50, and CGAN-TM uses Densenet-121. The methods using a generation technique are marked √.

We show results by using two different backbones for the identity encoder. IPES-GAN (ResNet-50) achieves 64.1% for Market-1501 and 53.5% for DukeMTMC-ReID in CMC-1 accuracy, which is better than previous approaches. Then, after using IBN-ResNet-50 we achieve **66.8% for Market-1501 and 55.7% for DukeMTMC-ReID in CMC-1 accuracy**. Our method outperforms all the GAN-based approaches with a margin of 5.4% in CMC-1 accuracy for Market-1501. Our proposed IPES-GAN also outperforms methods that do not perform generation by a large margin.

Table 3.1: The comparison with other state-of-the-art methods. The mAP, CMC-1, CMC-5, and CMC-10 (%) results are reported. `source` → `target` represents the setting of training on the source dataset and testing on the target dataset. **Red** and blue fonts indicate the best and the second best results. √ indicates that the corresponding method uses the generation technique to augment training samples.

Methods	Reference	Generation	DukeMTMC-ReID → Market-1501				Market-1501 → DukeMTMC-ReID			
			CMC-1	CMC-5	CMC-10	mAP	CMC-1	CMC-5	CMC-10	mAP
LOMO (Liao *et al.*, 2015)	CVPR'15		12.3	21.3	26.6	4.8	12.3	21.3	26.6	4.8
BoW (Zheng *et al.*, 2015)	ICCV'15		17.1	28.8	34.9	8.3	17.1	28.8	34.9	8.3
UMDL (Peng *et al.*, 2016)	CVPR'16		34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
CAMEL (Yu *et al.*, 2017)	ICCV'17		54.5	-	-	-	-	-	-	-
PUL (Fan *et al.*, 2018)	TOMM'18		45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
PTGAN (Wei *et al.*, 2018)	CVPR'18	√	38.6	-	66.1	-	27.4	-	50.7	-
SPGAN+LMP (Deng *et al.*, 2018)	CVPR'18	√	57.7	75.8	82.4	26.7	46.4	62.3	68.0	26.2
TJ-AIDL (Wang *et al.*, 2018*a*)	CVPR'18		58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
CamStyle (Zhong *et al.*, 2018*b*)	CVPR'18	√	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
HHL (Zhong *et al.*, 2018*a*)	ECCV'18	√	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
ATNet (Liu *et al.*, 2019*b*)	CVPR'19		55.7	73.2	79.4	25.6	45.1	59.5	64.2	24.9
SBSGAN (Huang *et al.*, 2019*a*)	ICCV'19	√	58.5	-	-	27.3	53.5	-	-	30.8
CR-GAN (Chen *et al.*, 2019)	CVPR'19	√	59.6	-	-	29.6	52.2	-	-	30.0
CASCL (Wu *et al.*, 2019)	ICCV'19		64.7	80.2	**85.6**	**35.6**	51.5	66.7	71.7	30.5
CSGLP (Ren *et al.*, 2019)	TIFS'19	√	59.2	76.2	83.2	31.1	47.8	62.3	68.3	27.1
CGAN-TM (IDE) (Tang *et al.*, 2020)	TIP'20	√	61.4	-	-	31.3	54.6	-	-	32.6
PPAN (Yang *et al.*, 2020)	TMM'20		62.7	77.2	82.5	30.2	55.6	68.1	73.2	**34.0**
IPES-GAN (ResNet-50)	Ours	√	64.1	79.3	83.1	33.6	53.5	69.1	73.1	32.9
IPES-GAN (IBN-ResNet-50)	Ours	√	**66.8**	**81.2**	85.4	34.4	**55.7**	**71.4**	**75.2**	33.3
Methods	Reference	Generation	DukeMTMC-ReID → MSMT17				Market-1501 → MSMT17			
			CMC-1	CMC-5	CMC-10	mAP	CMC-1	CMC-5	CMC-10	mAP
PTGAN (Wei *et al.*, 2018)	CVPR'18	√	11.8	-	27.4	3.3	10.2	-	24.4	2.9
SPGAN + LMP (Deng *et al.*, 2018)	CVPR'18	√	16.8	27.1	32.6	5.8	15.3	23.7	28.6	3.8
CamStyle (Zhong *et al.*, 2018*b*)	CVPR'18	√	19.8	29.2	34.1	6.3	17.6	26.2	31.4	4.9
HHL (Zhong *et al.*, 2018*a*)	ECCV'18	√	20.4	31.1	35.8	6.9	18.9	27.1	32.9	5.2
IPES-GAN (ResNet-50)	Ours	√	20.6	31.0	36.4	6.5	18.4	28.9	34.4	5.9
IPES-GAN (IBN-ResNet-50)	Ours	√	**23.1**	**32.6**	**39.1**	**7.4**	**20.2**	**31.2**	**37.2**	**6.9**

Performance on MSMT17 (Wei *et al.*, 2018). We evaluate the results of our proposed approach on another large-scale dataset, MSMT17, and compare it with GAN-based cross-dataset person re-identification methods in Table 3.1. We use the Market-1501 (DukeMTMC-ReID) as the source dataset and test the target MSMT17 dataset. Previous SOTA methods used backbone as ResNet-50. On the other hand, we report results using ResNet-50 and IBN-ResNet-50. We outperform the SOTA methods by a margin of 1.3% for Market-1501 → MSMT17 and 2.7% for DukeMTMC-ReID → MSMT17.

Table 3.2: Comparison with SOTA hybrid methods. `source` → `target` represents the setting of training on the source dataset and testing on the target dataset. 'Duke' refers to DukeMTMC-ReID and 'Market' refers to Market-1501 datasets.

Types	Methods	Reference	Generation	Duke → Market		Market→ Duke	
				CMC-1 (%)	mAP (%)	CMC-1 (%)	mAP (%)
Image	SPGAN (Deng *et al.*, 2018)	CVPR'18	√	57.7	26.7	46.4	26.2
	CR-GAN (Chen *et al.*, 2019)	CVPR'19	√	59.6	29.6	52.2	30.0
	IPES-GAN	Ours	√	**66.8**	**34.4**	**55.7**	**33.3**
Feature	TAUDL (Li *et al.*, 2018*b*)	CVPR'19		63.7	41.2	61.7	43.2
Hybrid	SPGAN + TAUDL	CVPR'18	√	66.5	38.5	66.1	47.2
	CR-GAN + TAUDL	CVPR'19	√	77.7	54.0	68.9	48.6
	IPES-GAN + TAUDL	Ours	√	**78.9**	**54.3**	**73.6**	**53.6**

Comparison with SOTA hybrid methods. We combine our IPES-GAN with TAUDL [6] and compare it with existing hybrid (image and feature level learning) methods in Table 3.2. We apply TAUDL with our trained identity encoder for unsupervised Re-ID in the target domain. Our IPES-GAN outperforms the SOTA methods in hybrid formulation by a margin of 1.2% for DukeMTMC-ReID → Market-1501 and by 4.7% for Market-1501 → DukeMTMC-ReID.

We should indicate that our method has a slightly lower mAP of 1.2% on Market-1501 than CASCL (Wu *et al.*, 2019). CASCL has better performance for CMC-10 and mAP on DukeMTMC-ReID → Market-1501. However, our method outperforms CASCL in CMC-1 and CMC-5. Our method also outperforms CASCL on Market-1501 → DukeMTMC-ReID. Furthermore, CASCL's performance heavily depends on the number of cameras in the source domain. Another method (Song *et al.*, 2020), has a mAP of 53% and CMC-1 of 75.8% for DukeMTMC-ReID → Market-1501 and a mAP of 49% and CMC-1 of 68.4% for Market-1501 → DukeMTMC-ReID. It belongs to a pseudo-label-based method, which makes guesses for unlabeled target data based on an encoder. The encoder is trained using the guessed labels, which are selected iteratively using clustering. However, it uses DBSCAN clustering method (Coifman, 1998) to generate data clusters. DBSCAN has parameters that need to be tuned for different datasets.

On the other hand, our method has a mAP of 34.4% and CMC-1 of 66.8% for DukeMTMC-ReID → Market-1501 and a mAP of 33.3% and CMC-1 of 55.7% for Market-1501 → DukeMTMC-ReID, it belongs to GAN based approach. The representation of the feature space by decoupling the identity-related and identity non-related features allows our network to easily generalize to unseen scenarios, which we also show by providing results on three large-scale Re-ID datasets- Market-1501 (Zheng *et al.*, 2015), DukeMTMC-ReID (Zheng *et al.*, 2017) and MSMT17 (Zheng *et al.*, 2017). Since our method falls under GAN based approach, a direct comparison may not give a complete idea of the performance. Further, our method also has the advantage that the images generated can be used by other Re-ID models to boost the performance.

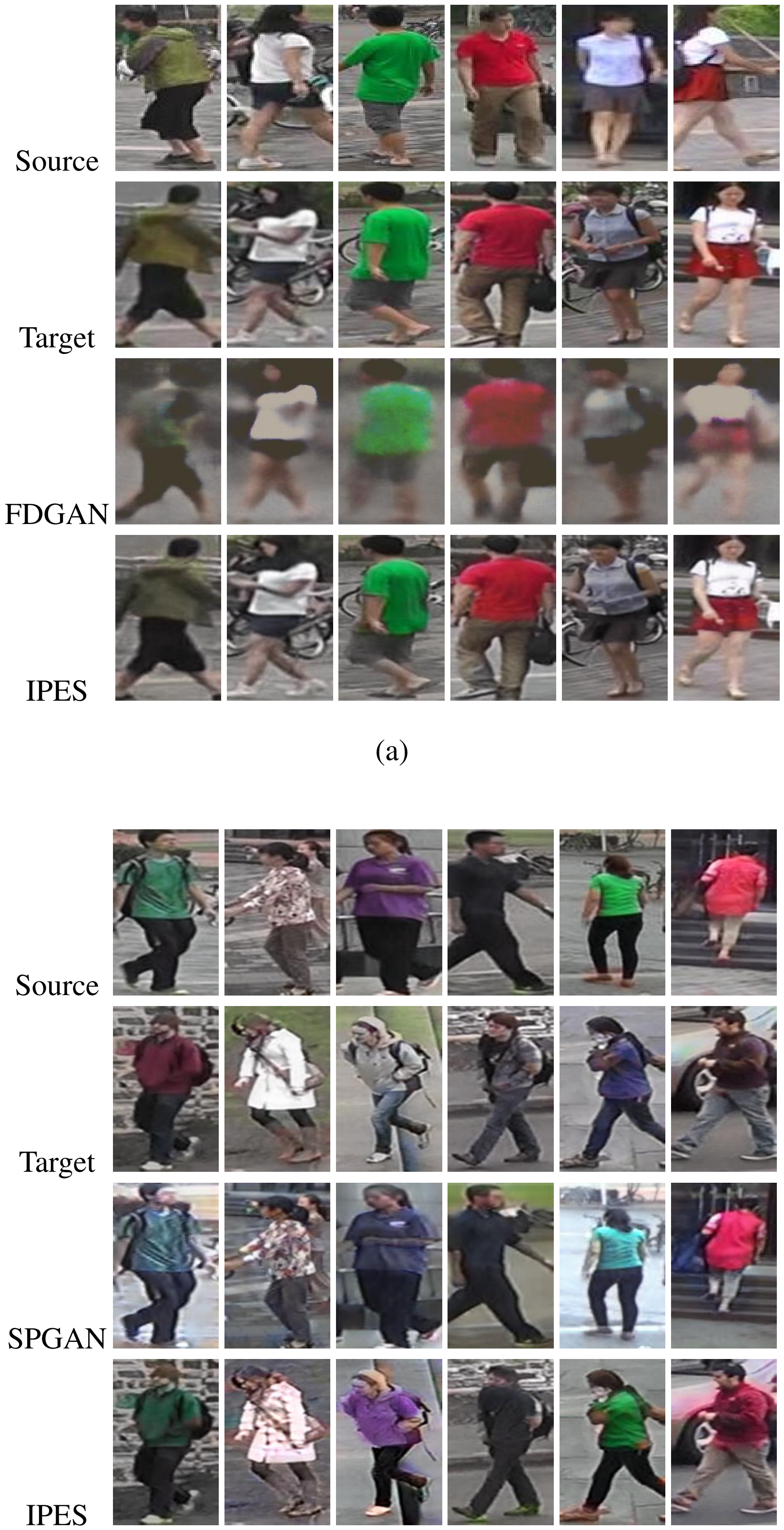
Qualitative Comparison. We show qualitative comparison with two previous generation-based methods as shown in Figure 3.10. The first and second rows are the source and target images, respectively. The third and fourth row show the translated images obtained by FDGAN (Ge *et al.*, 2018) and our method (IPES) respectively in Figure 3.10 (a), and SPGAN (Deng *et al.*, 2018) and our method in Figure 3.10 (b).

FD-GAN disentangles the pose information and generates more images with target poses. However, it needs paired images for training and cannot be extended for unsupervised domain adaptation. On the other hand, our method does not need any paired data and is trained in an unsupervised manner. Without paired information, we generate better quality images than FDGAN, which we show in Figure 3.10 (a). SPGAN attempts to translate source images to a target style, where person and background are treated as a whole. The translated image may obtain a similar target style, but the redundant source background still exists, as we show in Figure 3.10 (b). However, as shown in the figure, the images generated by IPES adapt to the target pose and background while preserving the identity-related features like the color and style of a person's clothing and shoes.

3.3.4 Ablation Study

We conduct an ablation study to show the influence of pose change, cyclic encoding, and loss functions in IPES-GAN.

Effectiveness of Pose Change. Re-ID aims to match persons across non-overlapping camera views. Among other factors like a person's appearance, illumination, occlusion, etc., large pose variation makes Re-ID a difficult task. Re-ID becomes more challenging in an unsupervised setting where labels of target data are unknown. Pose-variation



(a)

(b)

Figure 3.10: (a) Qualitative comparison of our method with FD-GAN (Ge *et al.*, 2018). (b) our method with SPGAN (Deng *et al.*, 2018). (**Best viewed in colors**).

between source and target domains is an important factor that precludes from learning a robust unsupervised domain adaptation method. We summarize the experimental results to show the effectiveness of pose change in Table 3.3. Our model performs much better unsupervised domain adaptation with pose change.

Table 3.3: Quantitative evaluation of effectiveness of pose change in unsupervised domain. 'M', 'D' and 'MSMT' denote the Market-1501 (Zheng *et al.*, 2015), DukeMTMC-ReID (Zheng *et al.*, 2017) and MSMT17 (Wei *et al.*, 2018) datasets respectively.

Methods	M → D		M → MSMT		D → M		D → MSMT	
	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP
Without Pose Change	63.9	30.6	14.5	4.3	53.7	31.2	18.0	5.6
With Pose Change	**66.8**	**34.4**	**20.2**	**6.9**	**55.7**	**33.3**	**23.1**	**7.4**

Cyclic Encoding. Cyclic encoding helps us control our unsupervised image translation in different data domains without identifying aligned image pairs. In order to prove this fact, in the first experiment, we generate images by single encoding and swapping of the environment features and, after that, training our identity encoder E_{id} with these images. It shows a significant boost of 8% in CMC-1 on DukeMTMC-ReID and 4.2% on Market-1501 compared to direct testing as we show in Table 3.4. In the next experiment, we encode our intermediate generated images to reconstruct the input images in a cyclic manner while training the identity encoder. With our cyclic encoding in the generative module we get a significant boost of $13.1\%/10.2\%$ in CMC-1 on DukeMTMC-ReID/Market-1501 over direct testing. From these results, we can conclude that our proposed model is successful in reducing the domain gap.

Loss Functions. Our model is trained with adversarial loss along with four other losses.

Table 3.4 shows the effect of various loss functions in optimizing the model. From Tables 3.4 and 3.5, it is clear that when we perform direct testing on target dataset with a model trained in a supervised setting with all the loss functions, we get better performance than in the setting where we remove $\mathcal{L}_{Cyc}$ or $\mathcal{L}_{KL}$ from our proposed network. Thus, all the loss functions help each other to attain an optimized performance.

To prove the significance of these loss functions, we conduct an ablation study by removing one loss at a time. First, as discussed earlier, cyclic encoding improves Re-ID accuracy by a large margin, and since it is achieved by cyclic reconstruction loss, there is a considerable decrease in performance of about 18.8% in DukeMTMC-ReID and 14.5% in Market-1501 in the absence of this loss. This decrease in per-

Table 3.4: Ablation study on different types of training samples. "Supervised" stands for the labeled source domain. "Direct Testing" denotes the unlabeled target domain by the model trained in supervised setting. "Single Encoding" refers to generation of intermediate image. "Cyclic Encoding" indicates our proposed IPES-GAN network. `source` → `target` represents the setting of training on the source dataset and testing on the target dataset. 'M' and 'D' denote the Market-1501 and DukeMTMC-ReID datasets, respectively. The mAP and CMC-1 (%) results are reported.

Methods	M → D		D → M	
	CMC-1	mAP	CMC-1	mAP
Supervised	84.4	70.1	93.4	80.6
Direct Testing	42.6	24.3	56.1	26.8
Single Encoding	50.6	30.9	60.3	29.5
Cyclic Encoding	55.7	34.1	66.8	34.4

formance is due to the absence of image-level information, due to which our model cannot generate human-perceptible images. Second, KL divergence combined with our identity losses on real and generated images helps our IPES-GAN network generate good quality images by preserving their identity-related features. Without these losses, there is around 10% decrease in accuracy for both Market-1501 and DukeMTMC-ReID datasets. Third, feature consistency compliments our cyclic reconstruction and helps generation without any image-level information. Finally, camera loss is introduced to reduce the effect of change in camera styles across different datasets. Conventional CE loss is adopted for real source images only, for which true labels are known. Using soft CE loss for real images would decrease the performance. We can also adopt CE loss for generated images. However, due to the environmental-switching feature in our generated images, we employ soft CE loss to reduce the noise and distribute the labels softly. If we use only one kind of loss (CE loss instead of soft CE loss as discussed

Table 3.5: Ablation study on loss functions. `source` → `target` represents the setting of training on the source dataset and testing on the target dataset. 'M' and 'D' denote the Market-1501 and DukeMTMC-ReID datasets respectively. The mAP and CMC-1 (%) results are reported.

Methods	M → D		D → M	
	CMC-1	mAP	CMC-1	mAP
w/o $\mathcal{L}_{Cyc}$	36.9	19.7	52.3	23.8
w/o $\mathcal{L}_{KL}$	40.2	23.4	54.0	24.6
w/o $(\mathcal{L}_{CE} + \mathcal{L}_{Soft\_CE})$	45.2	25.0	57.2	26.8
w/o $\mathcal{L}_{Con}$	51.0	30.2	60.3	29.5
w/o $\mathcal{L}_{Cam}$	51.5	29.8	62.9	30.0
Our Method	**55.7**	**34.1**	**66.8**	**34.4**

above) for both real and generated images, we will get a drop of CMC-1 accuracy by 3.9% in Market-1501 and 4.2% in DukeMTMC-ReID.

We minimize the KL divergence between the probability distribution predicted by the identity encoder and the CNN model trained with identity loss on the source dataset. Since the environment switched image should give a similar probability distribution for the output obtained from both the identity encoder and CNN model, we minimize the KL divergence between these output distributions. This loss helps to supervise the soft labeling of the generated images. Whereas, in the absence of KL loss, we apply Eq. 3.17 to obtain soft labels. Since the soft labels obtained by applying KL loss are richer in appearance information, the accuracy also improves. We also observe that the generated images are sharper and of better quality once KL loss is applied. Thus, in the absence of KL loss, low-quality images with poor representational soft labels drop accuracy compared to direct testing.

3.3.5 Evaluation of the Quality and Diversity of Generated Images

For performing the quantitative evaluation, we compare the synthetic data generated by IPES-GAN with SPGAN and CR-GAN. As shown in Table 3.6, compared to SPGAN and CRGAN, our method has lower FID (Heusel *et al.*, 2017) and higher LPIPS (Zhang *et al.*, 2018*c*). Thus, our method can generate images with better fidelity and diversity after translation between two entirely different data domains.

We also perform the Mean Opinion Score (MOS) test to assess the quality and diversity of generation methods. Specifically, we asked five reviewers to assign an integral score from 1 (bad quality/diversity) to 5 (excellent quality/diversity) to the generated images by different methods. Each reviewer rated 100 randomly selected images generated from DukeMTMC-ReID to Market-1501. The experimental results of the con-

Table 3.6: Quantitative evaluation to check quality of image. LPIPS: Learned perceptual image similarity, higher values indicate more similarity. FID: Fréchet Inception distance, lower the better.

Methods	M → D		D → M	
	LPIPS↑	FID↓	LPIPS↑	FID↓
SPGAN (Deng *et al.*, 2018)	0.099	0.171	0.099	0.115
CR-GAN (Chen *et al.*, 2019)	0.281	0.058	0.269	0.096
IPES-GAN	0.345	0.023	0.401	0.062

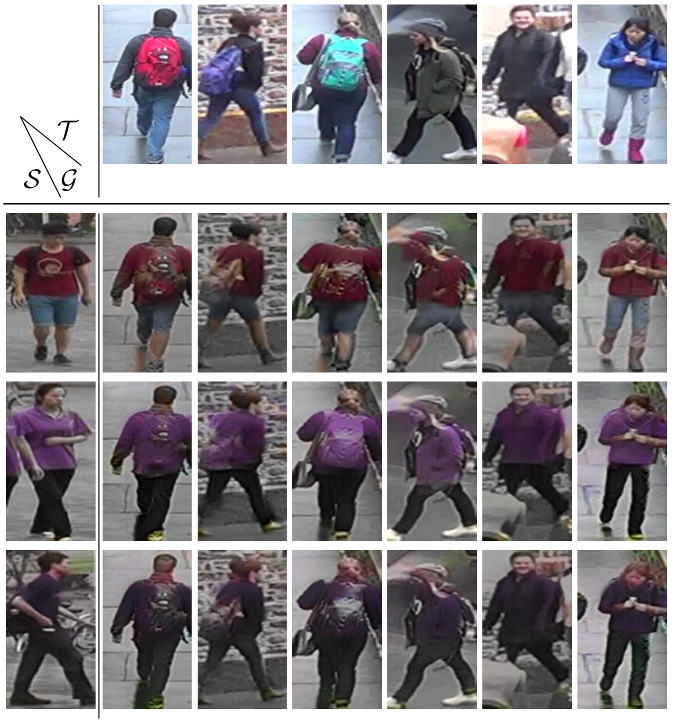


Figure 3.11: Visualization of translated images from the source (Market-1501, the first column) to the target (DukeMTMC-ReID, the first row). (**Best viewed in colors**).

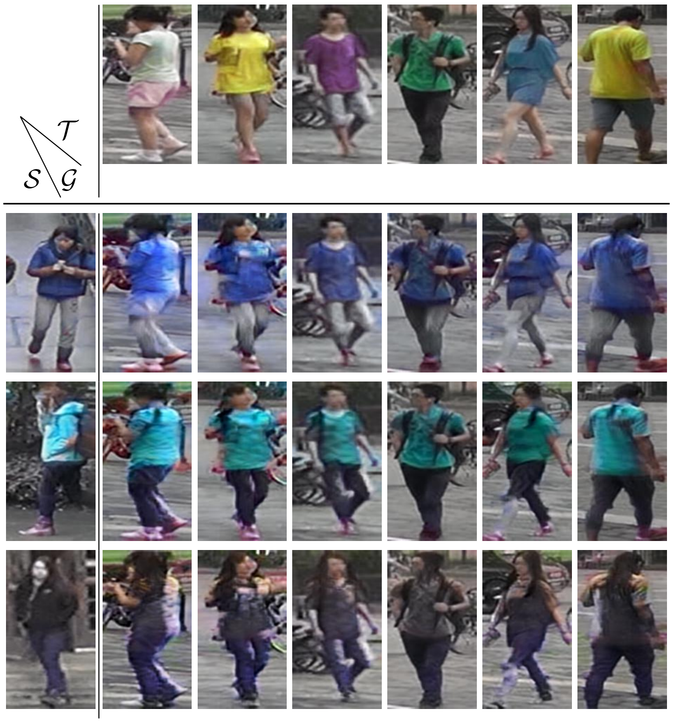


Figure 3.12: Visualization of translated images from the source (DukeMTMC-ReID, the first column) to the target (Market-1501, the first row). (**Best viewed in colors**).

Table 3.7: The Mean Opinion Score (MOS) test results on the generated images from DukeMTMC-ReID to Market-1501.

	SPGAN	SBSGAN	FD-GAN	CR-GAN	**IPES-GAN**
Quality	3.45	3.22	2.23	3.28	3.23
Diversity	1.95	1.44	2.89	2.15	3.34

Table 3.8: Maximum Mean Discrepancy (MMD). $\mathcal{S}$ and $\mathcal{T}$ represents source and target dataset, respectively. $\mathcal{S} \rightarrow \mathcal{T}$ ($\texttt{source} \rightarrow \texttt{target}$) represents the setting of training on the source dataset and testing on the target dataset. 'M', 'D' and 'MS' denote the Market-1501, DukeMTMC-ReID and MSMT17 datasets, respectively. **Red** and blue fonts indicates MMD after training and before training.

$\mathcal{T}$	$\mathcal{S}$	$\mathcal{S} \rightarrow \mathcal{T}$	MMD
D	M	-	0.295
D	-	M → D	**0.167**
M	D	-	0.306
M	-	D → M	**0.245**
MS	M	-	0.273
MS		M → MS	**0.183**
MS	D	-	0.280
MS	-	D → MS	**0.199**

ducted MOS tests are summarized in Table 3.7.

We use Maximum Mean Discrepancy (MMD) to quantify the domain gap between $\mathcal{T}$ (target) and $\mathcal{G}$ (generated) images, and $\mathcal{T}$ (target) and $\mathcal{S}$ (source) images. We show corresponding results in Table 3.8. MMD between DukeMTMC-ReID ($\mathcal{T}$) and Market-1501 ($\mathcal{S}$) before training the model is 0.295. After training, it reduces to 0.167 between DukeMTMC-ReID ($\mathcal{T}$) and (Market-1501 ($\mathcal{S}$) → DukeMTMC-ReID ($\mathcal{T}$)). Similarly, as illustrated in Table 3.8, MMD reduces after training for other dataset settings as well. This decrease in MMD shows a large difference between the source and target image styles before training, which leads to poor performance on the target dataset. However, after training, the source image adapts to the pose and background of the target image. Thus, source images with target environment style give better performance on target data. As we show, the generated images by our method obtain competitive quality and have the best diversity, comparing with all other representative methods.

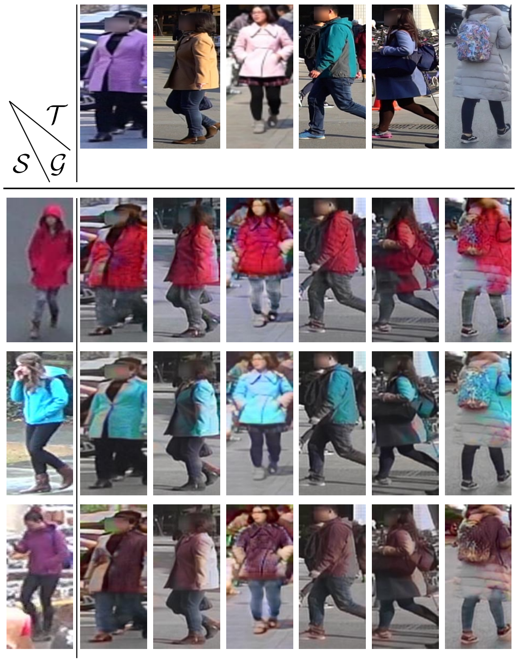


Figure 3.13: Visualization of translated images from the source (DukeMTMC-ReID, the first column) to the target (MSMT17, the first row). (**Best viewed in colors**).

3.3.6 Visualizations

We show the translated images in Figure 3.11 and 3.12. Figure 3.11 illustrates the results for translation from Market-1501 (source) to DukeMTMC-ReID (target). It is evident that the pose and background of the target image are adapted in the translated images while maintaining the appearance (color and style of clothes and shoes) of the source image. Similar observations can be inferred from translation of DukeMTMC-ReID (source) images to Market-1501 (target) in Figure 3.12.

In Figure 3.13, we illustrate the results for translation from Market-1501 (source) to MSMT17 (target). It is evident that the pose and background of the target image are adapted in the translated images while maintaining the appearance of the source image. Similar observations can be inferred from translation of DukeMTMC-ReID (source) images to MSMT17 (target) in Figure 3.14.

The Re-ID task primarily extracts features such as the color of clothes, additional useful cues such as a bag. While in some cases, cues such as bags may not be preserved, yet the clothing is efficiently adapted because few images have such additional cues, and the model may not be able to learn them. Further, the clothing appearance is the most significant cue, and the network primarily learns to match this. In some cases, gender may not be useful information as many shots are taken from different views, and quite

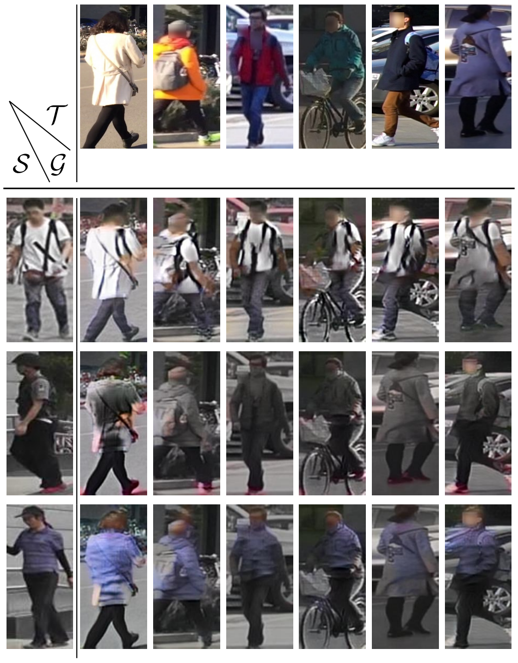


Figure 3.14: Visualization of translated images from the source (Market-1501, the first column) to the target (MSMT17, the first row). (**Best viewed in colors**).

a few are taken from an angle where identifying gender is very hard. Thus, while the network may inherently learn some gender-based cues, it may not consider it as a strong discriminative factor. Unlike previous approaches, our approach does not need separate training for image generation and discriminative learning, which reduces complexity.

Visualization under major occlusion In our unsupervised domain adaptation framework for person re-identification (re-id) task, we introduce a robust approach that effectively handles variations in camera viewing angles and occlusion within the target domain while preserving identities. By incorporating identity-preserving and environmental-switching mechanisms, our model adapts to diverse camera perspectives. The integration of pose and background style adaptation enables the model to generalize across different viewing angles. Leveraging shared identity features and private environment features, the model learns domain-invariant representations, ensuring resilience to background style differences. Our comprehensive approach shows promising results in addressing challenges posed by camera angles and occlusion, making it highly suitable for real-world re-id applications.

As per your suggestion, we provide a detailed demonstration of the robustness of our proposed model in handling a major occlusion scenario, as depicted in Figure 3.15. Despite the challenging conditions posed by significant occlusion, our model exhibits exceptional performance by accurately retrieving a majority of the correct samples

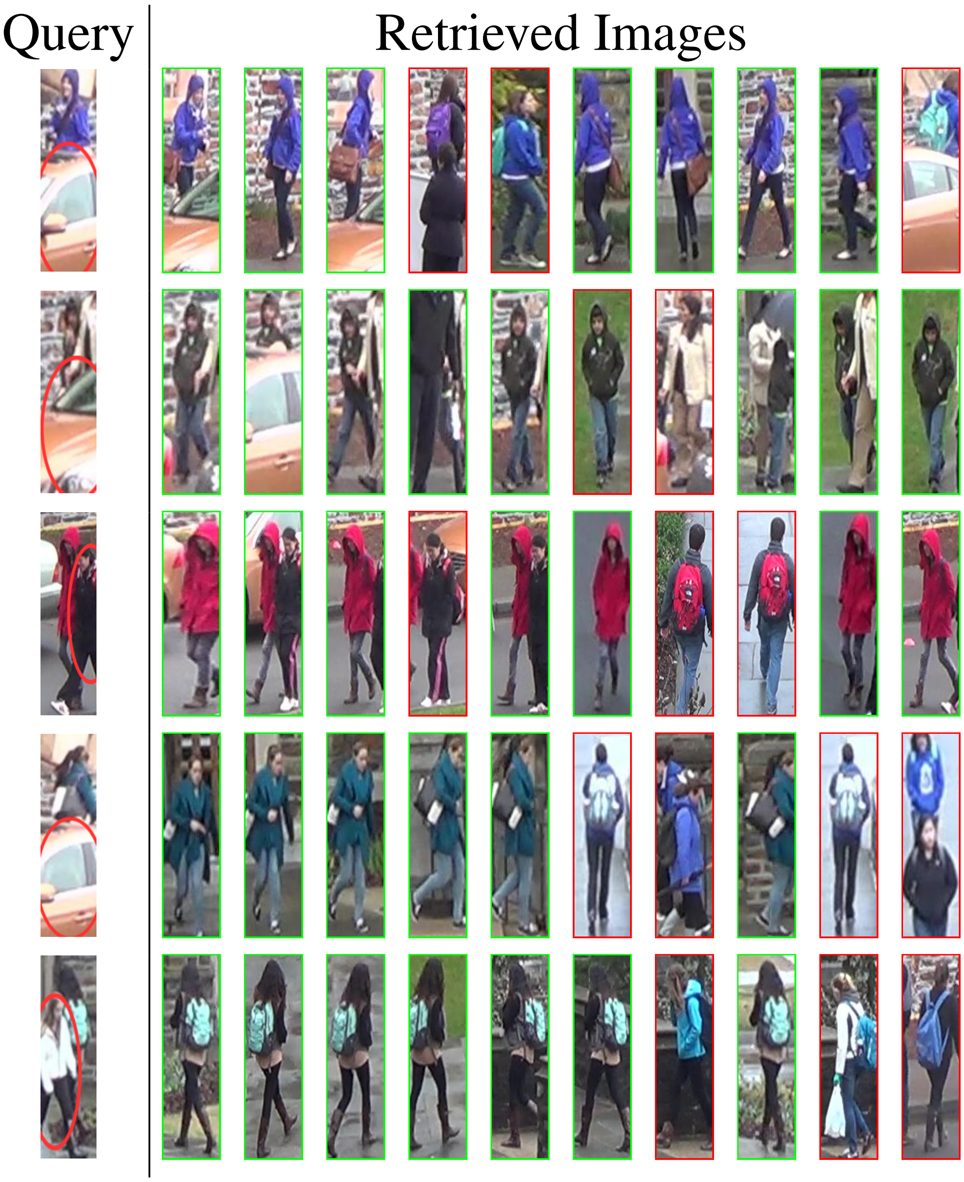
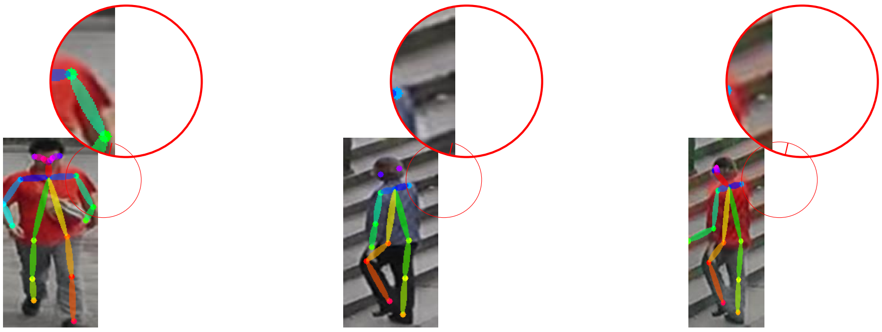


Figure 3.15: The top-10 predictions made by our proposed IPES-GAN model with a majorly occluded image as query. The green boxes represent correct predictions of query image from the gallery set and red represents the incorrect predictions. Red Circles in the query images represent the occlusion.



(a) Source (M)

(b) Target (D)

(c) Generated Image (M → D)

Figure 3.16: The figure shows the *Generated Image (M → D)* adapts to pose and background of *Target (D)* image while maintaining identity-related features of the *Source (M)* image. M and D denotes Market-1501 and DukeMTMC-ReID. (**Best viewed in colors**).

in the top-10 predictions. This outcome underscores the efficacy and reliability of our model in real-world scenarios, where occlusion is a common and complex challenge. Figure 3.15 visually illustrates the impressive retrieval results, highlighting the model's ability to maintain high accuracy even in the presence of substantial occlusion, thereby validating its practical applicability and significance in the field of person re-identification.

3.3.7 Background Adaptation of target domain

We illustrate in Figure 3.16 the magnified backgrounds of source, target, and generated image. It is visible that generated image (source → target) completely adapts to the background of the target image while preserving its identity-related features (appearance). We also superimpose the pose of the source, target, and generated image in this figure to show that generated images consist of the background and pose style of the target image.

3.4 Chapter Summary

This chapter proposes a novel person image generation network for the person re-identification task. Our model performs both generative and discriminative learning simultaneously. The proposed IPES-GAN encodes the pose information to incorporate the global structure of the target image. The target background and camera style are captured via environment encoder and camera style loss. The cycle consistency and adversarial loss further optimize the target image generation process. In order to obtain discriminative features for Re-ID, we apply soft cross-entropy loss and KL divergence loss. The elaborate qualitative and quantitative experiments demonstrate that our proposed approach generates images with better fidelity and diversity and achieves state-of-the-art Re-ID performance.

In this chapter, we explore the application of unsupervised domain adaptation to extend the generalizability of Re-ID models trained on a source dataset to an unlabeled target dataset. However, the efficacy of these models is significantly reduced when subject to adversarial perturbations. To address this issue, the subsequent chapter focuses on examining the robustness of these Re-ID models. In contrast to conventional adversaries that solely consider Euclidean space and ignore pixel geometry, our method incorporates the Wasserstein metric attack, resulting in a more potent attack.

CHAPTER 4

Adversarial Attack on Re-ID

4.1 Introduction

Deep neural networks (DNNs) have demonstrated tremendous performance improvement in Re-ID (Zhang *et al.*, 2017*b*; Sun *et al.*, 2018; Rombach *et al.*, 2022; Song *et al.*, 2023). However, they are vulnerable to adversarial attacks. Adversarial attacks have been extensively investigated on tasks which fall under closed-set setting like classification, object detection and segmentation (Szegedy *et al.*, 2013; Kurakin *et al.*, 2016; Dong *et al.*, 2018*b*; Guo *et al.*, 2020; Moosavi-Dezfooli *et al.*, 2016; Carlini and Wagner, 2017; Athalye *et al.*, 2018*a*; Su *et al.*, 2019; Yuan *et al.*, 2021; Sun *et al.*, 2022; Chen and Gu, 2020; Tashiro *et al.*, 2020; Li *et al.*, 2021*d*, 2023). However, except for a few attempts (Zheng *et al.*, 2018; Li *et al.*, 2019*b*; Zhao *et al.*, 2019*b*; Bai *et al.*, 2020*b*; Wang *et al.*, 2020*a*; Ding *et al.*, 2021; Yang *et al.*, 2021*a*; Zhao *et al.*, 2022), adversarial attacks have not been much investigated in open-set retrieval problems like person re-identification where the source and target dataset have completely non-overlapping labels (Refer Chapter 2 for adversarial attack formulation).

Previous works in the field of adversarial attack on person re-identification focus on attack based on l_∞ perturbations and its corresponding l_p generalization (Bai *et al.*, 2020*b*; Wang *et al.*, 2020*a*; Ding *et al.*, 2021; Yang *et al.*, 2021*a*; Zhao *et al.*, 2022). These works are based on projected gradient descent (PGD) method to find perturbations within a small l_p radius (Goodfellow *et al.*, 2014*b*). l_p threat model is known to be a poor metric to measure similarity of images which adjusts each pixel value in images independent of other pixels (Wu *et al.*, 2020*a*). Images which look similar under human perception are not necessarily close under l_p norm (Wong *et al.*, 2019). On the other hand, Wasserstein metric is a more perceptually-aligned metric for images (Hu *et al.*, 2020) (Peyré *et al.*, 2019). The set of allowable perturbations can differ significantly between the Wasserstein ball and the l_p ball. It is possible for examples that are near in terms of Wasserstein distance (Δ_W) to be far apart in l_p distance (Δ_∞), and conversely, for examples that are close in Δ_∞ to be far apart in terms of Δ_W as shown in Figure 4.1.

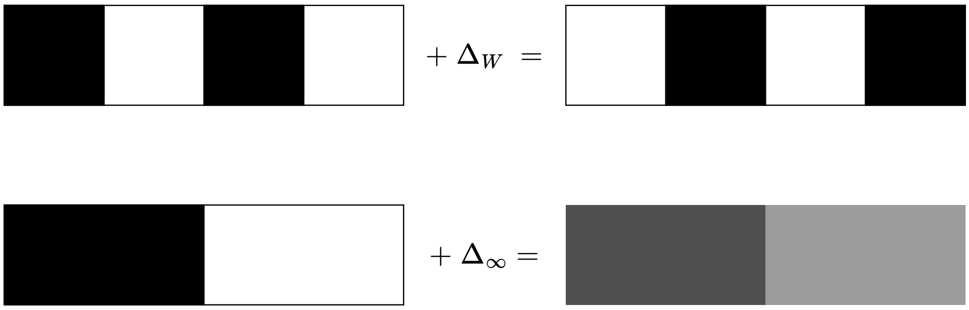


Figure 4.1: We show images with four pixels to demonstrate the difference between Wasserstein perturbations and Δ_∞ perturbations. A small perturbation Δ_W shifts the image one pixel to the right, which is minimal in Wasserstein distance but maximal in Δ_∞ distance. In contrast, a small perturbation Δ_∞ changes all pixels to be grayer, which is minimal in Δ_∞ distance but maximal in Wasserstein distance.

Wasserstein metric based perturbations provide more generalized image perturbations in the form of pixel mass movement. It redistributes the pixel mass instead of dealing with each pixel independently, as in case of l_p norm based threat models. Previous works in Re-ID mainly include l_∞ metric-based attacks (Bai *et al.*, 2020*b*; Zheng *et al.*, 2018), GAN-based attacks (Zhao *et al.*, 2019*b*; Wang *et al.*, 2020*a*) or meta-learning based attacks (Yang *et al.*, 2021*a*). However, to the best of our knowledge, Wassterstein ball perturbations are not yet proposed for open-set ranking problems like Re-ID. Wasserstein metric is used in few previous works against classification task (Wu *et al.*, 2020*a*; Wong *et al.*, 2019; Hu *et al.*, 2020). (Wong *et al.*, 2019) introduced Wasserstein perturbations as an alternative to l_p threat model for closed-set classification problem. An improvement over this work was proposed by (Hu *et al.*, 2020). However, we cannot directly apply the existing attack methods against classification tasks to open-set problems like Re-ID systems (Zheng *et al.*, 2018; Li *et al.*, 2019*b*).

In this chapter, we propose to obtain an adversarially perturbed sample by projecting clean images into the Wasserstein ball. Our approach is used to attack the entire ranking model in an open-set setting which is much more challenging problem than a closed-set task. Unlike previous attack methods on person Re-ID which deal with finding the adversarial samples in the l_p ball, we project adversarial samples on Wassterstein ball. Our method does not require any training to learn perturbations unlike previous approaches (Wang *et al.*, 2020*a*; Yang *et al.*, 2021*a*).

We provide an illustration of our Wasserstein non-targeted and targeted attacks in Figure 4.2. The aim of non-targeted attack is to increase the distance between the features of query and gallery samples of same identity. Different from non-targeted attack, targeted attack tries to minimize the distance between the features of query and a target adversary with identity different than that of query image.

Our contributions can be summarized as:

- To the best of our knowledge, we are the first to introduce Wasserstein metric for adversarial attack on Re-ID.
- We iteratively perturb the query images by performing l_∞ perturbation as the first step and then projecting the adversarial sample in the Wasserstein ball of radius ϵ followed by clamping so that perturbation lies in $[0, 1]$ pixel space.
- We show that our Wasserstein threat model can be easily generalized to attack state-of-the-art Re-ID models and unseen dataset scenarios.

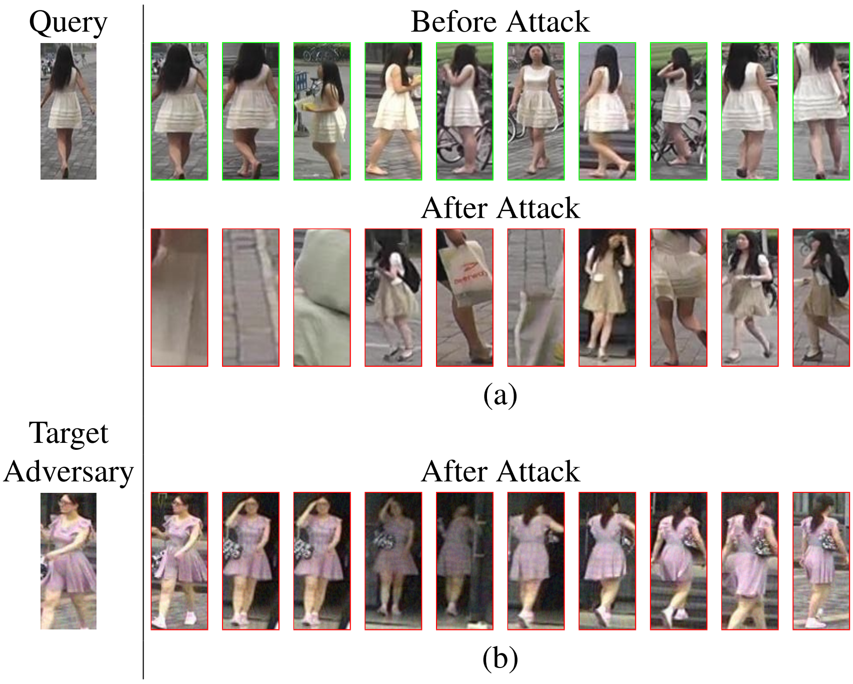


Figure 4.2: The top-10 predictions before and after our (a) non-targeted (b) targeted Wasserstein attack. As is visible in (a) after attack the top-10 predictions can be any other sample except the same identity (ID) as query whereas in (b) same IDs as target adversary sample are mainly included in the ranking list, degrading the performance of Re-ID model. The green boxes represent correct predictions of query image from the gallery set and red represents the incorrect predictions.

4.2 Preliminaries

Wasserstein Distance (W_D). is the minimum cost required to change the distributions with the help of probability mass. Let I, $\widetilde{I} \in \mathbb{R}^n_+$ be two vectorized images with equal mass such that $1^\top I = 1^\top \widetilde{I}$, which are converted to marginals by performing normalization, such that $\sum_i I_i$ = $\sum_i \widetilde{I}_i = 1$. Then, Wasserstein distance is defined as:

$$\begin{aligned} W_D(I, \widetilde{I}) &= \min_{T \in \mathbb{R}_+^{n \times n}} \langle T, C \rangle, \\ \text{s.t.} \quad T1 = I, &\quad T^\top 1 = \widetilde{I}, \end{aligned} \tag{4.1}$$

$\langle T, C \rangle$ represents inner product between T and C. T, C belong to $\mathbb{R}_+^{n \times n}$, T is the transportation plan and C is the cost matrix. T represents the amount of mass and C represents the cost of moving the mass between the pixels.

Wasserstein Ball (B_W)**.** The Wasserstein ball corresponding to Wasserstein distance is as shown in Figure 4.3. It is given as:

$$\widetilde{I} = I + \delta; B_W = \{\widetilde{I} : W_D(I, \widetilde{I}) \leq \epsilon\}. \tag{4.2}$$

where δ is the noise added to the clean image to obtain adversarial sample.

Transport Plan. Transport plan allows to move mass between any pair of pixels. We apply a local cost matrix and restrict the movement of mass within a kernel size of $k \times k$ of the original image.

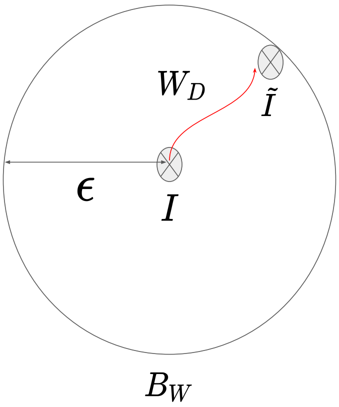


Figure 4.3: Wasserstein Ball.

4.3 Methodology

4.3.1 Wasserstein Metric Attack

Let the query image be I, gallery image be I_g and a Re-ID model f. The feature vector of query and gallery images are given as $f(I)$ and $f(I_g)$. In order to attack Re-ID task we employ a metric loss between the feature vectors to reduce the performance of Re-ID models. We push the features of generated adversarial sample $\widetilde{I} = I + \delta$, where I is the query image and δ is the noise, away from the features of same identity gallery image with the help of our metric loss function and project this sample on the Wasserstein ball of radius ϵ.

Metric Loss. We use Euclidean distance to increase the distance between the features of perturbed query images and images in gallery database containing the same identity as that of query image. It is given as,

$$l(f(I_g), f(\widetilde{I})) = f(I_g) - f(\widetilde{I}) \tag{4.3}$$

4.3.2 Iterative Wasserstein Space Projection

We perturb the query image in l_p space iteratively as given below,

$$\hat{I}^{t+1} = \widetilde{I}^t + \alpha sign(\frac{\partial l(f(I_g), f(\widetilde{I}^t))}{\partial f(\widetilde{I}^t)}), \tag{4.4}$$

where α is the step-size and t is the iteration number. $\hat{I}^{t+1}$ is the perturbed image and $\tilde{I}^0 = I$.

Wasserstein Projection. Projecting the perturbed image $\hat{I}$ onto the Wasserstein ball within ϵ distance of original image I can be defined as,

$$\widetilde{I}^{t+1} = \underset{B_W(I,\epsilon)}{proj}\,(\hat{I}^{t+1}), \tag{4.5}$$

where B_W is the Wasserstein ball with perturbation radius ϵ. $\widetilde{I}^{t+1}$ is the projected adversarial query image corresponding to the clean image obtained after t iterations. Wasserstein space projection can be obtained by solving the entropy-regularized optimal transport problem,

$$\begin{aligned} \min_{\widetilde{I}\in\mathbb{R}^n_+, T\in\mathbb{R}^{n\times n}_+} \quad & \frac{1}{2}\|\hat{I}-\widetilde{I}\|_2^2 + \frac{1}{\lambda}\sum_{ij} T_{ij}\log(T_{ij}) \\ \text{s.t.} \quad & T1 = I, \quad T^\top 1 = \widetilde{I} \quad \langle T, C\rangle \leq \epsilon, \end{aligned} \tag{4.6}$$

where the component $\frac{1}{2}\|\hat{I}-\widetilde{I}\|_2^2$ ensures that the output sample is close to the perturbed sample $\hat{I}$ in l_2 sense and at the same time lies in Wasserstein ball of radius ϵ with respect to clean sample I.

Dual of Entropy Regularized Optimization. An equivalent dual problem can be formed for the above entropy regularized optimization with the help of Lagrange multipliers where dual variables x, y and z can be introduced. It is given as (Wong *et al.*, 2019),

$$\underset{x,y\in\mathbb{R}^n, z\in\mathbb{R}_+}{max} \quad F(x,y,z) \tag{4.7}$$

where,

$$\begin{aligned} F(x,y,z) = & -\frac{1}{2\lambda}\|y\|_2^2 - z\epsilon + x^T I + y^T\hat{I} \\ & -\sum_{ij} e^{(x_i)}e^{(-zC_{ij}-1)}e^{(y_j)}. \end{aligned} \tag{4.8}$$

$F(x,y,z)$ can be maximized w.r.t. x, y and z to obtain these dual variables (Wong *et al.*, 2019). These dual variables can be substituted to obtain the adversarial sample $\widetilde{I}$ and transport plan T which are,

$$\begin{aligned} \widetilde{I}_i &= \hat{I}_i - \frac{y_i}{\lambda}. \\ T_{ij} &= e^{(x_i)}e^{(-zC_{ij}-1)}e^{(y_j)}. \end{aligned} \tag{4.9}$$

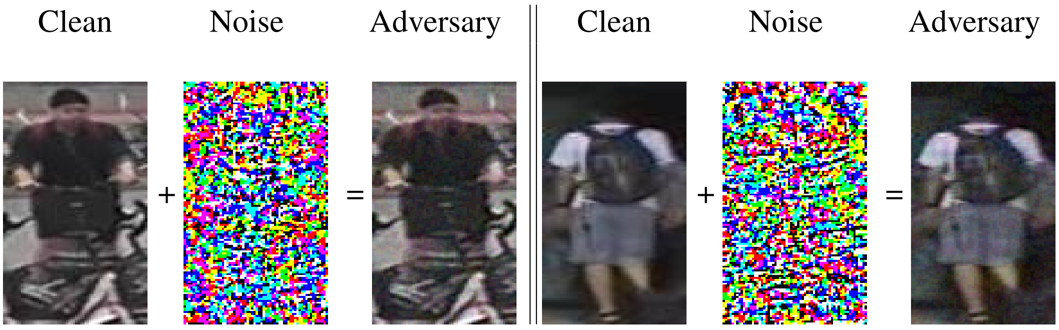


Figure 4.4: The figure illustrates the adversarial sample obtained by adding noise to clean images.

After projecting the perturbed image to Wasserstein space, we perform clamping to ensure that the adversarial sample is a valid image with pixels in the range $[0, 1]$.

In Figure 4.4, we show the adversarial sample corresponding to the clean image obtained after N iterations. We describe the Wasserstein threat model in Algorithm 1.

ALGORITHM 1: Wasserstein Threat Model

```
Require: Input query image $I$, gallery image $I_g$, Wasserstein radius $\epsilon$, metric loss $l$,
    iteration step size $\alpha$, $N$ is the total number of iterations, $\Gamma_I^\epsilon$ is the clip function
Ensure: Adversarial Sample $\widetilde{I}$
for $t = 0$ to $N$ : do
    $\nabla_W^t \leftarrow \frac{\partial l(f(I_g), f(\widetilde{I}^t))}{\partial f(\widetilde{I}^t)}$
    $\hat{I}^t \leftarrow \widetilde{I}^t + \alpha sign(\nabla_W^t)$
    $\widetilde{I}^t \leftarrow \underset{B_W(I,\epsilon)}{proj}(\hat{I}^t)$
    $\widetilde{I}^t \leftarrow \Gamma_I^\epsilon(\widetilde{I}^t)$
end for
```

4.4 Experiment

We discuss the datasets, evaluation metrics and experimental settings in this section.

4.4.1 Datasets and Settings

Datasets. Our proposed approach is tested on four large-scale Re-ID datasets: Market-1501 (Zheng *et al.*, 2015), DukeMTMC-ReID (Ristani *et al.*, 2016), MSMT17 (Wei *et al.*, 2018), and CUHK03 (Li *et al.*, 2014).

CUHK03 (Li *et al.*, 2014). It comprises of 14,097 images depicting 1,467 distinct individuals. To collect these images, 6 cameras were strategically placed across the campus, with each individual being captured by 2 of these cameras. The dataset contains two types of annotations: manually labeled bounding boxes and bounding boxes generated by an automatic detector. Additionally, the dataset offers 20 randomized train/test splits, where 100 of the identities are reserved for testing and the remainder are used for training.[1]

Implementation Details. We use a local cost matrix of 7×7 and entropy regularization constant of 3000 for all the experiments. We provide results for perturbation radii $\epsilon = 8$ and $\epsilon = 16$. We have used l_2 norm of distance in pixel space suppose going from pixels (i, j) to (k, l) for calculating cost matrix, thus we are using 1-Wasserstein distance. The

[1] The details of Market-1501, DukeMTMC-ReID and MSMT17 datasets are present in Chapter 3.

Table 4.1: White-Box Attack comparison with ODFA (Zheng *et al.*, 2018), TCIAA (Wang *et al.*, 2020*a*), UAP (Li *et al.*, 2019*b*) and Meta-Attack (Yang *et al.*, 2021*a*)

Market-1501 (Zheng *et al.*, 2015)						
Re-ID Models	mAP (%)					
	Before	ODFA	TCIAA	UAP	Meta-Attack	Ours
IDE (Zheng *et al.*, 2016*b*)	63.3	25.6	16.9	3.1	3.6	**0.4**
PCB (Sun *et al.*, 2018)	70.7	-	22.4	10.7	10.9	**0.9**
DukeMTMC-ReID (Ristani *et al.*, 2016)						
Re-ID Models	mAP (%)					
	Before	ODFA	TCIAA	UAP	Meta-Attack	Ours
IDE (Zheng *et al.*, 2016*b*)	66.7	23.5	17.6	4.2	3.6	**1.4**
PCB (Sun *et al.*, 2018)	68.0	-	25.2	14.3	11.2	**1.6**

maximum perturbation radius (ϵ) allowed is ≤ 10 for projected gradient descent on Wasserstein ball, unless otherwise stated. ϵ is the upper bound applied on the generated noise which determines the attack intensity of the threat model and visual quality of adversarial samples. We have performed all the qualitative results for non-targeted attack.

4.4.2 White-Box Attack

In white-box attack setting, the architecture and parameters θ of target model are known to us.

Comparison with SOTA Re-ID attack methods. We show in Table 4.1 comparison with other methods, and our performance exceeds by 3.2% on base model IDE and by 10% on robust model like PCB for Market1501 dataset and similarly our model outperforms by a significant range for DukeMTMC-ReID dataset.

Comparison with TCIAA on Re-ID backbones, Part-based and Data-augmentation based SOTA Re-ID models. We compare our performance with TCIAA for $\epsilon = 8$ and $\epsilon = 16$ in Table 4.2a for Market-1501 dataset. Similarly, we show results for DukeMTMC-ReID and MSMT17 in Tables 4.2b and 4.3, respectively. We can observe from the results that compared to TCIAA, our threat model decreases the performance of backbone models sharply, reaching almost to zero for $\epsilon = 16$. We also get robust attack performance for $\epsilon = 8$. Ensembling local features or data-augmentation methods also fail to provide defense against our attack. Our attack proves effective compared to

Table 4.2: White-Box Attack comparison with TCIAA (Wang *et al.*, 2020*a*) for perturbation radius ($\epsilon = 8$). Lower the rank accuracy, better is the performance of the threat model.

Market-1501 (Zheng *et al.*, 2015)

$\epsilon = 8$

Method	CMC-1 (%)								
	IDE	Inception-v3	AlignedRe-ID	PCB	HACNN	SPGAN	CamStyle	HHL	LSRO
Before	83.5	73.0	90.8	88.6	90.6	85.4	87.7	84.7	90.2
TCIAA	14.8	6.4	21.3	23.1	8.0	13.0	22.5	17.5	6.9
Ours	**1.6**	**0.2**	**0.9**	**1.6**	**1.9**	**1.9**	**1.1**	**1.3**	**0.4**

$\epsilon = 16$

Method	CMC-1 (%)								
	IDE	Inception-v3	AlignedRe-ID	PCB	HACNN	SPGAN	CamStyle	HHL	LSRO
Before	83.5	73.0	90.8	88.6	90.6	85.4	87.7	84.7	90.2
TCIAA	3.7	1.7	1.4	5.0	0.9	1.5	3.9	3.6	0.9
Ours	**0.1**	**0.0**	**0.3**	**0.2**	**0.2**	**0.4**	**0.2**	**0.2**	**0.1**

(a)

DukeMTMC-ReID (Ristani *et al.*, 2016)

	$\epsilon = 8$					$\epsilon = 16$				
Method	CMC-1 (%)					CMC-1 (%)				
	IDE	PCB	CamStyle	HHL	LSRO	IDE	PCB	CamStyle	HHL	LSRO
Before	86.2	85.8	76.4	72.4	72.0	86.2	85.8	76.4	72.4	72.0
TCIAA	15.3	23.9	12.6	8.1	9.2	1.3	1.4	1.2	1.0	0.7
Ours	**1.7**	**1.8**	**0.9**	**0.5**	**0.3**	**0.2**	**0.2**	**0.4**	**0.0**	**0.0**

(b)

TCIAA on all SOTA Re-ID systems. It is clearly visible from the results that in many cases our performance with smaller perturbation radius is better than previous works for larger radius.

Visualization. We show the original and corresponding adversarial samples generated by TCIAA (Wang *et al.*, 2020*a*) and our method in Figure 4.5. Compared to TCIAA, our Wasserstein threat model has better attack performance as well as it can generate high quality adversarial samples introducing imperceptible noise. It is also visible that TCIAA's quality of image is poor as we increase the perturbation radius as compared to our method. We also provide a quantitative analysis of comparison of the image quality with previous works in Section 4.5.

Performance on CUHK03 dataset. Table 4.4 shows comparison with state-of-the-art attack methods when attacking Re-ID backbones and robust Re-ID models.

Table 4.3: White-Box Attack comparison with TCIAA (Wang *et al.*, 2020*a*) on models trained on MSMT17 (Wei *et al.*, 2018).

	MSMT17							
	$\epsilon = 8$				$\epsilon = 16$			
	AlignedRe-ID		IDE		AlignedRe-ID		IDE	
	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1
Before	41.6	82.3	34.2	83.5	41.6	82.3	34.2	83.5
TCIAA	12.3	12.1	11.3	12.6	1.2	1.8	0.1	0.6
Ours	**0.5**	**1.3**	**0.1**	**0.1**	**0.1**	**0.1**	**0.0**	**0.0**

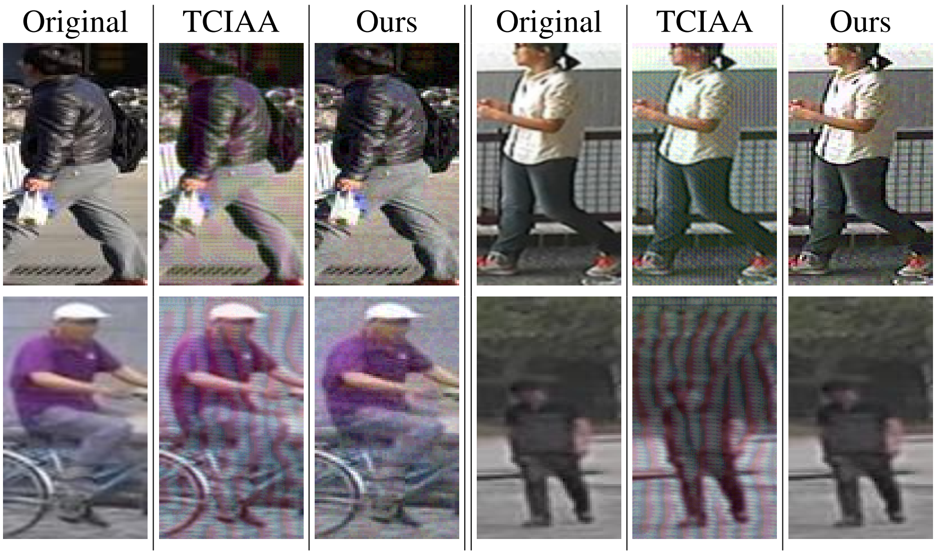


Figure 4.5: The adversarial samples obtained after attack. Above row shows images for perturbation radius $\epsilon = 8$ which are better than below row with radius $\epsilon = 16$.

4.4.3 Black-Box Attack

Black-box setting is a challenging problem as the attacker has no knowledge of target data or parameters of the target model. We perform a cross-model attack in black-box setting.

Cross-Model attack. In this attack method, noise is learned by generating adversarial samples with a known model and is used to attack an unknown model. We show an ablation study of our Wasserstein cross-model attack performance on three backbone models, IDE (Zheng *et al.*, 2016*b*), DenseNet-121 (Huang *et al.*, 2017) and Inception-v3 (Qian *et al.*, 2017) in Table 4.5a. Among the three backbones, Inception-v3 is most robust against our attack.

We compare our cross-model attack with SOTA attack method TCIAA (Wang *et al.*, 2020*a*) in Table 4.5b. Our overall cross-model attack performance is better than TCIAA for perturbation radius of $\epsilon = 16$ on SOTA Re-ID models as presented in Table 4.5b.

Table 4.4: Attacking Re-ID Backbones. IDE(ResNet-50) (Zheng *et al.*, 2016*b*) and Inception-v3 (Qian *et al.*, 2017) trained on CUHK03 dataset. We compare our proposed approach with GAP (Poursaeed *et al.*, 2018),U-PGD (Madry *et al.*, 2017), TCIAA (Wang *et al.*, 2020*a*) ($\epsilon = 16$).

Re-ID Models	mAP (%)					CMC-1 (%)				
	Before	GAP	U-PGD	TCIAA	Ours	Before	GAP	U-PGD	TCIAA	Ours
IDE	24.5	1.3	0.8	0.9	**0.1**	24.9	0.9	0.8	0.4	**0.0**
Inception-v3	30.1	2.0	0.8	0.3	**0.1**	32.1	1.1	0.4	0.1	**0.0**
Re-ID Models	mAP (%)					CMC-1 (%)				
	Before	GAP	U-PGD	TCIAA	Ours	Before	GAP	U-PGD	TCIAA	Ours
AlignedRe-ID	59.6	3.4	2.1	2.1	**0.2**	61.5	2.1	1.4	1.4	**0.1**
HACNN	47.6	1.8	0.8	0.4	**0.3**	48.0	0.9	0.4	0.1	**0.2**

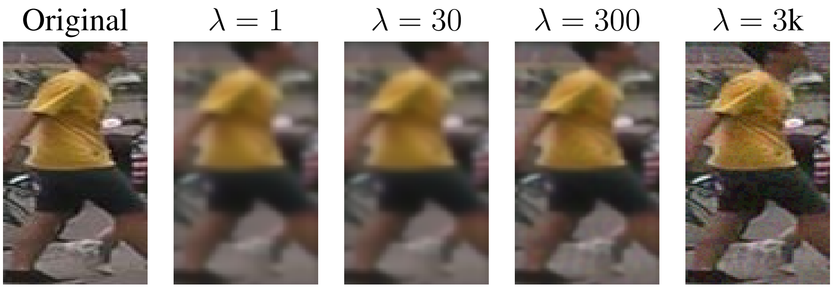


Figure 4.6: The adversarial samples generated for different values of regularization parameter λ.

HACNN seems to be more robust than other models, thus indicating that attention networks are better for robustness of Re-ID models.

4.4.4 Effect of Regularization Parameter

It is important to make sure that regularization parameter (λ) does not cause blurring of the adversarial samples as shown in Figure 4.6. It is clearly visible that for $\lambda = 3000$, we get better quality adversarial samples. The smaller value of regularization causes blurring in the images resulting in poor quality adversarial perturbations. Lower value of λ also makes tougher to find adversarial perturbations at lower values of radius (ϵ) in Wasserstein ball, thereby increasing the ϵ value.

4.4.5 Effect of size of local transport plan

Transport plan describes the movement of pixels from (i, j) to (k, l) with minimum cost. We show the performance of our Wasserstein attack for different sizes of transport

Table 4.5: **Cross-Model Attack:** Known N/W → Unknown N/W denotes the noise is learned by a known model and tested on an unknown model. All experiments carried out on Market-1501 (Zheng *et al.*, 2015).

Known N/W	Unknown N/W	R-1	R-5	R-10	mAP
IDE	→ DenseNet-121	15.3	30.2	39.1	11.5
	→ Inception-v3	36.5	58.2	67.1	18.9
DenseNet-121	→ IDE	36.8	57.3	65.7	19.4
	→ Inception-v3	51.2	72.8	80.3	29.2
Inception-v3	→ IDE	25.3	46.7	53.4	15.0
	→ DenseNet-121	14.5	22.7	38.4	6.8

(a) Ablation study of our cross-model attack on backbone networks IDE (ResNet-50) (Zheng *et al.*, 2016*b*), DenseNet-121 (Huang *et al.*, 2017) and InceptionNet-v3 (Qian *et al.*, 2017) ($\epsilon = 8$).

Known N/W	Unknown N/W	mAP (%)		CMC-1 (%)	
		TCIAA	Ours	TCIAA	Ours
AlignedRe-ID	→ PCB	31.7	**4.1**	22.9	**9.0**
	→ HACNN	13.4	**11.4**	14.8	**14.5**
	→ LSRO	14.8	**2.9**	17.0	**3.0**

(b) The mAP (%) and CMC-1 (%) comparison of our Cross-Model Attack on SOTA Re-ID models with TCIAA ($\epsilon = 16$).

plans in Figure 4.7. As it is clearly visible from the figure that transport plan of 7×7 gives best attack performance. Increasing the transport plan to 9×9 is not able to attack models as effectively, possibly due to the movement of more pixels by a large amount. Thus, with the help of hyper-parameter tuning, 7×7 transport plan gives us optimum attack performance. The attack performance increases for large values of ϵ reaching to almost zero.

4.5 Image Quality

SSIM (Sheikh *et al.*, 2004) is an image quality assessment metric which is used to measure the similarity between the images. Large SSIM score between real and synthetic images indicate that synthetic images have high quality and less noise. In our case, larger SSIM score indicates that high quality adversarial samples are generated with less distortion. We show the SSIM scores in Table 4.6. The results prove that in comparison to previous approaches we generate high quality images which are able to fool Re-ID models easily. We also provide maximum mean discrepancy (MMD). In comparison to TCIAA, our model produces adversarial examples with larger MMD. This

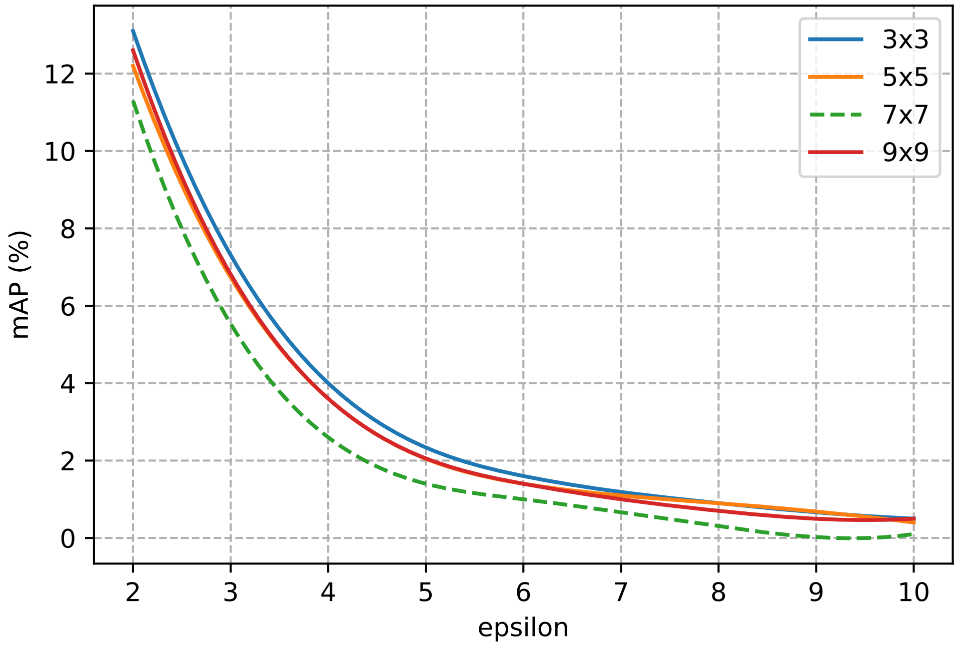


Figure 4.7: Attack performance for different transport sizes.

means that the domain of adversarial samples is farther away. Since our model achieves both larger SSIM and MMD, it can be inferred that it produces stronger adversarial attack with better perceptual quality.

Table 4.6: SSIM scores and MMD to compare the quality of adversarial samples ($\epsilon = 8$) on Market-1501 (Zheng *et al.*, 2015) and DukeMTMC-ReID (Ristani *et al.*, 2016) datasets. Larger SSIM scores coupled with larger MMD values indicates better performance.

		Market-1501	DukeMTMC-ReID
SSIM	TCIAA (Wang *et al.*, 2020*a*)	0.1889	0.1985
	Meta-Attack (Yang *et al.*, 2021*a*)	0.1963	0.2121
	Ours	**0.8198**	**0.7893**
MMD	TCIAA	0.4714	0.2813
	Ours	**0.4802**	**0.3867**

4.6 Chapter Summary

In this chapter, we propose threat model based on Wasserstein distance for person Re-Identification task. This perturbation is based on Wasserstein ball which is different from previous Re-ID works with l_∞ or more general l_p perturbation. Our Wasserstein attack does not require training to learn the noise. It generates the adversarial samples by adding noise to the clean image, and projecting the perturbations in Wasserstein ball. We perform extensive attacks on various backbones and SOTA Re-ID models trained on various datasets. Our performance in white-box as well as on more challenging black-box setting proves that our attack is very efficient.

In the method proposed in this chapter, we developed adversarial attack for analyzing the robustness of Re-ID models in white-box and black-box settings. In the next chapter, in order to defend against such adversarial attacks, we propose adversarial defense.

CHAPTER 5

Re-ID Defense

5.1 Introduction

DNNs show very good performance in closed-set and open-set computer vision tasks. However, it is a major concern that these DNNs are vulnerable to adversarial attacks. In order to make sure that the performance of DNNs does not get affected, an adversarial defense mechanism is proposed by many works. Previous works have proposed approaches like feature denoising (Xie *et al.*, 2019), adversarial training, and distillation (Goodfellow *et al.*, 2014*b*; Papernot and McDaniel, 2016; Madry *et al.*, 2017), fast adversarial training (Rice *et al.*, 2020; Chen *et al.*, 2022*b*; Jin *et al.*, 2022). Other approaches use techniques such as meta-learning (Goldblum *et al.*, 2020; Bartler *et al.*, 2022) or addition of noise (He *et al.*, 2019*b*; Eustratiadis *et al.*, 2021; Byun *et al.*, 2022) to the network to provide robustness. However, these approaches are mainly for closed-set tasks. To our knowledge, very few works have proposed adversarial defense for an open-set task like object Re-ID.

In this chapter, we propose a meta perturbed stochastic neural network (MP-SNN) to provide robust defense against adversarial attacks. MP-SNN learns anisotropic and isotropic noise distribution in a novel meta-learning defense algorithm. Previous works show that isotropic or anisotropic noise injection leads to better generalizability (Jeddi *et al.*, 2020; Eustratiadis *et al.*, 2021). However, the use of isotropic noise suffers from a sharp degradation in clean performance. At the same time, anisotropic noise fails to achieve competitive adversarial performance. Therefore, in contrast to previous works, which inject either isotropic or anisotropic noise (Jeddi *et al.*, 2020; Eustratiadis *et al.*, 2021), we make use of both noises, thus providing a richer noise distribution for a more challenging Re-ID task.

Our MP-SNN is trained with a novel meta-learning strategy where we propose tasks as **V**anilla, **P**erturbed, and **P**erturbed-adversarial training, respectively. We aim to increase the robustness and generalizability of our model by adapting well to these

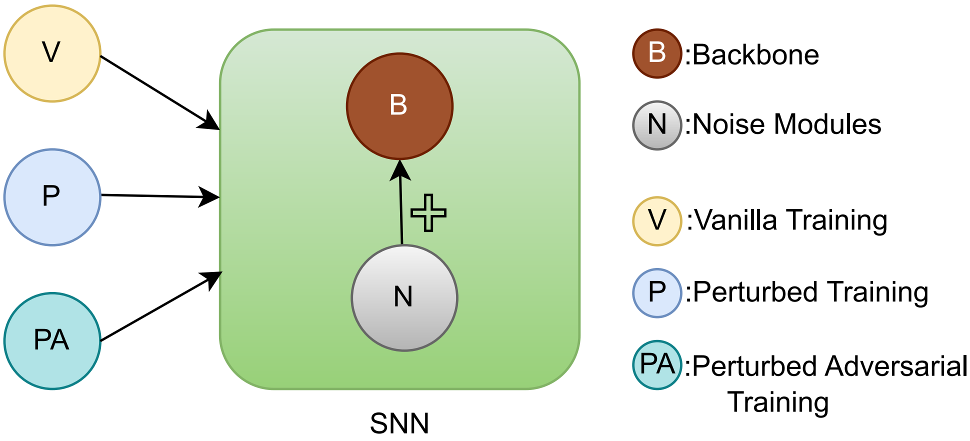


Figure 5.1: Meta Perturbed Defense: We propose a SNN by adding noise modules to a SOTA backbone. We train this SNN using a novel meta-learning strategy with tasks as 'V', 'P', and 'PA' to obtain our proposed framework MP-SNN.

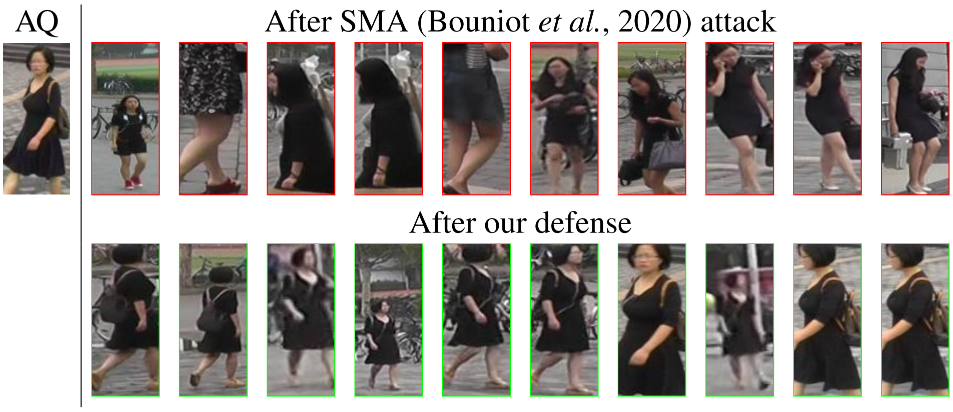


Figure 5.2: Top-10 predictions for Re-ID task on Market-1501 dataset after attack on ResNet-50 and after applying our defense. Green box - correct prediction; Red box - incorrect prediction. 'AQ' is attacked query image.

tasks. (Yang *et al.*, 2022) also proposed meta-learning based defense; however, it is computationally expensive as they use an additional dataset to capture variations in cross-domain. MP-SNN increases the adversarial robustness and helps to generalize better against unseen adversarial attacks. It provides a competitive performance in both clean and adversarial settings. We provide an overview of our proposed method in Figure 5.1.

Extensive experiments show that our method can be applied to architectures with varying complexity; ResNet-50 (He *et al.*, 2016) for person and vehicle Re-ID, and OSNet (Zhou *et al.*, 2019) for person Re-ID, and it shows SOTA performance across widely used person and vehicle Re-ID benchmarks. We show an illustration of our Meta Perturbed Defense in Figure 5.2.

Our contributions are as follows:

- We propose a robust meta perturbed stochastic neural network (MP-SNN) for defense against adversarial attacks in object Re-ID task. Our MP-SNN learns both anisotropic and isotropic noise distributions in a meta-learning framework.
- We propose a novel meta perturbed defense algorithm with tasks as vanilla, perturbed, and perturbed-adversarial training. MP-SNN increases the robustness and generalizability against wide variety of unseen attacks by adapting well to these tasks.
- We derive a novel feature covariance alignment loss which ensures high clean performance while providing robustness against wide variety of adversarial attacks.

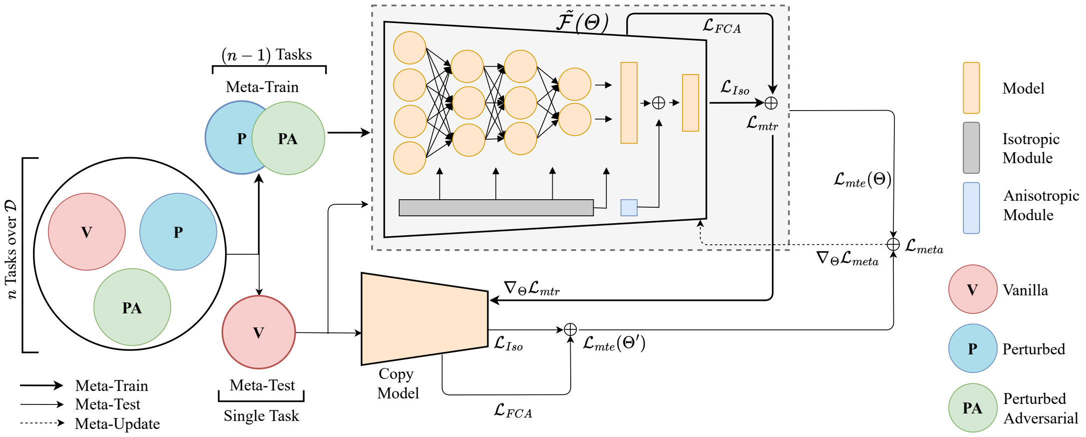


Figure 5.3: **Overview of MP-SNN.** Our SNN injected with anisotropic noise in the penultimate layer and isotropic noise in the inner layers is learnt via meta-learning. During training, we have n tasks over training data D, which we divide into $n-1$ meta-train tasks and a single meta-test task. In meta-train stage, L_{FCA} and L_{Iso} are calculated as meta-train loss $\mathcal{L}_{mtr}(\Theta)$. In meta-test stage, we copy the original model and update it using $\mathcal{L}_{mtr}(\Theta)$. We compute meta-test losses $\mathcal{L}_{mte}(\Theta)$ and $\mathcal{L}_{mte}(\Theta^{'})$ on original and updated model respectively. Finally, we update our original model with $L_{meta}(\Theta)$.

5.2 Methodology

In this section, we first describe our SNN. We then describe the training of SNN in our proposed meta perturbed defense algorithm. We show our proposed framework in Fig 5.3.

5.2.1 Proposed Stochastic Neural Network (SNN)

We first describe the model and the notations used. We then describe the Noise modules used in our network and training of our SNN.

Notations. The output of our SNN is drawn from a probabilistic rather than a deterministic distribution. We define our probabilistic model as $\tilde{\mathcal{F}}(x; \theta, \phi, \phi_a)$, where, x is the input, and θ, ϕ, ϕ_a are the model, isotropic noise, and anisotropic noise parameters, respectively. For simplicity we also represent our stochastic neural network as $\tilde{\mathcal{F}}(x; \Theta)$, where $\Theta = \{\theta, \phi, \phi_a\}$.

Vanilla Model. We use a standard deep neural network as our vanilla model. We use cross-entropy (Zhong *et al.*, 2018*b*) and triplet (Hermans *et al.*, 2017) loss to train the model with parameters θ. The vanilla loss is given as:

$$\mathcal{L}_V(\theta) = \mathcal{L}_{CE}(\theta) + \mathcal{L}_{Tri}(\theta) \tag{5.1}$$

Noise Modules.

We now discuss each component of noise separately in detail.

Isotropic Noise Injection. We inject isotropic noise (ϕ) into the inner layers of our model. It has a Gaussian distribution $\mathcal{N}(0, 1)$. The loss function for training our SNN on injecting isotropic noise is given as:

$$\mathcal{L}_{Iso}(\theta, \phi) = \mathcal{L}_{CE}(\theta, \phi) + \mathcal{L}_{Tri}(\theta, \phi) + \lambda r(\theta), \tag{5.2}$$

where $\lambda r(\theta)$ is the regularization term; $r(\theta) = -\theta^{1/2}/\tau$, where τ is determined using a harmonic series with input as the value of current epoch while training the network (Jeddi *et al.*, 2020).

Anisotropic Noise Injection. We add anisotropic Gaussian noise (ϕ_a) to the penultimate layer of our network which ensures comparative clean and adversarial performance (Eustratiadis *et al.*, 2021). We derive a new loss function called feature-covariance alignment (FCA) loss for optimizing the network with anisotropic perturbations similar to WCA-Net (Eustratiadis *et al.*, 2021). WCA-Net maximizes the loss by aligning the noise with weight vectors. However, it suffers from the fact that maximizing the loss inherently leads to large weight vectors which is clearly undesirable as large weights lead to overfitting. To address this, they add a regularizer that needs to be empirically tuned, which is difficult to tune for different datasets and models. Our loss function is more suited to Re-ID tasks where we utilize feature vectors of hard-batch triplets (Hermans *et al.*, 2017). It is given as:

$$\mathcal{L}_{FCA}(\theta, \phi_a) = \log\left(\frac{1}{B}\sum_{i=1}^{B}(f_p^i - f_n^i)^\top \Sigma (f_p^i - f_n^i)\right), \tag{5.3}$$

where B is the batch size, f_p and f_n are the feature vectors of positive and negative samples of the input image, and $\Sigma = \phi_a \phi_a^\top$ is the covariance matrix. Instead of aligning weight vectors with noise, we align feature vectors of the triplets associated with the final linear (penultimate) layer to Σ of the injected noise, avoiding the need of regularization. Thus, FCA loss is more computationally efficient than WCA-Net. FCA is maximized when features are well-aligned with eigen vectors of covariance matrix. FCA ensures that features and noise co-adapt and align efficiently, thereby giving us high clean and adversarial performance.

Derivation of Feature Covariance Alignment Loss (L_{FCA}). Let f be a feature extractor parameterized by θ. We denote an anchor or query sample by f_a, positive sample by f_p, and negative by f_n. Then, to obtain a correct prediction, we need,

$$m = f_a^\top (f_p - f_n) > 0$$

where m is referred to as margin. In presence of noise, we can represent the margin as,

$$\delta_m = (f_a + \delta)^\top [f_p - f_n]$$

Let the mean of m is given by,

$$\mu = \mathbb{E}[f_a^\top (f_p - f_n)]$$

Then, mean of δ_m is,

$$\mathbb{E}[f_a^\top (f_p - f_n)] + \underbrace{E(\delta^\top f_p) - E(\delta^\top f_n)}_{0} = \mu$$

The second term is 0 as we assume δ to be zero mean noise.

Similarly, we can compute the variance as,

$$\sigma_m^2 = \mathbb{E}[f_a^\top (f_p - f_n)(f_p - f_n)^\top f_a] - \mu^2 \tag{5.4}$$

and,

$$\begin{aligned}\sigma_\delta^2 &= \mathbb{E}[(f_a + \delta)^\top (f_p - f_n)] - \mu^2 \\ &= \mathbb{E}[(f_p - f_n)^\top (f_a f a^\top + \delta\delta^\top + f_a \delta^\top + \delta f_a^\top)\end{aligned} \tag{5.5}$$

From equations 5.4 and 5.5, we can see that the difference comes from $\mathbb{E}[(f_p - f_n)^\top \delta\delta^\top (f_p - f_n)]$ term. Thus, the difference can be represented as,

$$\sigma^2 = E[(f_p - f_n)^\top \Sigma (f_p - f_n)],$$

where $\Sigma = \delta\delta^\top$.

Therefore, our proposed feature-covariance alignment loss (FCA) loss is given as:

$$\mathcal{L}_{FCA}(\theta, \phi_a) = \log\left(\frac{1}{B}\sum_{i=1}^{B}(f_p^i - f_n^i)^\top \Sigma (f_p^i - f_n^i)\right),$$

where $\Sigma = \phi_a \phi_a^\top$ and B is the batch size.

Mixed Noise Injection. Isotropic noise generation has a significant shortcoming in that the generated noise has to be axis-aligned; that is, noise and weights of a given feature space need to align towards the same axis. This affects the SNN's ability to learn distributions that are not axis-aligned, thus affecting clean performance Eustratiadis *et al.* (2021). Anisotropic noise overcomes this limitation. However, this comes at the cost of a decrease in its performance against adversarial attacks. Hence, we propose a SNN that finds a balance between the respective limitations by injecting isotropic perturbations into all the intermediary feature space and anisotropic perturbation into the final feature space. The loss on injecting this mixed noise is:

$$\mathcal{L}_{Mix}(\theta, \phi, \phi_a) = \mathcal{L}_{Iso}(\theta, \phi) - \mathcal{L}_{FCA}(\theta, \phi_a). \tag{5.6}$$

We also perform adversarial training along with mixed noise perturbation. The total loss function of our proposed stochastic neural network is given as:

$$\mathcal{L}_{Tot}(\theta, \phi, \phi_a) = \mathcal{L}_{Mix}(\theta, \phi, \phi_a) + \eta \mathcal{L}_{Adv}(\theta, \phi, \phi_a), \tag{5.7}$$

where η is the adversarial loss weight, and $\mathcal{L}_{Adv}$ is the addition of triplet and cross-entropy loss. $\mathcal{L}_{Adv}$ is calculated on adversarial samples generated from attacking input data using untargeted l_∞ PGD attack (Madry *et al.*, 2017).

Training of our SNN.

We learn θ and $\{\phi, \phi_a\}$ alternately using a back-propagation strategy (Han *et al.*, 2017). In our proposed approach, we first inject anisotropic noise in the penultimate layer and update $\{\theta, \phi_a\}$ while keeping ϕ fixed. Then, we inject isotropic noise in the inner layers and update ϕ while keeping $\{\theta, \phi_a\}$ fixed. We describe the training of SNN in Algorithm 2.

ALGORITHM 2: Training of our SNN $\tilde{\mathcal{F}}(\Theta)$

```
Input: Training data $\mathcal{D}$, learning rates $\alpha, \beta$
Output: Parameters $\theta$, $\phi$ and $\phi_a$
while not done do
    $(\theta, \phi_a)$ are updated based on Eq. 5.7 while $\phi$ remains fixed:
    $(\theta, \phi_a) \leftarrow \alpha \nabla_{(\theta, \phi_a)} \mathcal{L}_{Tot}(\theta, \phi, \phi_a)$
    $\phi$ is updated based on Eq. 5.7 while $\theta$ and $\phi_a$ remain fixed:
    $\phi \leftarrow \beta \nabla_{\phi} \mathcal{L}_{Tot}(\theta, \phi, \phi_a)$
end while
```

5.2.2 Meta Training

We train our SNN using our novel meta-learning strategy for generalization and robustness against adversarial attacks by adapting well to noise modules and adversarial training. We propose a novel way to define tasks for meta-learning. These tasks $\mathbb{T}$ are: **Vanilla training (V):** Training on clean images using loss given in Eq. 5.1, **Perturbed training (P):** We add anisotropic and isotropic noise modules to model layers (Eq. 5.6), and **Perturbed-adversarial training (PA):** We train our model by adding noise modules and performing additional adversarial training (Eq. 5.7). We choose a random task as meta-test, and the remaining tasks as meta-train. Now, we discuss the three steps of our meta training in detail:

Step 1: Meta-Train. We select k samples from the training data D with total N samples in a mini-batch, for each task in meta-train for a given set of tasks $\mathbb{T}$. The meta-train loss $\mathcal{L}_{mtr}(\Theta)$ for our network $\tilde{\mathcal{F}}(x;\Theta)$ computed over all the meta-train tasks is formulated as,

$$\mathcal{L}_{mtr}(\Theta) = \sum_{j=1}^{k} \sum_{i=1}^{\mathbb{T}\backslash t} \mathcal{L}_{Tot}(\tilde{\mathcal{F}}(x_j;\Theta_i), y), \tag{5.8}$$

where y is the true label and L_{Tot} is as given in Eq. 5.7.

Step 2: Meta-Test. During meta-testing, we copy the original model $\tilde{\mathcal{F}}(x;\Theta)$ and update it with meta-train loss $\mathcal{L}_{mtr}$ to obtain $\tilde{\mathcal{F}}(x;\Theta^{'})$:$\Theta^{'} \leftarrow \nabla_{\Theta}(L_{mtr})$. We compute meta-test loss $\mathcal{L}_{mte}$ on meta-test set of $N-k$ samples over task t on both original and updated model. It is given as:

$$\mathcal{L}_{mte}(\Theta) = \sum_{j=1}^{N-k} \mathcal{L}_{Tot}(\tilde{\mathcal{F}}(x_j;\Theta), y) \tag{5.9}$$

$$\mathcal{L}_{mte}(\Theta^{'}) = \sum_{j=1}^{N-k} \mathcal{L}_{Tot}(\tilde{\mathcal{F}}(x_j;\Theta^{'}), y) \tag{5.10}$$

Step 3: Meta-Update. In the final step, our original model is updated using $\mathcal{L}_{meta}(\Theta)$ which is:

$$\mathcal{L}_{meta}(\Theta) = \mathcal{L}_{mte}(\Theta) + \mathcal{L}_{mte}(\Theta^{'}) \tag{5.11}$$

The meta training of our MP-SNN is summarized in Algorithm 3.

ALGORITHM 3: Meta Training

```
Input: Tasks $\mathbb{T}$, Training data $D$, Stochastic neural network (SNN) $\tilde{\mathcal{F}}(\Theta)$;

while not done do
    Randomly select a task $t$ from $\mathbb{T}$ as meta-test task
    Sample remaining tasks as meta-train tasks
    Meta-Train: (i) Sample $k$ non-overlapping images from $D$ as meta-train set for
    each training task
    (ii) Compute meta-train loss $\mathcal{L}_{mtr}(\Theta)$ (Eq. 5.8)
    Meta-Test: (i) Sample $N-k$ images from $D$ (exclusive of meta-train tasks' $k$
    images) as meta-test set
    (ii) Compute meta-test loss $\mathcal{L}_{mte}$ (Eq. 5.9 and 5.10)
    Meta-Update: (i) Compute $\mathcal{L}_{meta}(\Theta)$ (Eq. 5.11)
    (ii) Update $\Theta \leftarrow \nabla_{\Theta} L_{meta}(\Theta)$
end while
```

5.3 Experimental Settings

Datasets. We evaluate our method on three large-scale person Re-ID datasets Market-1501, DukeMTMC-ReID, MSMT17, and a vehicle Re-ID dataset Veri-776. [1].

VeRi-776 (Liu *et al.*, 2016). It is comprised of 49,357 images depicting 776 distinct vehicles captured by 20 different cameras. This dataset was collected under real-world traffic conditions, closely mimicking the environment of CityFlow. The dataset includes annotations for bounding boxes, vehicle types, colors, and brands.

Implementation Details. We set the hyperparameters as follows: the batch size is 16, and learning rates α and β are $1e-5$ and $3e-4$ respectively. We use the SGD optimizer with a momentum of 0.9 and a weight decay of $1e-5$. The hyperparameter λ used in Eq. 5.2 is set to $1e-4$. We perform PGD adversarial training and robustness test against attacks with 7 iterations and maximum $\epsilon = 8/255$, unless specified otherwise. Adversarial weight η is 0.5. For fair comparison with SOTA methods, we use 15 iterations with $\epsilon = 5/255$. We use Cumulative Matching Characteristic (CMC) precision and Mean Average Precision (mAP) metrics (Zheng *et al.*, 2015) to evaluate our work.

5.3.1 Performance against adversarial attacks

We evaluate the performance of our MP-SNN framework against wide variety of SOTA white-box attacks FGSM, PGD, MI-FGSM, NI-FGSM and SI-NI-FGSM, EOT-FGSM and EOT-PGD, and SOTA Re-ID specific attacks SMA (Bouniot *et al.*, 2020) and AMA (Bai *et al.*, 2020*b*) in Table 5.1. We show in Tables 5.1 and 5.2 that our MP-SNN is effective in improving the robustness of different network architectures trained on different datasets - Market-1501, DukeMTMC-ReID, MSMT17 and Veri-776. We report the performance of our proposed defense technique (MP-SNN) on two SOTA vanilla models, ResNet-50 (He *et al.*, 2016) and OSNet (Zhou *et al.*, 2019). MP-SNN (R-50) represents our defense mechanism MP-SNN is trained on 'ResNet-50', and MP-SNN (OS) represents that MP-SNN is trained on 'OSNet'. For the Market-1501 dataset, MP-SNN provides a boost of around **48%**, **55%** over SOTA ResNet-50 and **28%**, **20%** over OSNet under powerful EOT-PGD and Re-ID specific SMA attacks, respectively. Similar observations can be made on MSMT17 and VeRi-776 datasets as well for ResNet-50 and OSNet architectures.

[1] The details of Market-1501, DukeMTMC-ReID and MSMT17 datasets are present in Chapter 3

Datasets	Models	No Attack		FGSM		PGD		SMA		AMA	
		CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP
Market-1501	R-50	89.96	73.65	58.55	39.85	18.94	14.15	22.68	17.12	0.02	0.08
	MP-SNN (R-50)	**86.82**	**66.56**	**71.22**	**51.93**	**74.55**	**56.01**	**77.58**	**57.81**	**48.42**	**34.67**
DukeMTMC-reID	R-50	79.80	63.96	45.91	31.65	20.10	15.77	23.11	18.17	-	-
	MP-SNN (R-50)	**53.81**	**31.40**	**38.24**	**21.46**	**36.57**	**19.95**	**44.25**	**24.61**	-	-
MSMT17	R-50	52.30	33.62	5.08	3.10	0.26	0.20	0.59	0.40	-	-
	MP-SNN (R-50)	**37.34**	**22.72**	**22.02**	**13.11**	**15.91**	**9.54**	**25.79**	**15.34**	-	-
VeRi-776	R-50	95.17	75.40	69.78	49.85	59.53	45.66	64.00	49.34	-	-
	MP-SNN (R-50)	**89.21**	**64.92**	**79.61**	**54.74**	**79.32**	**55.22**	**85.57**	**60.57**	-	-

Datasets	Models	MI-FGSM		NI-FGSM		SI-NI-FGSM		EOT-FGSM		EOT-PGD	
		CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP
Market-1501	R-50	17.54	13.64	28.23	19.34	27.22	18.50	58.55	39.85	15.08	11.78
	MP-SNN (R-50)	**70.48**	**51.92**	**69.53**	**50.79**	**68.05**	**49.61**	**70.27**	**51.72**	**63.36**	**45.26**
DukeMTMC-reID	R-50	18.49	14.99	30.65	22.10	30.43	20.87	45.91	31.65	16.20	13.25
	MP-SNN (R-50)	**36.44**	**20.42**	**35.99**	**20.14**	**36.40**	**19.78**	**38.28**	**21.20**	**30.25**	**16.41**
MSMT17	R-50	0.31	0.25	0.65	0.47	0.75	0.45	5.08	3.10	0.20	0.17
	MP-SNN (R-50)	**19.09**	**11.23**	**18.59**	**10.68**	**16.30**	**9.43**	**22.71**	**13.55**	**13.71**	**8.30**
VeRi-776	R-50	61.79	46.84	61.85	45.56	60.78	44.44	69.78	49.85	55.42	43.20
	MP-SNN (R-50)	**76.40**	**53.27**	**77.23**	**54.21**	**78.42**	**54.22**	**78.54**	**54.74**	**73.77**	**50.98**

Table 5.1: Performance of Re-ID models against adversarial attacks after applying our defense method. ‘R-50’ is the baseline model of ResNet-50. ‘MP-SNN (R-50)’ is our proposed approach.

Datasets	Models	No Attack		FGSM		PGD		SMA		AMA	
		CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP
Market-1501	OS	94.23	83.60	59.08	46.73	45.48	36.13	54.45	43.74	1.51	0.87
	MP-SNN (OS)	**89.81**	**74.38**	**75.83**	**58.59**	**72.44**	**56.47**	**81.20**	**64.67**	**53.68**	**31.24**
MSMT17	OS	58.98	40.54	11.80	7.61	1.10	0.82	2.05	1.31	-	-
	MP-SNN (OS)	**47.79**	**31.24**	**27.83**	**17.40**	**16.87**	**10.65**	**28.23**	**18.00**	-	-

Datasets	Models	MI-FGSM		NI-FGSM		SI-NI-FGSM		EOT-FGSM		EOT-PGD	
		CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP
Market-1501	OS	50.80	40.21	56.82	44.14	58.28	45.51	59.08	46.73	42.42	33.13
	MP-SNN (OS)	**78.97**	**61.84**	**74.31**	**58.25**	**75.95**	**59.56**	**75.79**	**58.71**	**68.31**	**53.08**
MSMT17	OS	1.33	0.92	2.02	1.36	2.08	1.38	11.80	7.61	0.92	0.65
	MP-SNN (OS)	**23.30**	**14.48**	**20.49**	**12.79**	**18.92**	**11.76**	**27.42**	**17.13**	**15.68**	**9.72**

Table 5.2: Performance of Re-ID models against adversarial attacks after applying our defense method. ‘OS’ is the baseline model of OSNet. ‘MP-SNN (OS)’ is our proposed approach MP-SNN with OSNet as backbone.

Dataset	Model	Clean		WMA	
		CMC-1	mAP	CMC-1	mAP
Market-1501	R-50	89.96	73.65	13.25	10.55
	MP-SNN	86.82	66.56	**60.64**	**44.21**
	OS	94.23	83.60	42.07	30.25
	MP-SNN	89.81	74.38	**64.32**	**52.71**
MSMT17	R-50	52.30	33.62	0.13	0.12
	MP-SNN	37.34	22.72	**12.45**	**6.43**
	OS	58.98	40.54	0.67	0.46
	MP-SNN	47.79	31.24	**12.89**	**6.28**
VeRi-776	R-50	95.17	75.40	54.26	40.37
	MP-SNN	89.21	64.92	**72.47**	**47.57**

Table 5.3: Performance against our proposed Wasserstein Metric Attack (WMA). 'R-50' is SOTA ResNet-50 He *et al.* (2016) and 'OS' is OSNet Zhou *et al.* (2019). ($\epsilon = 8/255$).

Performance against Wasserstein Metric Attack (WMA) We evaluate our meta-perturbed defense on Wasserstein metric attack (WMA) in Table 5.3. We can observe from the results that our WMA attack drastically decreases the performance of ResNet-50 (R-50) and OSNet (OS) trained on three large-scale datasets Market-1501, MSMT17 and Vei-776. For Market-1501 dataset, the clean CMC-1 for R-50 is 89.96% which reduces to 13.25% after WMA attack. However, our proposed defense MPSNN increases the robustness of 'R-50' with a considerable **47.39%** increase in the CMC-1 metric. Similarly, when utilizing the 'OS' backbone architecture with clean CMC-1 as 94.23%, the performance reduces to 42.07% with the standard backbone, whereas after our proposed defense MPSNN, the CMC-1 increases to **64.32%**.

Generalization to Unseen Attacks. We show in Table 5.4 that the performance is still very low after doing PGD adversarial training of ResNet-50 represented as PGD-AT. Thus, vanilla adversarial training fails to provide generalization. Under MI-FGSM attack, PGD-AT gives a boost of only 7.1%, whereas MP-SNN gives a large boost of **52.94%**. Similarly under AMA attack, PGD-AT has poor performance with CMC-1 of 0.02%, whereas MP-SNN gives a high CMC-1 of **48.42%**. It is clear that MP-SNN generalizes well to various unknown white-box attacks as well as to Re-ID specific attacks while only being trained by adversarial samples generated from PGD attack. Thus, our proposed method does not require training against each attack and can be robust against various unknown attacks.

Model	Clean	FGSM	PGD	MI-FGSM	NI-FGSM	SI-NI-FGSM	EOT-FGSM	EOT-PGD	SMA	AMA
R-50	89.96	58.55	18.94	17.54	28.23	27.22	58.55	15.08	22.68	0.02
PGD-AT	86.60	62.11	27.49	24.64	36.72	37.82	62.11	24.16	28.02	0.05
MP-SNN	86.82	**71.22**	**74.19**	**70.48**	**69.53**	**68.05**	**70.27**	**63.36**	**77.58**	**48.42**

Table 5.4: Generalization of MP-SNN against various attacks. PGD-AT is adversarial training using adversarial samples generated using PGD attack. ‘R-50’ is the baseline SOTA ResNet-50 (He *et al.*, 2016). CMC-1 in %. Dataset is Market-1501 ($\epsilon = 8/255$).

Backbone	Defense	Clean	MI-FGSM	SMA
ResNet-50 (He *et al.*, 2016)	Baseline	89.96	17.54	22.68
	AMD (Bai *et al.*, 2020*b*)	84.67	21.49	25.41
	AMD*	84.76	20.72	24.40
	GOAT (Bouniot *et al.*, 2020)	86.65	26.33	33.45
	GOAT*	87.46	28.41	34.21
	JAD (Gong *et al.*, 2022)	88.65	27.04	30.55
	JAD*	90.46	30.25	34.53
	Meta-Def (Yang *et al.*, 2022)	88.90	53.34	62.40
	Meta-Def*	89.23	56.81	64.71
	MP-SNN	86.82	**69.09**	**75.44**
	MP-SNN*	87.36	**74.55**	**79.29**
OSNet (Zhou *et al.*, 2019)	Baseline	94.23	50.80	54.45
	AMD	91.77	47.20	47.86
	AMD*	92.36	44.12	44.47
	MP-SNN	89.81	**77.01**	**80.31**
	MP-SNN*	89.75	**80.26**	**81.99**

Table 5.5: Comparison with SOTA on Market-1501 dataset. CMC-1 in %. * denotes re-ranking. $\epsilon = 5/255$.

5.3.2 Comparison with State-of-the-art

We compare MP-SNN with SOTA Re-ID defense methods AMD (Bai *et al.*, 2020*b*), GOAT (Bouniot *et al.*, 2020), JAD (Gong *et al.*, 2022) and Meta-Def (Yang *et al.*, 2022) by implementing them in our setting in Table 5.5. We also provide results using re-ranking (Zhong *et al.*, 2017) during testing. Performance of AMD and GOAT after SMA attack reduces to 24.40% and 34.53% respectively, whereas MP-SNN achieves a high CMC-1 of **77.58%**. AMD and GOAT use vanilla adversarial training to provide defense which gives a poor performance against adversarial attacks, thereby showing its ineffectiveness. Similarly, JAD, which uses data-augmentation techniques, also performs poorly. Our method beats the SOTA Meta-Def by **17.74%** and **15.87%** under both MI-FGSM and SMA attacks respectively. Moreover, Meta-Def is a computationally expensive defense as it uses an additional virtual dataset to create a large training data.

We also compare MP-SNN with AMD with OSNet as backbone in Table 5.5. AMD uses adversarial training which gives a low performance for all the settings, as is visible from the results. These results show the effectiveness of our approach and its SOTA performance.

5.3.3 Ablation Study

Effect of FCA. In Table 5.6, we show the effect of our feature covariance alignment (FCA) in our MP-SNN framework. We observe that FCA gives high clean performance while maintaining its adversarial accuracy. Thus, FCA maximizes the alignment between noise and feature vectors without compromising its clean performance.

Method	Clean	PGD	SMA
MP-SNN w/o FCA	84.20	73.45	75.77
MP-SNN w/ FCA	**86.82**	**74.19**	**77.58**

Table 5.6: Effect of FCA on MP-SNN. $\epsilon = 8/255$

Loss		Clean	PGD	SMA
w/o ML	L_V	89.96	18.94	22.68
	L_{Adv}	86.60	27.49	28.02
	L_{Tot}	82.69	67.66	74.61
w/ ML	L_{Meta}	86.82	**74.19**	**77.58**

Table 5.7: Effect of loss functions and meta-training on our proposed MP-SNN. 'ML' is our proposed meta-learning strategy. Dataset is Market-1501. $\epsilon = 8/255$.

Effect of Losses and Meta-learning. We show in Table 5.7 that L_V (Vanilla Training) and L_{Adv} (Vanilla Adversarial Training) do not exhibit robustness against attacks. This is due to the fact that Re-ID datasets are complex, and only adversarial training fails to impart enough adversarial robustness. Under SMA attack, L_{Adv} gives a very low performance of 27.49%. We further show that L_{Tot} (Perturbed-Adversarial Training) increases the performance to **74.61%** under SMA attack. It shows the effectiveness of our noise modules in providing robustness against SOTA attacks. After injecting our anisotropic and isotropic noise modules, our network is able to learn better from adversarial examples. Our proposed MP-SNN with novel meta-Learning strategy (L_{Meta}) gives us a high performance of **74.19%** and **77.58%** under PGD and SMA attacks, respectively. We conclude from these experiments that our proposed MP-SNN tremendously increases the model robustness.

5.3.4 Challenges

We faced two main challenges in our meta-defense work.

Minor increase in training time: In our meta-learning-based defense, there is a small increase in training time compared to traditional single-task learning approaches. Our meta-defense typically had additional computations and iterations during meta-training, where our model learns from tasks defined as, (i) **Vanilla training (V):** Training on clean images, (ii) **Perturbed training (P):** We add anisotropic and isotropic noise modules to model layers, and (iii) **Perturbed adversarial training (PA):**, We train our model by adding noise modules and performing additional adversarial training. These tasks help to improve our model's ability to provide robustness against unknown adversarial attacks. While this can lead to a slight increase in training time compared to single-task learning, our technique and optimization algorithm aims to minimize this increase by efficiently leveraging the shared knowledge of tasks comprising of clean samples, noise parameters, and adversarial samples and reducing redundant computations. Despite the small increase in training time, the benefits gained from the our meta-defense process, such as improved adaptability and generalization to different attacks, often outweigh the additional time investment.

Defining meta-learning tasks suitable for our defense: In our meta perturbed defense, selecting relevant and suitable tasks during meta-training was crucial for the success of the defense system. Defining appropriate tasks for our meta-defense involved carefully curating a diverse set of attack-related scenarios that the model needs to be robust against. We define these tasks as: Vanilla training (V), Perturbed training (P), and Perturbed adversarial training (PA). These tasks are representative of potential threats, attack scenarios, or vulnerabilities that our defense system aims to address. Proper task selection ensures that our meta-learning model learns transferable knowledge from introduced noise parameters, clean and adversarial samples for robustness against new attack scenarios.

Despite the challenges mentioned, the application of meta-learning has resulted in optimized performance for our defense system. With meta-learning, we excel at learning from our defined defense-related tasks (V), (P), and (PA) and leveraging this knowledge to adapt to new adversarial attacks efficiently. By training on a variety of defense-related tasks, the meta-learning model acquires a more comprehensive understanding of defense strategies, which leads to improved performance and robustness. The novel task-definition (including noise parameters, clean and adversarial samples) for robust performance is a key advantage of our meta-learning-based defense over state-of-the art defense approaches.

In summary, there might be a slight increase in training time in our meta-learning-based defense due to the need for meta-training, the benefits of optimized performance and improved adaptability outweigh this challenge. Careful task selection remains essential to ensure that our meta-perturbed defense model learns relevant and transferable knowledge for the defense system to achieve its objectives effectively.

5.4 Chapter Summary

In this chapter, we propose a meta perturbed framework to defend Re-ID models against adversarial attacks. We leverage the adversarial robustness characteristics of anisotropic and isotropic noise modules along with adversarial training and efficiently learn them through our proposed meta-learning defense strategy. Our model generalizes well and is robust against various unseen attacks. Further, we derive a novel FCA loss related to anisotropic perturbations. FCA overcomes the decrease in clean performance while ensuring robust performance against adversarial attacks. Elaborate experiments show that MP-SNN provides SOTA adversarial defense for Re-ID task.

While Chapter 4 and this chapter introduced robustness techniques that rely on empirical (adversarial training) methods, these methods may not always provide a certification of defense (Cohen *et al.*, 2019) because they are based on data and observation rather than theoretical models. Although empirical methods can be useful for testing the effectiveness of defense mechanisms, they may not be sufficient to guarantee defense with high confidence, since they cannot account for all possible attack scenarios or ensure the system's continued effectiveness over time. In the next chapter we will focus on exploring certified defense in the black-box setting.

CHAPTER 6

Certified Zeroth-order Black-Box Defense

6.1 Introduction

A notable amount of success has been attained by machine learning (ML) models (Joshi *et al.*, 2022; Katz *et al.*, 2022), and deep neural networks (DNNs) in particular because of their better predictive capabilities. However, their lack of robustness and susceptibility to adversarial perturbations has caused serious worries about their wide-scale adaptation in a number of artificial intelligence (AI) applications (Goodfellow *et al.*, 2014*b*; Carlini and Wagner, 2017; Papernot *et al.*, 2016*a*; Brown *et al.*, 2017; Eykholt *et al.*, 2018; Antun *et al.*, 2020). These adversarial attacks have motivated various strategies to strengthen ML models as a key area of research (Madry *et al.*, 2017; Athalye *et al.*, 2018*a*; Zhang *et al.*, 2019*b*; Cui *et al.*, 2021). Among these techniques, adversarial training (AT) (Szegedy *et al.*, 2013; Madry *et al.*, 2017) is one of the prominent defense strategies. The advancements in AT led to various empirical defense methods (Athalye *et al.*, 2018*b*; Wang and Wang, 2022; Yan *et al.*, 2022; Chan *et al.*, 2019; Zhang *et al.*, 2019*b*), however, these methods may not always be certifiably robust (Uesato *et al.*, 2018; Croce and Hein, 2020). Another line of research is certified defense, where an off-the-shelf model's prediction is certified within the neighborhood of the input. These methods are called certified defense techniques (Wong and Kolter, 2018; Raghunathan *et al.*, 2018; Katz *et al.*, 2017; Salman *et al.*, 2019, 2020, 2022).

(Cohen *et al.*, 2019) first proposed randomized smoothing (RS), which certifies defense by forming a smoothed model from the empirical model by adding Gaussian noise to the input images. Few other works have been proposed which provide certified defense inspired by randomized smoothing (Salman *et al.*, 2020, 2022; Addepalli *et al.*, 2021). (Salman *et al.*, 2020) pre-pended a custom-trained denoiser to the predictor for increased robustness. In another work, (Salman *et al.*, 2022) apply visual transformers within the smoothing network in order to provide certified robustness to adversarial patches. While previous works in adversarial defense have achieved promising advancements, robustness is provided over white-box models with known architectures and parameters. The white-box assumption, however, has high computational complexity as models are trained end-to-end as in AT, thus limiting the practicality and scalability of the defense method. For instance, it becomes impractical to retrain complex ML models trained on a vast number of MRIs or CT scans (Sinha *et al.*, 2022; Hussain *et al.*, 2023).

Moreover, privacy concerns may arise when implementing white-box defense since the owner may not wish to reveal model information. This is because attacks such as membership inference and model inversion attacks expose the vulnerabilities of the training data (Fredrikson *et al.*, 2015). Due to the scalability and privacy issues, few previous works tackled the highly non-trivial problem of adversarial defense in the black-box setting ('black-box defense') (Salman *et al.*, 2020; Zhang *et al.*, 2022*d*).

(Salman *et al.*, 2020) used surrogate models as approximates of the black-box models, over which defense may be done using the white-box setup. However, this setup requires information on the target model type and its function, which may be not be available practically. In another recent work, (Zhang *et al.*, 2022*d*) proposed a more authentic black-box defense of DNN models with the help of the zeroth-order optimization perspective. They pre-pended a custom-trained denoiser as in (Salman *et al.*, 2020) followed by an autoencoder architecture to the target model and trained it with ZO optimization. However, (Zhang *et al.*, 2022*d*) fail to perform for high dimensional datasets like Tiny imagenet with the dimensionality of $299 \times 299 \times 3$ as the denoiser used may fail to preserve spatial information. The main contribution of (Zhang *et al.*, 2022*d*), that is, the addition of autoencoder to existing technique (Salman *et al.*, 2020), enhances the robustness of black-box model to some extent for low-dimensional datasets like CIFAR-10 ($32 \times 32 \times 3$). However, this limits its usage to only the coordinate-wise gradient estimation (CGE) ZO optimization technique. However, our proposed approach can utilize two main existing ZO optimization techniques: randomized gradient estimate (RGE) and CGE. We discuss in detail in Section 6.4.2, where we prove the limitations of (Zhang *et al.*, 2022*d*) on high dimensional datasets experimentally in Figure 6.4.

In this chapter, we propose a certified black-box defense ZO-RUDS using the ZO optimization technique, where we pre-pend a novel robust UNet denoiser (RDUNet) to the target model. We further pre-pend our RDUNet and custom-trained autoencoder and propose the ZO-AE-RUDS defense mechanism. Our proposed methods require only input queries and output feedback and provide defense in a pure black-box setting. Unlike

SOTA (Zhang *et al.*, 2022*d*), which leads to high model variance on direct application of ZO optimization on the custom-trained denoiser, our proposed RDUNet, due to its architectural advantage over previous denoisers decreases model variance and provides better performance with direct application of ZO optimization. We experiment with various denoisers and prove that our RDUNet denoiser provides improved performance for both low-dimensional and high-dimensional datasets. Since we are dealing with a difficult ZO optimization and cannot back-propagate through the model, we optimize our proposed model by utilizing the black-box model's predicted labels and softmax probabilities. In order to further increase the certification of our proposed approach, we utilize maximum mean discrepancy (MMD) to bring the distributions of original input images closer to obtained denoised output.

We provide an illustration of certified defense techniques and compare them to our approach in Figure 6.1. We compare our proposed approaches, ZO-RUDS and ZO-AE-RUDS, with SOTA certified defense techniques in white-box (W) and black-box (B) settings. The input to the defense framework is sample x and noise η. RS (Cohen *et al.*, 2019) and DS (Salman *et al.*, 2020) provide certified defense in the white-box setting. In the 'RS' technique, noisy images ($x + \eta$) are input to the white-box model for certified robustness. 'DS' uses an additional custom-trained denoiser and pre-pends it to the predictor for certified robustness in the white-box setting. In addition to defense in the white-box setting, 'DS' proposed certified black-box robustness using a surrogate model.

In order to provide black-box defense without the use of a surrogate model (as it uses the target model as its proxy and it is not always possible to have access to the information on the target model and its function), (Zhang *et al.*, 2022*d*) proposed black-box defense with zeroth-order (ZO) optimization. They proposed ZO-DS by direct application of ZO on 'DS' and further append an autoencoder in the ZO-AE-DS technique. However, their proposed techniques have low performance on high-dimension datasets as the custom-trained denoiser fails to learn fine-scale information leading to poor performance. In order to overcome these limitations, we propose a robust UNet denoiser RDUNet inspired from the conventional UNet used for image segmentation (Ronneberger *et al.*, 2015). Our robust denoiser RDUNet with the up-sampling and downsampling layers, and lateral skip connections enables our defense to learn complex structures and fine-scale information and makes it invariant to changes

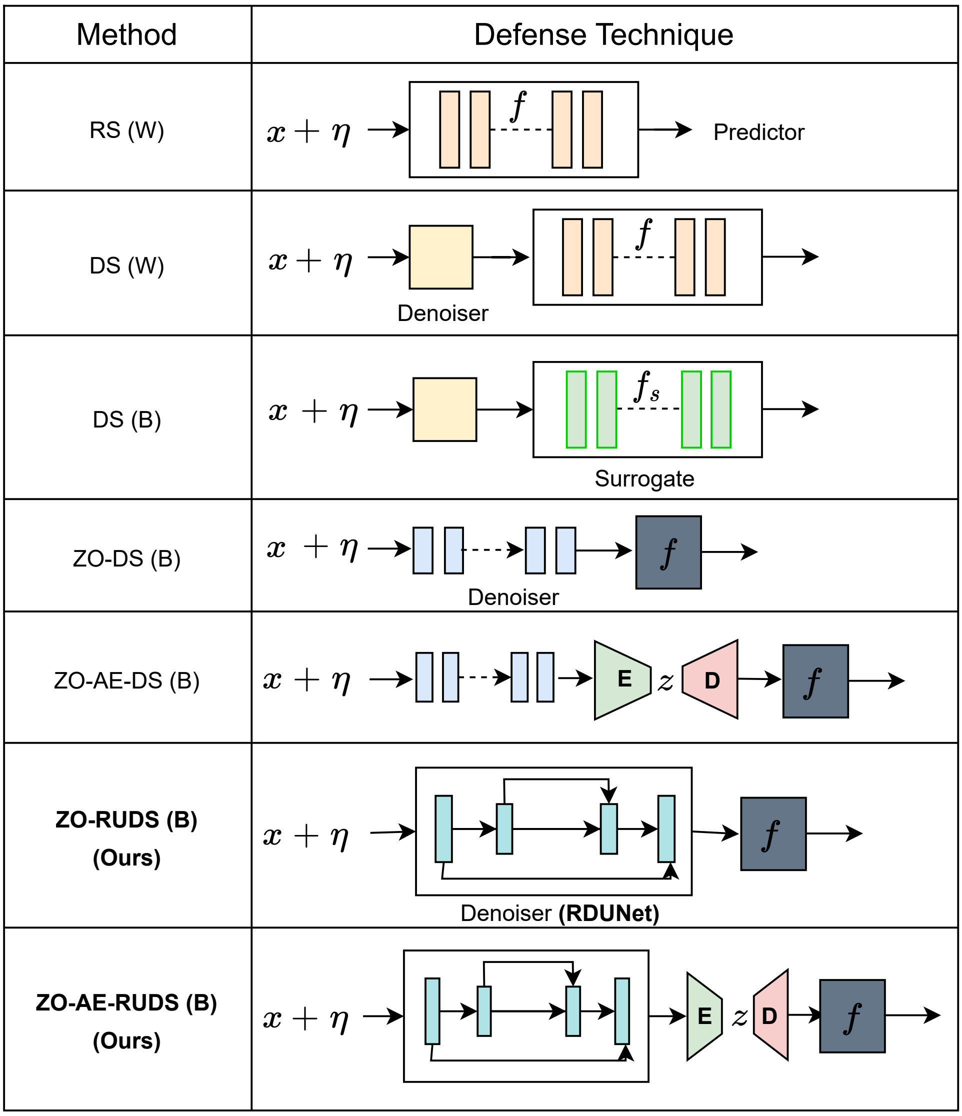


Figure 6.1: We make a comparison with four previous certified defense methods, including RS (Cohen *et al.*, 2019), DS(W) (Salman *et al.*, 2020), DS(B) (Salman *et al.*, 2020), ZO-DS (Zhang *et al.*, 2022*d*) and ZO-AE-DS (Zhang *et al.*, 2022*d*) (ZO-optimization approaches). 'W' and 'B' refer to white-box (defense technique can utilize weights of target model f) and black-box settings. 'x' - input sample, 'η' - noise, 'E' - Encoder, 'D' - Decoder, 'f' - target model, f_s - surrogate model (proxy of f) and 'z' - latent feature vector.

in image dimensions, thus giving a high performance for high-dimension images.

We summarize our contributions as follows:

- We propose a certified black-box defense mechanism based on the preprocessing technique of pre-pending a robust denoiser to the predictor to remove adversarial noise using only the input queries and the feedback obtained from the model.
- We design a novel robust UNet denoiser RDUNet which defends a black-box model with ZO optimization approaches. Unlike previous ZO optimization-based defense approaches, which give a poor performance on high-dimensional data due to high model variance, our UNet-based robustification model gives high performance for both low-dimensional and high-dimensional datasets.
- We conduct extensive experiments and show that our proposed defense mechanism beats SOTA by a huge margin on four classification datasets, CIFAR-10, CIFAR-100, STL-10, Tiny Imagenet, and on MNIST dataset for reconstruction task.

6.1.1 ZO Optimization for adversarial learning.

ZO optimization is useful in solving black-box problems where gradients are difficult to compute or infeasible to obtain (Wei *et al.*, 2022; Yin *et al.*, 2023). These methods are gradient-free counterparts of first-order (FO) optimization methods (Liu *et al.*, 2020). Recently, ZO optimization has been used for generating adversarial perturbations in black-box setting (Chen *et al.*, 2017; Ilyas *et al.*, 2018*b*,*a*; Tu *et al.*, 2019; Liu *et al.*, 2019*c*, 2020a; Huang and Zhang, 2020; Cai *et al.*, 2021, 2022). Similar to attack methods, ZO optimization can also be applied to black-box defense methods with access only to the inputs and outputs of the targeted model. (Zhang *et al.*, 2022*d*) proposed black-box defense using ZO optimization and leveraged autoencoder architecture for optimizing the defense approach with CGE optimization. However, their approach fails to perform for high-dimension datasets. Inspired from (Zhang *et al.*, 2022*d*), we propose a better defense mechanism with a robust UNet denoiser which gives high performance for high-dimension images.

6.2 Preliminaries

Notations and Basics. Let $x \in \mathbb{R}^d$ is the input sample and $l \in \{1, 2, ..., \mathbb{Y}\}$ be the label. An adversarial attack can perturb x by adding an adversarial noise. In order to defend model f against these adversarial attacks and to provide certified robustness, (Cohen *et al.*, 2019) proposed randomized smoothing (RS), a technique to construct a smoothed classifier f_s from f. It is given as,

$$f_s(x) = \underset{l \in \mathbb{Y}}{\arg\max}\ \mathbb{P}_{\eta \in N(0,\sigma^2 I)}[f(x+\eta) = l], \tag{6.1}$$

where η is the Gaussian noise with standard deviation σ.

Randomized Smoothing and Certified Robustness. (Lecuyer *et al.*, 2019) and (Li *et al.*, 2019*a*) first gave robustness guarantees for the smoothed classifier f_s using RS, but it was loosely bounded. (Cohen *et al.*, 2019) proposed a tight bound on l_2 robustness guarantee for the smoothed classifier f_s. They used Monte Carlo sampling and proposed an effective statistical formulation for predicting and certifying f_s. If the prediction of the base classifier for noise perturbed input samples $(x + \eta)$ is the probability p_f as the topmost prediction, and p_s is the runner-up prediction. Then the smoothed classifier is robust within the radius R_c assuming that f_s gives correct prediction. R_c is the certified radius within which the predictions are guaranteed to remain constant. (Cohen *et al.*, 2019) gave lower and upper bound estimates for p_f and p_s as $\underline{p_f}$ and $\overline{p_s}$ respectively using Monte Carlo technique.

Theorem 1 ((Cohen *et al.*, 2019)) *Given f is the base classifier which returns the target label of the input sample, and f_s is the smoothed classifier, then assuming that f_s classifies correctly, the probabilities of topmost and runner-up predictions are given as:*

$$p_f = max\ \mathbb{P}_\eta[f(x+\eta) = l] \tag{6.2}$$

$$p_s = \underset{l' \neq l}{max}\ \mathbb{P}_\eta[f(x+\eta) = l'], \tag{6.3}$$

where η is the noise sampled from the Gaussian distribution $N(0, \sigma^2)$. Then, f_s is robust inside a radius R_c, which is given as:

$$R_c = \frac{\sigma}{2}[\phi^{-1}(p_f) - \phi^{-1}(p_s)], \tag{6.4}$$

where ϕ^{-1} is the inverse of standard Gaussian CDF. If p_f and p_s hold the below inequality:

$$p_f \geq \underline{p_f} \geq \overline{p_s} \geq p_s \tag{6.5}$$

Then, R_c is given as:

$$R_c = \frac{\sigma}{2}[\phi^{-1}(\underline{p_f}) - \phi^{-1}(\overline{p_s})] \tag{6.6}$$

The enhancement of the smoothed classifier's robustness relies on the application of the "Neyman-Pearson" lemma, which is used to derive the above expressions (Cohen *et al.*, 2019).

Denoised Smoothing (DS). (Salman *et al.*, 2020) proposed that naively applying randomized smoothing gives low robustness, as the standard classifiers are not trained to be robust to the Gaussian perturbation of the input sample. 'DS' augments a custom-trained denoiser D_θ to base classifier f. In this approach, an image-denoising preprocessing step is employed before input samples are passed through f. The denoiser pre-pended smoothed classifier, which is effective at removing the Gaussian noise, is given as,

$$f_s(x) = \arg\max_{l \in \mathbb{Y}} \mathbb{P}_{\eta \in N(0,\sigma^2 I)}[f(D_s^\theta(x+\eta)) = l]. \tag{6.7}$$

In order to obtain the optimal denoiser D_s^θ, DS proposed a stability regularized denoising loss in the first-order optimization setting. It is given as,

$$\begin{aligned} \mathcal{L}_{MSE+STAB}(\theta) = & \mathop{\mathbb{E}}_{T,\eta} \|D_s^\theta(x+\eta) - x\|_2^2 \\ & + \mathop{\mathbb{E}}_{T,\eta}[\mathcal{L}_{CE}(f(D_s^\theta(x+\eta)), f(x))], \end{aligned} \tag{6.8}$$

where T is the training dataset and $\mathcal{L}_{CE}$ is the cross-entropy loss. However, previous approaches provide certified robustness in white-box setting with access to the target model's architectures and parameters. (Salman *et al.*, 2020) first proposed certified defense in black-box setting. However, they utilized surrogate model which requires the information about model type and its function. Recently, (Zhang *et al.*, 2022*d*) proposed certified black-box defense with ZO optimization.

6.3 Methodology

In this section, we first describe the architecture of our proposed robust defense model. We then describe the objective function of our proposed defense mechanism. Lastly, we discuss our two proposed defense mechanisms using RGE and CGE ZO optimization approaches. Our proposed framework is as shown in Figure 6.2.

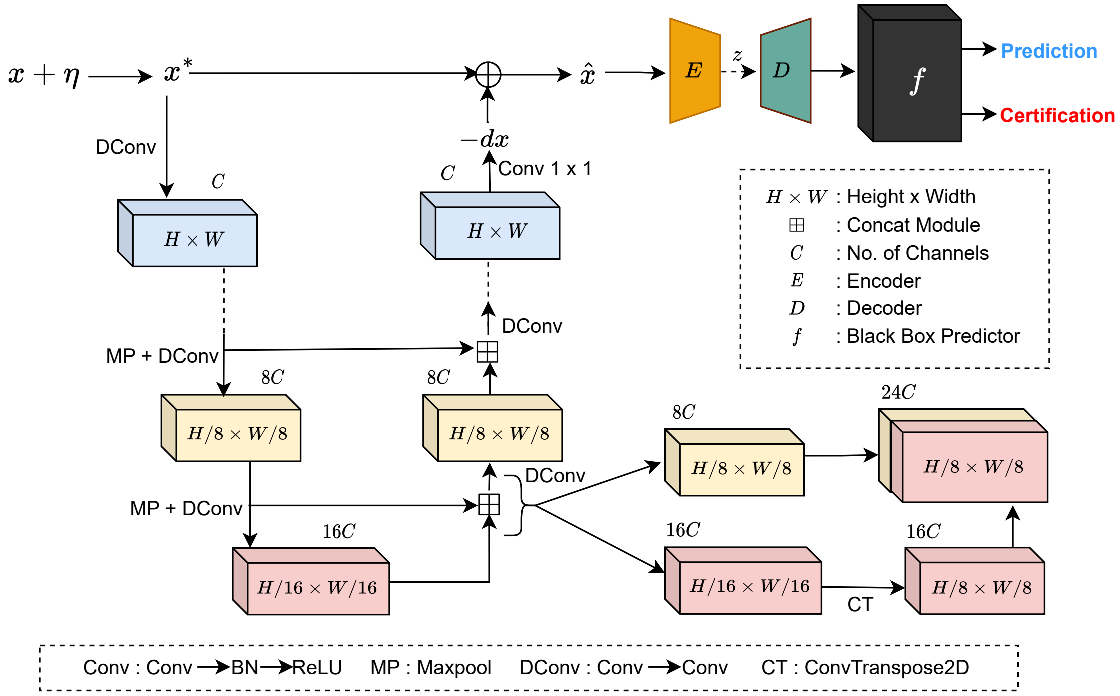


Figure 6.2: An overview of proposed certified defense mechanism via robust UNet denoiser RDUNet. Noise is added to input sample x which is given as input to the robust denoiser. The output of denoiser is the residual map which when added to noisy image x^* gives denoised output $\hat{x}$. The denoised output is input of the autoencoder architecture which is then send as input to black-box model f.

Problem Statement. We aim to defend a black-box model f, where f is used for classification or reconstruction purposes. We consider l_2 norm-ball constrained adversarial attacks as our threat (Goodfellow *et al.*, 2014*b*). **Notations.** We consider input samples as x and and $l \in \{1, 2, ..., L\}$ be the predicted label. Noise and noise-perturbed images are represented as η and x^*, respectively. We denote our proposed learnable RDUNet as D_{θ}^{u}. We denote encoder and decoder as E_{θ_e} and D_{θ_d} respectively. We represent our black-box predictor as f. We denote RGE and CGE ZO optimization as 'R' and 'C', respectively.

6.3.1 Proposed Robust Architecture

We provide a robust black-box defense by pre-pending our proposed robust denoiser RDUNet, followed by a custom-trained AE to the black-box predictor. We show the architecture of our proposed RDUNet in Figure 6.2. The network has a feedforward and a feedback path. We have a stack of conv layers, and each conv layer contains a convolution layer with kernel size 3×3 and padding of 1, followed by batch normalization layer (Ioffe and Szegedy, 2015) and rectified linear unit (Krizhevsky *et al.*, 2012). DConv represents two consecutive conv layers. Our feedforward path consists of five blocks: one is DConv block, and four are Maxpooling + DConv.

RDUNet has four blocks with a fusion module followed by DConv operation in the feedback path. The last block in the feedback path is a convolution layer with kernel size 1×1. Fusion module receives two inputs, one feedback input from the feedback path and the lateral input from the feedforward path. We use ConvTranspose2D (Dumoulin *et al.*, 2016) to upsample the feedback input to the same size as the lateral input. The feedforward path generates feature maps of increasingly lower resolutions, and along the feedback path, the feature maps have an increasingly higher resolution, as is visible from Figure 6.2. The denoised output $\hat{x}$ is the sum of noisy image x^* and $-d\hat{x}$. It is represented as, $\hat{x} = x^* + (-d\hat{x})$.

6.3.2 Proposed Objective Function

Our proposed robustification model is trained by using different losses designed for ensuring three objectives, (i) correct predictions by f on the denoised output $\hat{x}$, (ii) the similarity between the features of clean training samples x and denoised output $\mathrm{D}_\theta^u(x^*)$, and (iii) decrease in the domain gap between the probability distributions of clean samples and denoised output.

Robust Prediction. We use cross-entropy loss $\mathcal{L}_{CE}$ (Zhong *et al.*, 2018*b*) to make sure that the label predicted by the black-box model of the original input sample is same as that of the denoised output of its Gaussian-perturbed counterpart.

$$\mathcal{L}_{CE}(\theta) = \mathbb{E}[-p(f(x)^l)\log(p(f(\mathrm{D}_\theta^u(x^*))^l))], \tag{6.9}$$

where $f(x)$ is a black-box predictor which takes an input x and makes predictions. $p(f(x)^l)$ and $p(f(\mathrm{D}_\theta^u(x^*))^l)$ are the probabilities predicted by f for clean samples and denoised output.

Feature Similarity. We leverage the information that the black-box model trained on the training dataset is highly discriminative. Thus, we propose cosine similarity to learn a mapping between the logit features of the original input images $(f(x))$ and the logits of the output obtained from the denoiser of the Gaussian perturbed inputs $(f(\mathrm{D}^u_\theta(x^*)))$.

$$\mathcal{L}_{CS}(\theta) = \mathbb{E}[\frac{f(x)^\top f(\mathrm{D}^u_\theta(x^*))}{||f(x)||||f(\mathrm{D}^u_\theta(x^*))||}]. \tag{6.10}$$

Domain Similarity. In addition to maintaining the label and feature consistency at the sample level, we want to bring the domain distribution of synthesized denoised images closer to the original input sample by using the maximum mean discrepancy $(MMD(\mu, v))$ (Gretton *et al.*, 2012) on the features of input images $(f_{feat}(x))$ and features of denoised output $(f_{feat}(\mathrm{D}^u_\theta(x^*)))$. This distribution pulling of the original samples and denoised output is inspired by the task of domain adaptation (Mekhazni *et al.*, 2020; Zhang and Wu, 2020).

$$\mathcal{L}_{MMD}(\theta) = MMD(f_{feat}(x), f_{feat}(\mathrm{D}^u_\theta(x^*))). \tag{6.11}$$

Let $\mu = f_{feat}(x)$ and $v = f_{feat}(\mathrm{D}^u_\theta)$, then MMD is the distance between the feature means of μ and v. It is given as (Gretton *et al.*, 2012),

$$MMD(\mu, v) = \| \sum_i \phi(p_i) - \sum_i \phi(q_i)\|^2_{\mathcal{H}}, \tag{6.12}$$

where $\| \cdot \|$ is the norm, ϕ is a function that maps datapoints to a kernel Hilbert space (RKHS) $\mathcal{H}$, $\{p_i\}$ and $\{q_i\}$ are samples drawn from distributions μ and v, respectively.

MMD measures the distance between the expected feature map of the samples from the distributions μ and v in the RKHS ($\mathcal{H}$) induced by ϕ. Minimization of MMD between distributions of clean images and denoised output ensures reconstruction of denoised output that are as close as possible to the clean images. Therefore, the utilization of MMD leads to robust training of the denoiser model which when pre-pended to the black-box target model provides certified black-box defense.

Our Overall Objective Function. We optimize our certified defense mechanism with loss functions given in Eq. 6.9, 6.10 and 6.11

$$\mathcal{L}_{Tot}(\theta) = \mathcal{L}_{CE} + \lambda_{CS}\mathcal{L}_{CS} + \lambda_{MMD}\mathcal{L}_{MMD}, \tag{6.13}$$

where λ_{CS} and λ_{MMD} are the weights assigned to the loss functions. However, since we cannot access the parameters or weights of the model, we cannot optimize our model using standard optimizers like SGD (Amari, 1993) or ADAM (Zhang, 2018) as that would require back-propagation through the predictor. Thus, we utilize ZO optimization approaches where values of functions are approximated instead of using true gradients.

6.3.3 Proposed Black-Box Defense Methods

We discuss in this section our two defense methods in detail:

ZO Robust UNet Denoised Smoothing (ZO-RUDS) Defense (RGE Optimization). In order to achieve ZO-RUDS with our proposed pre-pended RDUNet denoiser to the black-box model f, we represent our objective function as,

$$\mathcal{L}^{R}_{Tot}(\theta) := \mathcal{L}_{Tot}(f(\mathrm{D}^{u}_{\theta}(x^{*})), \tag{6.14}$$

where $\mathcal{L}^{R}_{Tot}(\theta)$ represents that we optimize $\mathcal{L}_{Tot}$ (Eq. 6.13) with RGE ZO optimization (R) (Liu *et al.*, 2020). We calculate gradient estimate of $\mathcal{L}^{R}_{Tot}(\theta)$ as,

$$\hat{\nabla}_{\theta}\mathcal{L}^{R}_{Tot}(\theta) \approx \frac{d\mathrm{D}^{u}_{\theta}(x^{*})}{d\theta}\hat{\nabla}_{z}f(z)|_{z=\mathrm{D}^{u}_{\theta}(x^{*})}, \tag{6.15}$$

where $\hat{\nabla}_{z}f(z)$ is the ZO gradient estimate of f. We calculate the RGE ZO gradient estimate of $\mathcal{L}^{R}_{Tot}(\theta)$ by the difference of two function values along a set of random direction vectors. It is represented as:

$$\hat{\nabla}_{\theta}\mathcal{L}^{R}_{Tot}(\theta) = \sum_{k=0}^{q-1}[\frac{d}{\xi \cdot q}(\mathcal{L}_{Tot}(\theta + \xi u_{k}) - \mathcal{L}_{Tot}(\theta))u_{k}], \tag{6.16}$$

q are the querying directions, and $u \in \{1, 2, ..., q\}$ are q independently and uniformly drawn random vectors from a unit Euclidean sphere $\xi > 0$ is the smoothing parameter with a small step size of 0.005. We show the corresponding algorithm of our ZO-RUDS defense in Algorithm 4. (Zhang *et al.*, 2022*d*) directly applied ZO to the previous approach (Salman *et al.*, 2020), with poor performance in the RGE optimization approach. However, we show in Table 6.1 in Section 6.4.1, that after using our proposed RDUNet in ZO-RUDS defense mechanism, we achieve a huge increase of **35%** in certified accuracy compared to (Zhang *et al.*, 2022*d*). This proves that our proposed RDUNet type of architecture enables the model to learn fine-scale information while maintaining a low reconstruction error in comparison to previous custom-trained denoisers (Salman *et al.*, 2020; Zhang *et al.*, 2022*d*).

ZO Autoencoder-based Robust UNet Denoised Smoothing (ZO-AE-RUDS). We further pre-pend RDUNet followed by an encoder E_{θ_e} and a decoder D_{θ_d} to the black-box predictor for better performance on datasets like Tiny Imagenet and STL-10 with large dimensionality as shown in Figure 6.3. We extend our defensive operation where the new black-box is $D + f$, and the new white-box system is RDUNet + E by plugging AE between our RDUNet and black-box predictor f. It ensures that ZO optimization can be carried out in a feature embedding space with low dimensions. However, the autoencoder can also lead to over-reduced features for these datasets leading to poor performance as in (Zhang *et al.*, 2022*d*).

We observe from Figure 6.4 in Section 6.4.2 that for low-dimension datasets like CIFAR-10, there is an increase of approximately 2%, whereas, for high-dimension dataset STL-10, there is an increase of approximately 10% between performances of ZO-RUDS and ZO-AE-RUDS. Thus, AE enables us to conduct ZO optimization in a feature-embedding space which ensures the feasibility of least-variance CGE. Our ZO-AE-RUDS with RDUNet denoiser overcomes the curse of dimensionality, and due to the ability of RDUNet to learn fine-scaled information leads to high performance. The objective function for our ZO-AE-RUDS defense is represented as,

$$\mathcal{L}^{C}_{Tot}(\theta) := \mathcal{L}_{Tot}(f(D_{\theta_d}(z)); z = E_{\theta_e}(\mathrm{D}^{u}_{\theta}(x^{*})), \tag{6.17}$$

where $\mathcal{L}^{C}_{Tot}(\theta)$ represents that we optimize $\mathcal{L}_{Tot}$ (Eq. 6.13) with CGE ZO optimization

ALGORITHM 4: ZO-RUDS Defense (RGE)

```
Require: Input $x$, noise $\eta$, smoothing parameter $\xi$, query directions $q$, dimensionality
    $d$, black-box predictor $f$, and initial parameters $\theta$ of RDUNet.
Ensure: Trained RDUNet $\mathrm{D}^{u}_{\theta}$
$\hat{x} = \mathrm{D}^{u}_{\theta}(x + \eta) = \mathrm{D}^{u}_{\theta}(x^{*})$,
Calculate $\mathcal{L}_{Tot}(f(x), f(\hat{x}))$ (Eq. 6.13),
for $k = 0$ to $q - 1$ do
    Obtain a random direction vector $u_k$ with Normal distribution $N(\mu, \sigma)$,
    Calculate $\hat{x}_q = \hat{x} + \xi \cdot u_k$,
    Calculate $\mathcal{L}_{Tot}(f(x), f(\hat{x}_q))$ (Eq. 6.13),
    Calculate gradient estimation using Eq. 6.16,
end for
```

(C) (Liu *et al.*, 2020). We calculate gradient estimate of $\mathcal{L}^C_{Tot}(\theta)$ as,

$$\hat{\nabla}_\theta \mathcal{L}^C_{Tot}(\theta) \approx \frac{dE_{\theta_e}(\mathrm{D}^u_\theta(x^*))}{d\theta} \hat{\nabla}_z f(z)|_{z=E_{\theta_e}(\mathrm{D}^u_\theta(x^*))}, \tag{6.18}$$

where z is the latent feature vector with reduced dimension $d_r < d$. This reduction in dimension makes the CGE ZO optimization approach feasible. Thus, we utilize CGE (Lian *et al.*, 2016; Liu *et al.*, 2018*b*) and calculate the gradient estimate of training objective function $\mathcal{L}_{Tot}(\theta)$ as,

$$\hat{\nabla}_\theta \mathcal{L}^C_{Tot}(\theta) = \sum_{k=0}^{d-1} [\frac{(\mathcal{L}_{Tot}(\theta + \xi e_k) - \mathcal{L}_{Tot}(\theta - \xi e_k))}{\xi} e_k], \tag{6.19}$$

where $e^k \in \mathbb{R}^d$ is the kth elementary basis vector, with 1 at the kth coordinate and 0s elsewhere. We show the corresponding algorithm of our ZO-AE-RUDS defense using CGE optimization in Algorithm 5.

ALGORITHM 5: ZO-AE-RUDS Defense (CGE)

```
Require: Input $x$, noise $\eta$, smoothing parameter $\xi$, query directions $q$, dimensionality
    $d$, black-box predictor $f$, decoder $D_\theta$, initial parameters $\theta$ of RDUNet, and $\theta_e$ of
    white-box encoder $E_{\theta_e}$
Ensure: Trained $\mathrm{D}^u_\theta + E_{\theta_e}$
$z = E_{\theta_e}(\mathrm{D}^u_\theta(x + \eta))$,
$\hat{x} = \mathrm{D}^u_\theta(z)$,
Calculate $\mathcal{L}_{Tot}(f(x), f(\hat{x}))$ (Eq. 6.13),
for $k = 0$ to $d - 1$ do
    Obtain a elementary basis direction vector $e_k$,
    Calculate $\hat{x}^+_q = \hat{x} + \xi \cdot e_k$ and $\hat{x}^-_q = \hat{x} - \xi \cdot e_k$,
    Calculate $\mathcal{L}_{Tot}(f(x), f(E_{\theta_e}(\hat{x}^+_q)))$ (Eq. 6.13),
    Calculate $\mathcal{L}_{Tot}(f(x), f(E_{\theta_e}(\hat{x}^-_q)))$ (Eq. 6.13),
    Calculate gradient estimation using Eq. 6.19,
end for
```

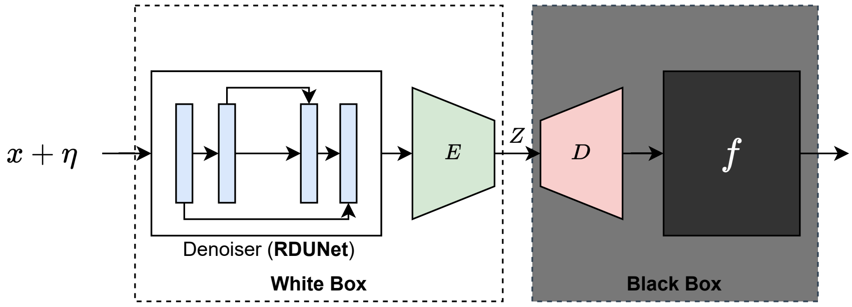


Figure 6.3: Architecture of our defense technique ZO-AE-RUDS.

6.4 Experimental Settings

Datasets and Models. We evaluate the results on CIFAR-10 (Krizhevsky and Hinton, 2009*a*), CIFAR-100 (Krizhevsky and Hinton, 2009*b*), Tiny Imagenet (Liu *et al.*, 2017) and STL-10 (Coates *et al.*, 2011) datasets for classification task. For image reconstruction, we focus on MNIST (LeCun *et al.*, 1998*b*) dataset. We consider pre-trained models ResNet-110 for CIFAR-10 and ResNet-18 for STL-10. For CIFAR-100, we use ResNet-50 as the target classifier.

CIFAR-10. It comprises 60,000 color images, 32×32 pixels in size, classified into 10 classes, each with 6,000 images. 50,000 images are used for training, and 10,000 for testing. The dataset is divided into five training batches and one test batch, each containing 10,000 images.

CIFAR-100. The dataset shares similarities with CIFAR-10, with the exception that it comprises 100 classes, each consisting of 600 images. For every class, there are 500 training images and 100 testing images.

Tiny ImageNet. It comprises 100,000 colored images that are downsized to 64×64 pixels. The dataset consists of 200 classes, with each class containing 500 images. For every class, there are 500 training images, 50 validation images, and 50 test images.

STL-10. It is an image dataset with 100,000 unlabeled and 13,000 labeled images from 10 classes. It is popular for evaluating self-taught learning algorithms and contains color images of 96×96 pixels.

MNIST. MNIST database, short for Modified National Institute of Standards and Technology database, consists of handwritten digits with a training set of 60,000 examples and a test set of 10,000 examples.

Implementation Details. We use learning rate 10^{-4} and weight decay by 10 at every 100 epochs with total 600 epochs. We set the smoothing parameter $\zeta = 0.005$ for a fair comparison with SOTA. We sample noise η with mean $\mu = 0$ and variance $\sigma^2 = 0.25$ from Normal distribution. We optimize ZO-RUDS with RGE and ZO-AE-RUDS with CGE optimization unless otherwise stated. We set the values of hyperparameters $\lambda_{MMD} = 4$ and $\lambda_{CS} = 1$. We use batch-size of 256. RGE ZO optimization is denoted as 'R', and CGE ZO optimization is 'C' in all the experiments. It is to be noted that we do not use original labels of the datasets, thus making our technique practical and scalable as it is not practically possible to annotate large-scale datasets. We measure the robustification using standard certified accuracy (SCA) (%) and robust certified accuracy (RCA) (%). In each tabular result, we have shown best results for certified black-box defense in bold. We also show the results of certified defense in white-box setting and prove that our results are comparative or better than previous white-box certified defense methods even with no information of model weights or parameters.

Evaluation Metrics. We measure the robustification of our model using standard certified accuracy (SCA) at l_2-radius $(r) = 0$ and robust certified accuracy (RCA) at $r = \{0.25, 0.50, 0.75\}$. Higher certified accuracy (CA) ensures that for a given r, more percentage of correctly predicted samples have certified radii larger than r (Cohen *et al.*, 2019).

Computational Complexity. We optimize our training time with the help of parallel processing and matrix operations. The averaged training time on NVIDIA A100-SXM4 for one epoch is approximately $\sim$ 30sec for FO-AE-RUDS in white-box setting. For our proposed certified black-box defense approach ZO-RUDS (RGE) the averaged training time is $\sim$ 30min and $\sim$ 33min for ZO-AE-RUDS (CGE), on the CIFAR-10 dataset.

6.4.1 Comparison with SOTA

We compare our proposed defense methods ZO-RUDS and ZO-AE-RUDS with previous certified defense approaches in white-box (First-order (FO)) and black-box (Zeroth-order (ZO)) settings in Table 6.1. In white-box setting, we compare our approach with RS (Cohen *et al.*, 2019), DS (FO-DS) (Salman *et al.*, 2020) and FO-AE-DS (Zhang *et al.*, 2022*d*). Our proposed defense methods ZO-RUDS and ZO-AE-RUDS with RDUNet comfortably outperform SOTA (Zhang *et al.*, 2022*d*) by a large margin of **35%** and **9%** in RGE and CGE optimization approaches, respectively. It consistently achieves higher certified robustness across different r. The high performance of ZO-RUDS over SOTA signifies that our method leads to an effective DS-oriented robust defense even without additional custom-trained autoencoder. We observe that our method provides better results in the black-box setting and better performance than RS (Cohen *et al.*, 2019) and DS (Salman *et al.*, 2020), which defend the model in the white-box setting, thus proving the effectiveness of our approach.

FO-AE-DS is the first-order implementation of ZO-AE-DS. We compare our pro-

Method	Type	SCA	RCA (r)		
			0.25	0.50	0.75
RS (Cohen *et al.*, 2019)	W	76.22	61.20	43.23	25.67
DS (FO-DS) (Salman *et al.*, 2020)	W	70.80	53.31	40.89	25.90
DS (Salman *et al.*, 2020)	B	74.89	44.56	18.20	14.39
FO-AE-DS (Zhang *et al.*, 2022*d*)	W	75.92	60.54	46.45	32.19
ZO-DS (R) (Zhang *et al.*, 2022*d*)	B	42.34	18.12	5.01	0.19
ZO-AE-DS (R) (Zhang *et al.*, 2022*d*)	B	60.90	43.25	26.23	7.78
ZO-AE-DS (C) (Zhang *et al.*, 2022*d*)	B	70.90	53.45	33.21	12.45
FO-RUDS	W	77.32	61.78	49.43	34.56
FO-AE-RUDS	W	79.95	63.24	49.21	35.82
ZO-RUDS (R)	B	**77.89**	**58.92**	**38.31**	**21.93**
ZO-AE-RUDS (R)	B	**76.87**	**58.23**	**37.72**	**20.65**
ZO-AE-RUDS (C)	B	**79.87**	**61.32**	**42.90**	**23.21**

Table 6.1: Comparison with SOTA certified defense techniques in white-box (W) and black-box (B) settings on CIFAR-10 dataset. 'R' and 'C' are RGE and CGE optimization techniques. 'q'=192.

posed ZO-AE-RUDS in RGE and CGE ZO optimization techniques as ZO-AE-RUDS (R) and ZO-AE-RUDS (C), respectively. We observe that our prepended robust denoiser RDUNet followed by autoencoder leads to better performance with the CGE optimization approach (ZO-AE-RUDS (C)). We show the results for FO-RUDS and FO-AE-RUDS with our proposed robust denoiser RDUNet in white-box setting. We observe that our FO-RUDS and FO-AE-RUDS show a performance improvement of approximately 7% and 3% on FO-DS (Salman *et al.*, 2020) and FO-AE-DS (Zhang *et al.*, 2022*d*) respectively, in standard certified accuracy (SCA) and similarly provide robustness for all radii. We show that our black-box defense techniques (ZO-RUDS, ZO-AE-RUDS (R), and ZO-AE-RUDS (C)) achieve comparative performance to our proposed defenses in white-box settings (FO-RUDS and FO-AE-RUDS).

6.4.2 Performance on Image Classification

Performance on different number of queries. We show the performance for other queries $q = 20, 100$ for CIFAR-10 and $q = 576$ for STL-10 dataset in Table 6.2. Our proposed robustification outperforms SOTA (Zhang *et al.*, 2022*d*) on all these queries by a huge margin for SCA and RCA evaluation metrics at different l_2 radii. We observe that (Zhang *et al.*, 2022*d*) has poor performance for high-dimension images even after increasing the number of queries or using CGE optimization with auto-encoder to decrease the variance caused by RGE optimization. This may happen due to two reasons;

CIFAR-10					
q	Model	SCA	RCA (r)		
			0.25	0.50	0.75
20	ZO-DS (R) (Zhang *et al.*, 2022*d*)	18.56	3.88	0.53	0.26
	ZO-AE-DS (C) (Zhang *et al.*, 2022*d*)	40.67	27.90	17.84	7.10
	ZO-RUDS (R)	**62.10**	**42.43**	**36.23**	**24.56**
	ZO-AE-RUDS (C)	**63.59**	**51.34**	**39.01**	**30.23**
100	ZO-DS (R) (Zhang *et al.*, 2022*d*)	39.82	17.90	4.71	0.29
	ZO-AE-DS (C) (Zhang *et al.*, 2022*d*)	54.32	40.90	23.98	9.35
	ZO-RUDS (R)	**74.60**	**58.71**	**39.56**	**27.82**
	ZO-AE-RUDS (C)	**76.34**	**62.46**	**44.29**	**30.22**
STL-10					
576	ZO-DS (R) (Zhang *et al.*, 2022*d*)	37.59	21.23	8.67	2.56
	ZO-AE-DS (C) (Zhang *et al.*, 2022*d*)	44.78	33.41	26.10	16.43
	ZO-RUDS (R)	**58.20**	**47.83**	**40.32**	**29.89**
	ZO-AE-RUDS (C)	**68.29**	**57.93**	**47.31**	**33.22**

Table 6.2: Comparison of our defense with previous ZO defense approaches. ‘q’ is the number of queries.

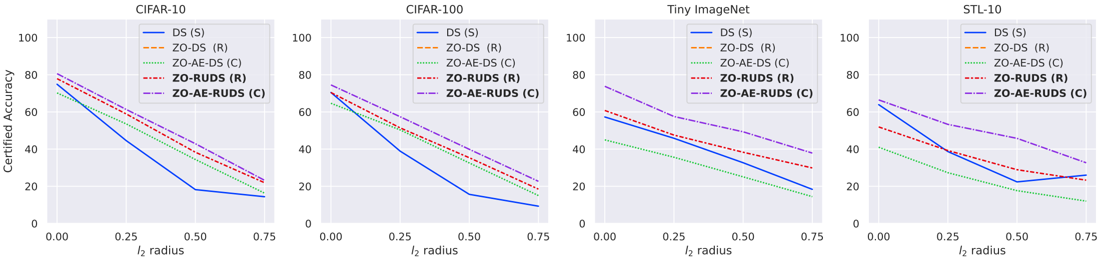


Figure 6.4: Comparison of Certified Accuracy on low-dimension (CIFAR-10, CIFAR-100) and high-dimension (STL-10 and Tiny Imagenet) datasets for different l_2 radius at query $q = 192$.

First, the bottleneck in the autoencoder architecture constrains the fine-scaled information necessary for reconstructing denoised images, and Second, the over-reduced feature dimension in high-dimension images could hamper the performance. Our proposed denoiser RDUNet with lateral connections between the encoder and decoder ensures better reconstruction of denoised output, which ensures prediction with high certified accuracy. RDUNet decreases model variance as it consists of downsampling and upsampling layers in the encoding and decoding path, which makes the model invariant to changes in image dimensions and thus performs better for high dimensions as well. The lateral skip connections help the model learn fine-scale information, thus overcoming the disadvantage of auto-encoder constraining fine-scale information.

Performance on low-dimension (CIFAR-10, CIFAR-100) and high-dimension (STL-

10, Tiny Imagenet) classification datasets. We compare our proposed defense methods ZO-RUDS (R) and ZO-AE-RUDS (C) with previous black-box defense methods DS (S) (Salman *et al.*, 2020), ZO-DS (R) (Zhang *et al.*, 2022*d*) and ZO-AE-DS (C) (Zhang *et al.*, 2022*d*) at various r for low and high-dimension classification datasets in Figure 6.4. Our proposed approaches ZO-RUDS and ZO-AE-RUDS beat the SOTA (Zhang *et al.*, 2022*d*) by a large margin of **30.21%** and **8.87%** in RGE, and CGE optimization approaches respectively for CIFAR-100 dataset. It outperforms SOTA by a huge margin for all other radii as well. We show that unlike (Zhang *et al.*, 2022*d*), which gives better performance only for the CGE optimization approach after appending autoencoder in ZO-AE-RUDS, our proposed RDUNet denoiser when appended to predictor performs better for both RGE and CGE optimization approaches. We observe that SOTA (Zhang *et al.*, 2022*d*) fails to perform even after the addition of an autoencoder to the network for high-dimension Tiny Imagenet and STL-10 datasets. Our defense methods ZO-RUDS and ZO-AE-RUDS beat SOTA (Zhang *et al.*, 2022*d*) by a huge margin of **24.81%** and **25.84%** respectively, for Tiny Imagenet dataset. Similar observations can be made for CIFAR-10 and STL-10 datasets at different certified radii, as shown in Figure 6.4.

6.4.3 Performance on Image Reconstruction

Previous works (Antun *et al.*, 2020; Raj *et al.*, 2020; Wolf) show that image reconstruction networks are vulnerable to adversarial attacks like PGD attacks (Madry *et al.*, 2017). We compare our proposed defense methods ZO-RUDS and ZO-AE-RUDS with previous white-box and black-box defense methods in Table 6.3. We follow the settings of (Zhang *et al.*, 2022*d*) and aim to recover the original sample using a pre-trained reconstruction network (Raj *et al.*, 2020) under adversarial perturbations generated by a 40-step l_2 PGD attack under $||\delta||_2 = \{0, 1, 2, 3, 4\}$. We use root mean square error (RMSE) and structural similarity (SSIM) (Hore and Ziou, 2010) to find the similarity between the original and reconstructed image. We observe that our method beats SOTA (Zhang *et al.*, 2022*d*) with low RMSE and high SSIM scores for all the values of $||\delta||_2$. Our defense beats FO defense methods as well as the vanilla model, proving the robustness provided by our method for the reconstruction task. We observe that at high values of $||\delta||_2$ perturbation the performance of vanilla model (Raj *et al.*, 2020) decreases drastically.

Method	$\|\delta\|_2 = 0$		$\|\delta\|_2 = 1$		$\|\delta\|_2 = 2$		$\|\delta\|_2 = 3$		$\|\delta\|_2 = 4$	
	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
Vanilla	0.1213	0.7934	0.3251	0.4367	0.4629	0.1468	0.6129	0.0494	0.5976	0.0168
FO-DS	0.1596	0.7415	0.1692	0.6934	0.2182	0.5421	0.2698	0.3956	0.3245	0.3178
FO-AE-DS	0.1475	0.7594	0.1782	0.7025	0.2182	0.5421	0.2693	0.4163	0.3176	0.3293
ZO-DS (R)	0.1892	0.5345	0.2267	0.4634	0.2634	0.3689	0.3092	0.2792	0.3482	0.2177
ZO-AE-DS (C)	0.1398	0.6894	0.1634	0.7099	0.2126	0.5472	0.2689	0.4188	0.3367	0.3294
ZO-RUDS (R)	**0.1232**	**0.7924**	**0.1465**	**0.7991**	**0.2053**	**0.5966**	**0.2380**	**0.4591**	**0.3082**	**0.3811**
ZO-AE-RUDS (C)	**0.1219**	**0.7926**	**0.1392**	**0.8346**	**0.1872**	**0.6648**	**0.2174**	**0.6102**	**0.2679**	**0.5236**

Table 6.3: Performance comparison with SOTA image reconstruction tasks which are Vanilla Model (Raj *et al.*, 2020), FO-DS (Salman *et al.*, 2020), FO-AE-DS (Zhang *et al.*, 2022*d*), ZO-DS (R) (Zhang *et al.*, 2022*d*) and ZO-AE-DS (C) (Zhang *et al.*, 2022*d*) on MNIST dataset. ($q = 192$)

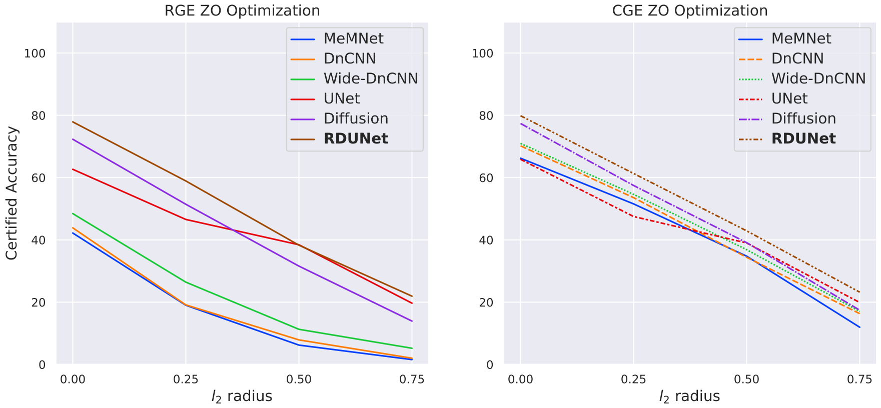


Figure 6.5: Effect of different denoisers on our RGE and CGE ZO optimization-based defense approaches for different l_2-radii at q=192. Dataset is CIFAr-10.

6.4.4 Ablation Study

We show the effect of different denoiser architectures and loss functions in our proposed defense mechanism.

Effect of various denoisers. We compare our proposed robust denoiser RDUNet with previous denoisers MemNet (Tai *et al.*, 2017), DnCNN (Zhang *et al.*, 2017*a*), wide-DnCNN (Zhang *et al.*, 2017*a*), UNet (Ronneberger *et al.*, 2015) and Diffusion (Ho *et al.*, 2020) in Figure 6.5. We show that after appending denoiser DnCNN as proposed in (Salman *et al.*, 2020; Zhang *et al.*, 2022*d*) with 17 layers primarily including Conv+BN+ReLU layers and wide-DnCNN with 128 deep layers to the black-box model in the RGE optimization leads to poor performance. We observe that appending these denoisers to the autoencoder architecture in the CGE optimization improves the robust-

Loss	SCA	RCA (r)		
		0.25	0.50	0.75
$\mathcal{L}_{CE}$	72.67	54.90	36.01	17.03
$\mathcal{L}_{CE} + \mathcal{L}_{CS}$	73.21	56.04	36.45	18.79
$\mathcal{L}_{CE} + \mathcal{L}_{CS} + \mathcal{L}_{MMD}$	**79.87**	**61.32**	**42.90**	**23.21**

Table 6.4: Effect of loss functions on our proposed (ZO-AE-RUDS) for $q = 192$. Dataset is CIFAR-10.

Training Strategy	SCA	RCA (l_2-radius)		
		0.25	0.50	0.75
$(\mathrm{D}^u_\theta)_{finetune}$	67.52	53.44	29.56	12.87
$(\mathrm{D}^u_\theta)_{scratch}$	**79.87**	**61.32**	**42.90**	**23.21**

Table 6.5: Effect of RDUNet training strategies on ZO-AE-RUDS for $q = 192$. Dataset is CIFAR-10.

fication of the model to some extent. We show that the diffusion model gives comparatively better performance than other conventional denoisers. However, RDUNet gives better performance than diffusion models. It may happen due to multiple noise addition, which makes ZO optimization of the diffusion model unstable. These results conclude that our proposed RDUNet provides robustness to a black-box model against adversarial perturbations.

Effect of Loss Functions. We show in Table 6.4, the effect of cosine similarity loss $\mathcal{L}_{CS}$ at the sample level and MMD loss $\mathcal{L}_{MMD}$ between the probability distributions of the original samples and denoised output. We observe in Table 6.4 that after applying $\mathcal{L}_{CS}$ in addition to cross-entropy loss, our model's performance increases by 1% and using $\mathcal{L}_{MMD}$ the performance increases by approximately 6%. This shows that bringing closer the features of the original sample and denoised output at the instance and domain level increase the robustness of the black-box model.

Effect of Training Strategies. We show the effect of our training schemes (over RDUNet) on ZO-AE-RUDS across different r in Table 6.5. $(\mathrm{D}^u_\theta)_{finetune}$ represents pre-training RDUNet and then fine-tuning it in our defense method, and $(\mathrm{D}^u_\theta)_{scratch}$ is training RDUNet from scratch. We show that training from scratch gives better performance than pre-training and fine-tuning RDUNet. We observe during training that $(\mathrm{D}^u_\theta)_{finetune}$ achieves very high performance at the initial training stage; however, it does not improve as training progresses. Pre-training the denoiser causes the optimization to get stuck at a local optima leading to decreased performance.

CIFAR-10					
	Model	SCA	RCA (r)		
(σ)			0.25	0.50	0.75
0.5	FO-DS (Salman *et al.*, 2020)	66.93	45.78	23.34	6.20
	ZO-DS (R) (Zhang *et al.*, 2022*d*)	35.67	12.45	2.58	0.01
	FO-AE-DS (Zhang *et al.*, 2022*d*)	67.78	48.90	26.76	9.21
	ZO-AE-DS (C) (Zhang *et al.*, 2022*d*)	63.24	42.98	25.67	4.56
	ZO-RUDS (R)	**70.44**	**49.63**	**31.21**	**12.48**
	ZO-AE-RUDS (C)	**71.52**	**49.68**	**32.37**	**13.56**
1.0	FO-DS (Salman *et al.*, 2020)	51.29	30.47	8.36	1.67
	ZO-DS (R) (Zhang *et al.*, 2022*d*)	21.89	6.34	0.12	0.00
	FO-AE-DS (Zhang *et al.*, 2022*d*)	54.97	37.83	20.63	5.64
	ZO-AE-DS (C) (Zhang *et al.*, 2022*d*)	48.90	37.21	18.39	0.79
	ZO-RUDS (R)	**55.25**	**38.98**	**23.54**	**3.58**
	ZO-AE-RUDS (C)	**55.26**	**38.84**	**22.49**	**4.29**

Table 6.6: Comparison of our proposed approach with previous ZO optimization approaches on the CIFAR-10 dataset for different noise levels. Query ‘q’=192.

In this work, we study the problem of certified black-box defense, aiming to robustify the black-box model with access only to input queries and output feedback. First, we proposed two novel defense mechanisms, ZO-RUDS and ZO-AE-RUDS, which substantially enhance the defense and optimization performance by reducing the variance of ZO gradient estimates. Second, we proposed a novel robust denoiser RDUNet that provides a scalable defense by directly integrating denoised smoothing with RGE ZO optimization, which was not feasible in previous works. We show that RDUNet gives high performance by further appending autoencoder (AE) as in ZO-AE-RUDS defense. Lastly, we proposed an objective function with MMD loss, bringing the distribution of denoised output closer to clean data. Our elaborate experiments demonstrate that ZO-RUDS and ZO-AE-RUDS achieve SOTA-certified defense performance on classification and reconstruction tasks.

Effect of Noise Level (σ)**.** We show the performance of our proposed defense techniques for $\sigma = \{0.50, 1.0\}$ in Tables 6.6 and 6.7, in addition to the certified robustness calculated at noise level $\sigma = \{0.25\}$ unless otherwise stated. Our defense techniques consistently outperform the SOTA black-box defense approaches for low-dimension CIFAR-10 and high-dimension STL-10 datasets for all values of σ.

Variation of ZO techniques over number of epochs. We show the RGE and CGE optimization of our black-box defenses ZO-RUDS and ZO-AE-RUDS, as training progresses in Figure 6.6. We observe that ZO-AE-RUDS (CGE) gives high performance at

STL-10					
	Model	SCA	RCA (r)		
(σ)			0.25	0.50	0.75
0.5	ZO-DS (R) (Zhang *et al.*, 2022*d*)	31.56	16.34	1.57	0.00
	ZO-AE-DS (C) (Zhang *et al.*, 2022*d*)	34.67	22.65	18.42	9.81
	ZO-RUDS (R)	**49.97**	**39.81**	**33.54**	**20.10**
	ZO-AE-RUDS (C)	**61.20**	**48.71**	**31.23**	**20.51**
1.0	ZO-DS (R) (Zhang *et al.*, 2022*d*)	12.56	1.24	0.34	0.09
	ZO-AE-DS (C) (Zhang *et al.*, 2022*d*)	22.65	13.42	6.32	0.05
	ZO-RUDS (R)	**33.28**	**24.51**	**16.53**	**7.62**
	ZO-AE-RUDS (C)	**43.27**	**31.27**	**23.33**	**9.80**

Table 6.7: Comparison of our proposed approach with previous ZO optimization approaches on STL-10 dataset for different noise levels. Query ‘q’=192.

initial epochs, however as training progresses both defense approaches give comparative performance. We observe high increase in performance at initial epochs and slight increase in accuracy at higher epochs.

Performance on different classifiers. We show the performance of our proposed defense techniques ZO-RUDS (R) and ZO-AE-RUDS (C) on different classifiers (VGG-16 and Vision Transformer (ViT-16-L(224)) in Table 6.8. These results in addition to the results provided in Section 6.4.2 prove that our proposed defense method leads to robustification of wide variety target models with various architectures.

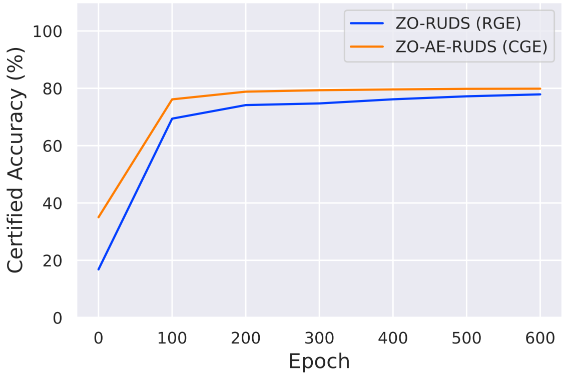


Figure 6.6: Training progress of our certified black-box defense ZO-RUDS (RGE) and ZO-AE-RUDS (CGE) over number of epochs. Dataset used is CIFAR-10. Query ‘q’ = 192, $\sigma = 0.25$.

Model	SCA	RCA (r)		
VGG-16				
		0.25	0.50	0.75
ZO-RUDS (R)	77.24	59.93	38.01	21.35
ZO-AE-RUDS (C)	79.83	61.34	43.66	23.90
ViT-16-L(224) (Tseng *et al.*, 2022)				
ZO-RUDS (R)	77.91	60.24	38.75	21.26
ZO-AE-RUDS (C)	79.89	62.36	43.87	24.01

Table 6.8: Certified Accuracy for different classifiers on CIFAR-10 dataset for noise level $\sigma = \{0.25\}$. Query 'q'=192.

6.5 Chapter Summary

In this chapter, we focus on the problem of certified black-box defense, which aims to improve the robustness of a black-box model by using only input queries and output feedback. Two new defense mechanisms, ZO-RUDS and ZO-AE-RUDS, are proposed to enhance defense and optimization performance by reducing the variance of ZO gradient estimates. In addition, a novel robust denoiser called RDUNet is introduced, which integrates denoised smoothing with RGE ZO optimization, providing a scalable defense that was not possible in previous works. The high performance of RDUNet is demonstrated by appending an autoencoder (AE) as in the ZO-AE-RUDS defense. Finally, an objective function with MMD loss is proposed to bring the distribution of denoised output closer to clean data. The experiments conducted in this study show that ZO-RUDS and ZO-AE-RUDS achieve state-of-the-art certified defense performance on classification and reconstruction tasks.

CHAPTER 7

Limitations

7.1 GAN-based Unsupervised Domain Adaptation

Failure Cases:

- Significant differences in the feature distributions between the source and target domains causes our adapted model to fail in capturing finer discriminative details like bags and gender across domains. The Re-ID task primarily extracts features such as colour of clothes, additional useful cues such as bag. While in some cases, cues such as bag may not be preserved, yet the clothing is adapted efficiently. This maybe due to the reason that only few images have such additional cues and the model may not be able to learn them. Further, the clothing appearance is the most significant cue and the network primarily learns to match this. In some cases gender may not be useful information as many shots are taken from different views and quite a few of them are taken from an angle where identifying gender is very hard. Thus, while the network may inherently learn some gender based cues, it may not consider it as a strong discriminative factor.

Restrictions and Assumptions:

- The target domain is assumed to have no labeled data available, requiring adaptation techniques to rely solely on the knowledge from the labeled source domain.
- We assume that the source and target domains share some common identity features and that the domain discrepancy mainly arises from different environment (pose and background) distributions.

7.2 Wasserstein Metric Attack

Failure Cases:

- In certain instances, achieving a significant Wasserstein distance requires the introduction of substantial perturbations into the input data. While effective in manipulating model predictions, these large perturbations can inadvertently result in visually conspicuous adversarial examples. This, in turn, poses a challenge to the robustness of the attack, as these visually noticeable artifacts may be easily detected by human observers, potentially undermining the adversarial attack's effectiveness. Balancing the need for large perturbations to enhance Wasserstein distance while maintaining inconspicuousness remains an important area of research, seeking to bolster the robustness of machine learning models against increasingly sophisticated adversarial attacks.

Restrictions and Assumptions:

- A significant limitation of Wasserstein metric attack is it's reliance on having complete knowledge of the target model's architecture and training mechanism, which includes awareness of the specific use of Wasserstein distance as a training objective. It is effective under white-box setting where training information is accessible.

7.3 Meta Perturbed Defense

Failure Cases:

- Despite the significant strides made in fortifying our defense model against adversarial attacks through meta-learning and the introduction of specific noise patterns during training, it is essential to acknowledge its inherent limitations. Adversarial attacks are continually evolving, and some may possess the capability to adapt to the noise patterns established during meta-learning, thereby circumventing our defense mechanisms. These adaptive attacks challenge the efficacy of our model and highlight the need for ongoing research and innovation to develop more robust and adaptive defense strategies. While our defense model represents a valuable step forward, we remain vigilant in recognizing that it may not be universally effective against all forms of adversarial threats.

Restrictions and Assumptions:

- Firstly, meta perturbed defense relies on the assumption that the noise patterns learned through meta-learning will effectively generalize to both unseen data and various types of adversarial attacks. However, it is essential to acknowledge that there might be instances where this generalization does not hold, potentially limiting the defense's efficacy. Secondly, the meta-learning approach can introduce an additional computational overhead during the training process, making it more resource-intensive compared to conventional training methods. Despite these challenges, we continue to explore ways to optimize the defense model's performance while considering the trade-offs between robustness and computational costs.

7.4 Certified Defense

Failure Cases:

- Failure cases in certified defense methods stem from their reliance on guarantees within a predefined perturbation space established during certification. While these defenses can offer robustness within the certified bounds, they become susceptible to attacks exploring perturbation spaces beyond those limits, leaving the model vulnerable to exploitation. Moreover, the complexity of the mathematical computations involved in certified defenses makes them computationally expensive and challenging to scale effectively to larger models and datasets. Balancing the trade-offs between robustness and computational efficiency remains a key challenge in developing certified defense mechanisms that can provide reliable protection against a wide range of adversarial threats.

Restrictions and Assumptions:

- Firstly, it assumes that the model's predictions exhibit continuity and smoothness concerning the input space. However, in cases where the model's decision boundaries contain discontinuities, the defense may be weakened. Although our certified defense method proves effective against attack strategies it is specifically certified for, there is a concern about its generalization to unseen or adapted attack methods. The defense's effectiveness may vary depending on the magnitude of perturbations applied. Extremely large or subtle perturbations could fall outside the certified bounds, potentially compromising the defense's overall robustness. Addressing these assumptions and limitations requires a comprehensive approach that accounts for various attack scenarios and perturbation magnitudes to develop a more resilient and adaptive defense mechanism.

CHAPTER 8

Thesis Summary and Future Directions

8.1 Introduction

This thesis explores the application of generative learning and adversarial attack and defense techniques for Re-ID and classification tasks in computer vision. With the growing availability of large datasets and advances in deep learning algorithms, Re-ID and classification tasks have become increasingly important in areas such as surveillance, security, and autonomous vehicles. However, these systems are vulnerable to adversarial attacks, where imperceptible perturbations can cause misclassification or Re-ID errors. Generative learning and adversarial defense strategies have shown promise in mitigating the impact of these attacks, but their effectiveness in real-world scenarios remains unclear. In this research, our aim is to evaluate the performance of different generative models and adversarial defense techniques in improving the robustness and reliability of Re-ID and classification systems under various types of attacks.

8.2 Summary

In this thesis, four novel algorithms for generative and discriminative learning are proposed, namely IPES-GAN, WMA, MP-SNN, ZO-AE-RUDS. In the first approach, we proposed a novel person image generation network for the person Re-ID task in Chapter 3. The proposed IPES-GAN encodes the pose information to incorporate the global structure of the target image. The target background and camera style are captured via environment encoder and camera style loss. The cycle consistency and adversarial loss further optimize the target image generation process. In order to obtain discriminative features for Re-ID, we apply soft cross-entropy loss and KL divergence loss.

The generative models learned in Chapter 3 provide robustness to some extent. However, the performance of these DNNs decreases drastically under the influence of adversarial perturbations. This motivated us to introduce a threat model based on

Wasserstein distance for person Re-ID task (WMA in Chapter 4). This perturbation is based on Wasserstein ball which is different from previous Re-ID works with l_∞ or more general l_p perturbation. WMA does not require training to learn the noise. It generates the adversarial samples by adding noise to the clean image, and projecting the perturbations in Wasserstein ball. It uses Euclidean distance as metric loss to increase the distance between features of perturbed query images from the gallery database.

In order to design defense mechanism against the adversarial attacks, we designed a meta perturbed framework to defend Re-ID models against adversarial attacks in Chapter 5. We leverage the adversarial robustness characteristics of anisotropic and isotropic noise modules along with adversarial training and efficiently learn them through our proposed meta-learning defense strategy. Our model generalizes well and is robust against various unseen attacks. Further, we derive a novel FCA loss related to anisotropic perturbations. FCA overcomes the decrease in clean performance while ensuring robust performance against adversarial attacks in Chapter 4.

We further introduced black-box certified defense methods against adversarial perturbations have been recently investigated in the black-box setting with a zeroth-order (ZO) perspective in Chapter 6). We proposed a robust UNet denoiser (RDUNet) that ensures the robustness of black-box models trained on high-dimensional datasets. We proposed a novel black-box denoised smoothing (DS) defense mechanism, ZO-RUDS, by prepending our RDUNet to the black-box model, ensuring black-box defense. We further proposed ZO-AE-RUDS in which RDUNet followed by autoencoder (AE) is prepended to the black-box model.

The research has yielded three accepted publications and one submitted paper.

1. Verma, Astha, A. V. Subramanyam, Zheng Wang, Shin'ichi Satoh, and Rajiv Ratn Shah. "Unsupervised domain adaptation for person re-identification via individual-preserving and environmental-switching cyclic generation." IEEE Transactions on Multimedia (2021).

2. Verma, Astha, A. V. Subramanyam, and Rajiv Ratn Shah. "Wasserstein Metric Attack on Person Re-ID." 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2022.

3. Meta Perturbed Re-Id Defense, in IEEE International Conference on Multimedia and Expo, 2023. (Accepted)

4. Certified Zeroth-order Black-Box Defense with Robust UNet Denoiser, International Conference on Computer Vision, 2023. (Submitted)

8.3 Future Work

The objective of this research is to develop algorithms for generative learning and adversarial learning for Re-Id and classification tasks. Generative learning has the potential to revolutionize Re-ID and classification tasks in computer vision. However, adversarial attacks pose a significant threat to the performance and security of these models. To address these challenges, researchers are developing new techniques for generating and defending against adversarial attacks. The future of generative learning and adversarial attack and defense in Re-ID and classification tasks will depend on continued innovation and collaboration among researchers and practitioners, with a focus on improving the accuracy and robustness of these models while maintaining security and privacy for individuals and communities.

8.3.1 Unsupervised Domain Adaptation for Video Re-ID

We proposed GAN-based unsupervised domain adaptation in Chapter 3. GANs have become a popular approach for generating realistic images and videos, but they also have several limitations. One major limitation is that GANs can suffer from mode collapse, where the generator learns to produce only a limited set of samples that fail to capture the diversity of the true data distribution. Additionally, GANs require large amounts of data to train and can be sensitive to hyperparameters. Finally, there are concerns about the interpretability and transparency of GANs, as they do not provide clear explanations for how they generate their outputs (Goodfellow *et al.*, 2014*a*; Bau *et al.*, 2019; Karras *et al.*, 2020).

In order to overcome the above limitations, we plan to investigate unsupervised adversarial domain adaptation with similarity diffusion (Zhang *et al.*, 2020*b*; Xu *et al.*, 2021*b*) for video networks. It is a technique that can be applied to video Re-ID, where the goal is to recognize individuals across different cameras or environments. In this technique, a pre-trained video Re-ID model is adapted to a new target domain using adversarial training and similarity diffusion, without using any labeled data from the target domain.

Here are the main steps involved in unsupervised adversarial domain adaptation with similarity diffusion for video Re-ID:

- Pre-train a video Re-ID model on a large dataset, such as the MARS (Zheng *et al.*, 2016*a*) or DukeMTMC-VideoReID (Wu *et al.*, 2018*a*) datasets. The pre-training process involves training the model to recognize individuals across different cameras and environments.

- Select a small amount of unlabeled videos from the target domain, such as surveillance footage from a new location or environment.

- Use a generator network similar to our proposed IPES-GAN in Chapter 3 to create synthetic videos that are similar to the target domain. The generator network takes in the pre-trained model and a set of random noise vectors as input, and outputs a set of synthetic videos that are similar to the target domain. The generator network is trained using adversarial training to create videos that are difficult for a discriminator network to distinguish from real videos from the target domain.

- Use a diffusion model (Wang *et al.*, 2021*b*; Luo *et al.*, 2022) to propagate similarity information between the synthetic videos and the real videos from the target domain. The diffusion model takes in the synthetic videos and the real videos, and outputs a set of similarity scores that indicate how similar each pair of videos is to each other. The similarity scores are used to update the weights of the pre-trained model, such that the model becomes more adept at recognizing individuals in the target domain.

- Evaluate the performance of the adapted model on a set of labeled videos from the target domain. The adapted model should be able to recognize individuals in the target domain with high accuracy, despite not being explicitly trained on labeled data from that domain.

8.3.2 Future of Adversarial Attacks in Wasserstein Space

Inspired from our work in Chapter 4, we plan to develop robust adversarial attack in Wasserstein space. Here are some potential future directions for robust adversarial attack in Wasserstein space.

- Adversarial attacks are often developed for specific models and datasets, and may not transfer well to other models or datasets (Papernot *et al.*, 2016*a*; Ilyas *et al.*, 2019; Wang *et al.*, 2021*a*). This limits the practical applicability of adversarial attacks. One potential future direction is to develop more transferable Wasserstein attacks, such as those that target specific features or properties of the model that are likely to be shared across different models.

- Many existing adversarial attacks rely on small perturbations to the input data, which may not reflect the full range of possible manipulations (Hosseini and Poovendran, 2019; Qin *et al.*, 2019; Xiao *et al.*, 2021). This limits the ability to test the robustness of machine learning models to more complex and diverse attacks. Another future direction can be the development of adversarial attacks in Wasserstein space with the help of generative models that can generate more diverse and complex adversarial examples.

8.3.3 Robustness to Adaptive Attacks

One of the main limitations of our existing adversarial defense technique proposed in Chapter 5 is its susceptibility to adaptive attacks (Athalye *et al.*, 2018*a*; Tramer *et al.*, 2020), where the attacker modifies their attack strategy to bypass the defense mechanism. Many existing defenses rely on the assumption that the attacker's goal is to produce the smallest possible perturbation, but this assumption may not hold for adaptive attackers that can generate more powerful attacks that can overcome the defense mechanism.

One potential future direction to mitigate this limitation is to develop more robust defenses that can withstand a broader range of attack strategies. This can involve exploring more advanced defense mechanisms, such as ensemble models (Tramèr *et al.*, 2017; AprilPyone and Kiya, 2020) or adversarial training with diverse attacks (Jang *et al.*, 2019; Jia *et al.*, 2022). Another approach can be to incorporate more randomness into the training process, which can make the model more resistant to attacks that rely on deterministic gradients. In chapter 5, we add noise to feature vectors, similarly in order to make our method robust against adaptive attacks, we can incorporate randomness in other ways as mentioned below:

- Adding a small perturbation to the input data and minimizing the difference between the probability distributions of the original and perturbed output (Miyato *et al.*, 2018; Li *et al.*, 2021*c*). Minimization in the probability distributions can be achieved by KL divergence (Goldberger *et al.*, 2003) or Maximum Mean Discrepancy (MMD) (Gretton *et al.*, 2012) loss. It can improve the robustness of the model against adaptive adversarial attacks.
- Adding random noise to the input data where the output is the most common class predicted by the target model over multiple random perturbations, which can improve the robustness of the model against adversarial attacks (Cohen *et al.*, 2019; Salman *et al.*, 2022).

LIST OF PAPERS BASED ON THESIS

1. Verma, Astha, A. V. Subramanyam, Zheng Wang, Shin'ichi Satoh, and Rajiv Ratn Shah. "Unsupervised domain adaptation for person re-identification via individual-preserving and environmental-switching cyclic generation." IEEE Transactions on Multimedia (2021).

2. Verma, Astha, A. V. Subramanyam, and Rajiv Ratn Shah. "Wasserstein Metric Attack on Person Re-identification." 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2022.

3. Meta Perturbed Re-Id Defense, in IEEE International Conference on Multimedia and Expo, 2023. (Accepted)

4. Certified Zeroth-order Black-Box Defense with Robust UNet Denoiser, Submitted in International Conference on Computer Vision, 2023. (Under Review)

REFERENCES

1. **Addepalli, S.**, **V. BS**, **A. Baburaj**, **G. Sriramanan**, and **R. V. Babu**, Towards achieving adversarial robustness by enforcing feature consistency across bit planes. *In Proceedings of the CVPR*. 2020.

2. **Addepalli, S.**, **S. Jain**, **G. Sriramanan**, and **R. V. Babu**, Boosting adversarial robustness using feature level stochastic smoothing. *In Proceedings of the CVPR*. 2021.

3. **Addepalli, S.**, **S. Jain**, *et al.* (2022). Efficient and effective augmentation strategy for adversarial training. *Advances in Neural Information Processing Systems*, **35**, 1488–1501.

4. **Akhtar, N.**, **A. Mian**, **N. Kardan**, and **M. Shah** (2021). Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, **9**, 155161–155196.

5. **Amari, S.-i.** (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, **5**(4-5), 185–196.

6. **Andriushchenko, M.**, **F. Croce**, **N. Flammarion**, and **M. Hein**, Square attack: a query-efficient black-box adversarial attack via random search. *In ECCV*. Springer, 2020.

7. **Antun, V.**, **F. Renna**, **C. Poon**, **B. Adcock**, and **A. C. Hansen** (2020). On instabilities of deep learning in image reconstruction and the potential costs of ai. *PNAS*, **117**(48), 30088–30095.

8. **AprilPyone, M.** and **H. Kiya** (2020). Ensemble of models trained by key-based transformed images for adversarially robust defense against black-box attacks. *arXiv preprint arXiv:2011.07697*.

9. **Athalye, A.**, **N. Carlini**, and **D. Wagner**, Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *In ICML*. PMLR, 2018*a*.

10. **Athalye, A.**, **L. Engstrom**, **A. Ilyas**, and **K. Kwok**, Synthesizing robust adversarial examples. *In ICML*. PMLR, 2018*b*.

11. **Bai, J.**, **B. Chen**, **Y. Li**, **D. Wu**, **W. Guo**, **S.-t. Xia**, and **E.-h. Yang**, Targeted attack for deep hashing based retrieval. *In ECCV*. Springer, 2020*a*.

12. **Bai, S.**, **X. Bai**, and **Q. Tian**, Scalable person re-identification on supervised smoothed manifold. *In CVPR*. 2017.

13. **Bai, S.**, **X. Bai**, and **Q. Tian**, Adversarial deep domain adaptation for person re-identification. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

14. **Bai, S.**, **Y. Li**, **Y. Zhou**, **Q. Li**, and **P. H. Torr** (2020*b*). Adversarial metric attack and defense for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**(6), 2119–2126.

15. **Bai, T.**, **J. Luo**, **J. Zhao**, **B. Wen**, and **Q. Wang** (2021*a*). Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*.

16. **Bai, Y.**, **J. Jiao**, **W. Ce**, **J. Liu**, **Y. Lou**, **X. Feng**, and **L.-Y. Duan**, Person30k: A dual-meta generalization network for person re-identification. *In CVPR*. 2021*b*.

17. **Bai, Y.**, **C. Wang**, **Y. Lou**, **J. Liu**, and **L.-Y. Duan** (2021*c*). Hierarchical connectivity-centered clustering for unsupervised domain adaptation on person re-identification. *IEEE Transactions on Image Processing*, **30**, 6715–6729.

18. **Bak, S.**, **P. Carr**, and **J.-F. Lalonde** (2018). Domain adaptation through synthesis for unsupervised person re-identification. *arXiv preprint arXiv:1804.10094*.

19. **Ballard, D. H.** (1981). Generalizing the hough transform to detect arbitrary shapes. *Pattern recognition*, **13**(2), 111–122.

20. **Bartler, A.**, **A. Bühler**, **F. Wiewel**, **M. Döbler**, and **B. Yang**, Mt3: Meta test-time training for self-supervised test-time adaption. *In AISTATS*. PMLR, 2022.

21. **Bau, D.**, **J.-Y. Zhu**, **H. Liu**, and **L. Torresani**, Understanding and improving interpretable gans. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

22. **Bay, H.**, **T. Tuytelaars**, and **L. Van Gool** (2006). Surf: Speeded up robust features. *Lecture notes in computer science*, **3951**, 404–417.

23. **Baytaş, İ. M.** and **D. Deb** (2023). Robustness-via-synthesis: Robust training with generative adversarial perturbations. *Neurocomputing*, **516**, 49–60.

24. **Bouniot, Q.**, **R. Audigier**, and **A. Loesch**, Vulnerability of person re-identification models to metric adversarial attacks. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020.

25. **Bousmalis, K.**, **N. Silberman**, **D. Dohan**, **D. Erhan**, and **D. Krishnan**, Unsupervised pixel-level domain adaptation with generative adversarial networks. *In CVPR*. 2017.

26. **Brock, A.**, **J. Donahue**, and **K. Simonyan** (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

27. **Brown, T. B.**, **D. Mané**, **A. Roy**, **M. Abadi**, and **J. Gilmer** (2017). Adversarial patch. *arXiv preprint arXiv:1712.09665*.

28. **Bunel, R. R.**, **I. Turkaslan**, **P. Torr**, **P. Kohli**, and **P. K. Mudigonda** (2018). A unified view of piecewise linear neural network verification. *Advances in Neural Information Processing Systems*, **31**.

29. **Byun, J.**, **H. Go**, and **C. Kim**, On the effectiveness of small input noise for defending against query-based black-box attacks. *In CVPR*. 2022.

30. **Cai, H.**, **Y. Lou**, **D. McKenzie**, and **W. Yin**, A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization. *In ICML*. PMLR, 2021.

31. **Cai, H.**, **D. Mckenzie**, **W. Yin**, and **Z. Zhang** (2022). Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling. *SIAM Journal on Optimization*, **32**(2), 687–714.

32. **Cai, M.**, **Z. Li**, and **Y. Li**, Unsupervised person re-identification via soft multilabel learning. *In Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

33. **Cao, Z.**, **G. Hidalgo**, **T. Simon**, **S.-E. Wei**, and **Y. Sheikh** (2018). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*.

34. **Carlini, N.** and **D. Wagner**, Towards evaluating the robustness of neural networks. *In (sp)*. IEEE, 2017.

35. **Carmon, Y.**, **A. Raghunathan**, **L. Schmidt**, **J. C. Duchi**, and **P. S. Liang** (2019). Unlabeled data improves adversarial robustness. *Adv. neural inf. process. syst*, **32**.

36. **Chan, A.**, **Y. Tay**, **Y. S. Ong**, and **J. Fu** (2019). Jacobian adversarially regularized networks for robustness. *arXiv preprint arXiv:1912.10185*.

37. **Chen, C.**, **J. Li**, **X. Han**, **X. Liu**, and **Y. Yu**, Compound domain generalization via meta-knowledge encoding. *In CVPR*. 2022*a*.

38. **Chen, J.**, **Y. Cheng**, **Z. Gan**, **Q. Gu**, and **J. Liu**, Efficient robust training via backward smoothing. *In AAAI*, volume 36. 2022*b*.

39. **Chen, J.** and **Q. Gu**, Rays: A ray searching method for hard-label adversarial attack. *In Proceedings of the 26th ACM SIGKDD*. 2020.

40. **Chen, J.**, **M. I. Jordan**, and **M. J. Wainwright**, Hopskipjumpattack: A query-efficient decision-based attack. *In 2020 ieee symposium on security and privacy (sp)*. IEEE, 2020*a*.

41. **Chen, L.-C.**, **Y. Zhu**, **G. Papandreou**, **F. Schroff**, and **H. Adam**, Encoder-decoder with atrous separable convolution for semantic image segmentation. *In ECCV*. 2018.

42. **Chen, P.-Y.**, **H. Zhang**, **Y. Sharma**, **J. Yi**, and **C.-J. Hsieh**, Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. *In Proceedings of the 10th AISEC*. 2017.

43. **Chen, T.**, **S. Kornblith**, **M. Norouzi**, and **G. Hinton**, A simple framework for contrastive learning of visual representations. *In International Conference on Machine Learning*. PMLR, 2020*b*.

44. **Chen, T.**, **S. Liu**, **S. Chang**, **Y. Cheng**, **L. Amini**, and **Z. Wang**, Adversarial robustness: From self-supervised pre-training to fine-tuning. *In Proceedings of the CVPR*. 2020*c*.

45. **Chen, Y.**, **X. Zhu**, and **S. Gong**, Instance-guided context rendering for cross-domain person re-identification. *In ICCV*. 2019.

46. **Chen, Z.**, **X. Luo**, **Y. Hu**, and **Y. He**, Domain-aware generative adversarial networks. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35. 2021.

47. **Cheng, D.**, **J. Li**, **Q. Kou**, **K. Zhao**, and **R. Liu** (2022). H-net: Unsupervised domain adaptation person re-identification network based on hierarchy. *Image and Vision Computing*, **124**, 104493.

48. **Cheng, Z.**, **Y. Cao**, and **S. Gong**, Revisiting temporal modeling for video-based person re-identification. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

49. **Cheng, Z.**, **F. Zhu**, **X.-Y. Zhang**, and **C.-L. Liu** (2023). Adversarial training with distribution normalization and margin balance. *Pattern Recognition*, **136**, 109182.

50. **Choi, S.**, **T. Kim**, **M. Jeong**, **H. Park**, and **C. Kim**, Meta batch-instance normalization for generalizable person re-identification. *In CVPR*. 2021.

51. **Choi, Y.**, **M. Choi**, **M. Kim**, **J.-W. Ha**, **S. Kim**, and **J. Choo**, Stargan v2: Diverse image synthesis for multiple domains. *In Computer Vision and Pattern Recognition*. 2020.

52. **Coates, A.**, **A. Y. Ng**, and **H. Lee** (2011). The stl-10 dataset. *https://cs.stanford.edu/ acoates/stl10/*.

53. **Cocchiarella, N.** (1977). Sortals, natural kinds and re-identification. *Logique et analyse*, **20**(80), 439–474.

54. **Cohen, J.**, **E. Rosenfeld**, and **Z. Kolter**, Certified adversarial robustness via randomized smoothing. *In ICML*. PMLR, 2019.

55. **Coifman, B.** (1998). Vehicle re-identification and travel time measurement in real-time on freeways using existing loop detector infrastructure. *Transportation Research Record*, **1643**(1), 181–191.

56. **Crevier, D.** and **R. Lepage** (1997). Knowledge-based image understanding systems: A survey. *Computer vision and image understanding*, **67**(2), 161–185.

57. **Croce, F.** and **M. Hein**, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *In Proceedings of the 37th ICML*, ICML'20. JMLR.org, 2020.

58. **Cui, J.**, **S. Liu**, **L. Wang**, and **J. Jia**, Learnable boundary guided adversarial training. *In Proceedings of the ICCV*. 2021.

59. **Dalal, N.** and **B. Triggs**, Histograms of oriented gradients for human detection. *In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1. IEEE, 2005.

60. **de Jorge Aranda, P.**, **A. Bibi**, **R. Volpi**, **A. Sanyal**, **P. Torr**, **G. Rogez**, and **P. Dokania** (2022). Make some noise: Reliable and efficient single-step adversarial training. *Advances in Neural Information Processing Systems*, **35**, 12881–12893.

61. **Deb, D.**, **X. Liu**, and **A. K. Jain**, Faceguard: A self-supervised defense against adversarial face images. *In 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023.

62. **Deng, W.**, **L. Zheng**, **Q. Ye**, **G. Kang**, **Y. Yang**, and **J. Jiao**, Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

63. **Devaguptapu, C.**, **D. Agarwal**, **G. Mittal**, **P. Gopalani**, and **V. N. Balasubramanian**, On adversarial robustness: A neural architecture search perspective. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

64. **Dhillon, G. S.**, **K. Azizzadenesheli**, **A. Ravichandran**, **A. Farhadi**, and **O. O. Koyejo**, Stochastic activation pruning for robust adversarial defense. *In Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

65. **Ding, K.**, **Z. Liu**, **S. Han**, **G. Ateniese**, and **H. Yang** (2019). Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*.

66. **Ding, W.**, **X. Wei**, **R. Ji**, **X. Hong**, **Q. Tian**, and **Y. Gong** (2021). Beyond universal person re-identification attack. *IEEE transactions on information forensics and security*, **16**, 3442–3455.

67. **Ding, Y.**, **H. Fan**, **M. Xu**, and **Y. Yang** (2020). Adaptive exploration for unsupervised person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, **16**(1), 1–19.

68. **Dinh, L.**, **D. Krueger**, and **Y. Bengio** (2014). Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

69. **Donahue, J.**, **S. Singh**, **D. Gurari**, and **N. Sünderhauf** (2019). Large scale adversarial representation learning. *arXiv preprint arXiv:1907.02544*.

70. **Dong, H.**, **X. Liang**, **K. Gong**, **H. Lai**, **J. Zhu**, and **J. Yin**, Soft-gated warping-gan for pose-guided person image synthesis. *In NeurIPS*. 2018*a*.

71. **Dong, J.**, **Y. Wang**, **J.-H. Lai**, and **X. Xie**, Improving adversarially robust few-shot image classification with generalizable representations. *In CVPR*. 2022.

72. **Dong, X.**, **J. Huang**, and **S. Yang**, Domain guided dropout for person re-identification. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

73. **Dong, Y.**, **F. Liao**, **T. Pang**, **H. Su**, **J. Zhu**, **X. Hu**, and **J. Li**, Boosting adversarial attacks with momentum. *In CVPR*. 2018*b*.

74. **Dumoulin, V.**, **F. Visin**, **M. A. G. Mazzeo**, **M. Matteucci**, and **Y. N. Dauphin** (2016). A guide to convolution arithmetic for deep learning. *CoRR*, **abs/1603.07285**. URL `http://arxiv.org/abs/1603.07285`.

75. **Dutta, S.**, **S. Jha**, **S. Sanakaranarayanan**, and **A. Tiwari** (2017). Output range analysis for deep neural networks. *arXiv preprint arXiv:1709.09130*.

76. **Elaalami, I. A.**, **S. O. Olatunji**, and **R. M. Zagrouba** (2022). At-bod: An adversarial attack on fool dnn-based blackbox object detection models. *Applied Sciences*, **12**(4), 2003.

77. **Esser, P.**, **R. Rombach**, and **B. Ommer**, Taming transformers for high-resolution image synthesis. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

78. **Eustratiadis, P.**, **H. Gouk**, **D. Li**, and **T. Hospedales**, Weight-covariance alignment for adversarially robust neural networks. *In ICML*. PMLR, 2021.

79. **Eykholt, K.**, **I. Evtimov**, **E. Fernandes**, **B. Li**, **A. Rahmati**, **C. Xiao**, **A. Prakash**, **T. Kohno**, and **D. Song**, Robust physical-world attacks on deep learning visual classification. *In Proceedings of the CVPR*. 2018.

80. **Fan, H.**, **L. Zheng**, **C. Yan**, and **Y. Yang** (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, **14**(4), 1–18.

81. **Fan, H.**, **L. Zheng**, and **Y. Yang** (2017). Unsupervised person re-identification: Clustering and fine-tuning. *arXiv preprint arXiv:1705.10444*.

82. **Fang, S.**, **J. Li**, **X. Lin**, and **R. Ji**, Learning to learn transferable attack. *In AAAI*, volume 36. 2022.

83. **Farenzena, M.**, **L. Bazzani**, **A. Perina**, **V. Murino**, and **M. Cristani**, Person re-identification by symmetry-driven accumulation of local features. *In CVPR*. 2010.

84. **Feng, H.**, **M. Chen**, **J. Hu**, **D. Shen**, **H. Liu**, and **D. Cai** (2021*a*). Complementary pseudo labels for unsupervised domain adaptation on person re-identification. *IEEE Transactions on Image Processing*, **30**, 2898–2907.

85. **Feng, W.**, **B. Wu**, **T. Zhang**, **Y. Zhang**, and **Y. Zhang**, Meta-attack: Class-agnostic and model-agnostic physical adversarial attack. *In CVPR*. 2021*b*.

86. **Feng, Y.**, **B. Chen**, **T. Dai**, and **S.-T. Xia**, Adversarial attack on deep product quantization network for image retrieval. *In AAAI*, volume 34. 2020.

87. **Finn, C.**, **P. Abbeel**, and **S. Levine**, Model-agnostic meta-learning for fast adaptation of deep networks. *In ICML*. PMLR, 2017.

88. **Folz, J.**, **S. Palacio**, **J. Hees**, and **A. Dengel**, Adversarial defense based on structure-to-signal autoencoders. *In 2020 WACV*. IEEE, 2020.

89. **Fredrikson, M.**, **S. Jha**, and **T. Ristenpart**, Model inversion attacks that exploit confidence information and basic countermeasures. *In Proceedings of the 22nd ACM CCS*. 2015.

90. **Fu, Y.**, **Y. Wei**, **G. Wang**, **Y. Zhou**, **H. Shi**, and **T. S. Huang**, Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. *In proceedings of the IEEE/CVF international conference on computer vision*. 2019.

91. **Ganin, Y.** and **V. Lempitsky**, Unsupervised domain adaptation by backpropagation. *In International conference on machine learning*. PMLR, 2015.

92. **Ganin, Y.**, **E. Ustinova**, **H. Ajakan**, **P. Germain**, **H. Larochelle**, **F. Laviolette**, **M. Marchand**, and **V. Lempitsky** (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, **17**(1), 2096–2030.

93. **Gatys, L. A.**, **A. S. Ecker**, and **M. Bethge**, Image style transfer using convolutional neural networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

94. **Ge, P.**, **C.-X. Ren**, **X.-L. Xu**, and **H. Yan** (2023). Unsupervised domain adaptation via deep conditional adaptation network. *Pattern Recognition*, **134**, 109088.

95. **Ge, Y.**, **D. Chen**, and **H. Li** (2020). Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*.

96. **Ge, Y.**, **Z. Li**, **H. Zhao**, **G. Yin**, **S. Yi**, **X. Wang**, *et al.*, Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *In NeurIPS*. 2018.

97. **Ge, Y.**, **F. Zhu**, **D. Chen**, **R. Zhao**, **X. Wang**, and **H. Li** (2022). Structured domain adaptation with online relation regularization for unsupervised person re-id. *IEEE Transactions on Neural Networks and Learning Systems*.

98. **Girshick, R.**, **J. Donahue**, **T. Darrell**, and **J. Malik**, Rich feature hierarchies for accurate object detection and semantic segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

99. **Goldberger, J.**, **S. Gordon**, **H. Greenspan**, *et al.*, An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. *In ICCV*, volume 3. 2003.

100. **Goldblum, M.**, **L. Fowl**, and **T. Goldstein** (2020). Adversarially robust few-shot learning: A meta-learning approach. *NIPS*, **33**, 17886–17895.

101. **Gong, Y.**, **L. Huang**, and **L. Chen**, Person re-identification method based on color attack and joint defence. *In CVPR*. 2022.

102. **Gong, Z.**, **T. Liu**, **X. Wang**, and **D. Tao** (2021). Understanding the difficulty of adversarial robustness via landscape analysis. *IEEE Transactions on Neural Networks and Learning Systems*.

103. **Goodfellow, I.**, **J. Pouget-Abadie**, **M. Mirza**, **B. Xu**, **D. Warde-Farley**, **S. Ozair**, **A. Courville**, and **Y. Bengio**, Generative adversarial nets. *In Advances in neural information processing systems*. 2014*a*.

104. **Goodfellow, I.**, **J. Pouget-Abadie**, **M. Mirza**, **B. Xu**, **D. Warde-Farley**, **S. Ozair**, **A. Courville**, and **Y. Bengio** (2020). Generative adversarial networks. *Communications of the ACM*, **63**(11), 139–144.

105. **Goodfellow, I. J.**, **J. Shlens**, and **C. Szegedy** (2014*b*). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

106. **Gray, D.** and **H. Tao** (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. *Computer Vision–ECCV 2008*, 262–275.

107. **Gregor, K.**, **I. Danihelka**, **A. Graves**, **D. J. Rezende**, and **D. Wierstra**, Draw: A recurrent neural network for image generation. *In Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 2015.

108. **Gretton, A.**, **K. M. Borgwardt**, **M. J. Rasch**, **B. Schölkopf**, and **A. Smola** (2012). A kernel two-sample test. *Journal of Machine Learning Research*, **13**(Mar), 723–773.

109. **Gu, X.**, **H. Chang**, **B. Ma**, **S. Bai**, **S. Shan**, and **X. Chen**, Clothes-changing person re-identification with rgb modality only. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

110. **Guo, C.**, **Y. Li**, **P. Luo**, **W. Huang**, and **L. Zhang** (2021). Adversarial metric learning: Enabling robustness of person re-identification models against adversarial attacks. *IEEE Transactions on Information Forensics and Security*, **16**, 1807–1822.

111. **Guo, J.**, **X. Zhu**, **C. Zhao**, **D. Cao**, **Z. Lei**, and **S. Z. Li**, Learning meta face recognition in unseen domains. *In CVPR*. 2020.

112. **Gupta, R.**, **R. Stanforth**, and **P. Hennig**, Verifiable robustness of neural networks with contractive activation functions. *In Proceedings of the 38th ICML*. PMLR, 2021.

113. **Han, T.**, **Y. Lu**, **S.-C. Zhu**, and **Y. N. Wu**, Alternating back-propagation for generator network. *In AAAI*, volume 31. 2017.

114. **He, B.**, **J. Li**, **Y. Zhao**, and **Y. Tian**, Part-regularized near-duplicate vehicle re-identification. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019*a*.

115. **He, K.**, **X. Zhang**, **S. Ren**, and **J. Sun**, Deep residual learning for image recognition. *In CVPR*. 2016.

116. **He, Z.**, **A. S. Rakin**, and **D. Fan**, Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. *In CVPR*. 2019*b*.

117. **Hermans, A.**, **L. Beyer**, and **B. Leibe** (2017). In defense of the triplet loss for person re-identification. *CoRR*.

118. **Heusel, M.**, **H. Ramsauer**, **T. Unterthiner**, **B. Nessler**, and **S. Hochreiter**, Gans trained by a two time-scale update rule converge to a local nash equilibrium. *In NeurIPS*. 2017.

119. **Hinton, G.**, **O. Vinyals**, **J. Dean**, *et al.* (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, **2**(7).

120. **Hirzer, M.**, **P. M. Roth**, and **H. Bischof**, Person re-identification by efficient impostor-based metric learning. *In 2012 IEEE ninth international conference on advanced video and signal-based surveillance*. IEEE, 2012.

121. **Ho, J.**, **A. Jain**, and **P. Abbeel** (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, **33**, 6840–6851.

122. **Hoffman, J.**, **E. Tzeng**, **T. Park**, **J.-Y. Zhu**, **P. Isola**, **K. Saenko**, **A. A. Efros**, and **T. Darrell** (2017). Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*.

123. **Hore, A.** and **D. Ziou**, Image quality metrics: Psnr vs. ssim. *In 2010 20th ICPR*. IEEE, 2010.

124. **Hosseini, H.** and **R. Poovendran** (2019). Semantic adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, **30**(9), 2805–2824.

125. **Hu, J. E.**, **A. Swaminathan**, **H. Salman**, and **G. Yang** (2020). Improved image wasserstein attacks and defenses. *arXiv preprint arXiv:2004.12478*.

126. **Huang, G.**, **Z. Liu**, **L. Van Der Maaten**, and **K. Q. Weinberger**, Densely connected convolutional networks. *In CVPR*. 2017.

127. **Huang, T.** and **S. Russell**, Object identification in a bayesian context. *In IJCAI*, volume 97. Citeseer, 1997.

128. **Huang, X.**, **M.-Y. Liu**, **S. Belongie**, and **J. Kautz**, Multimodal unsupervised image-to-image translation. *In ECCV*. 2018.

129. **Huang, Y.**, **Q. Wu**, **J. Xu**, and **Y. Zhong**, Sbsgan: Suppression of inter-domain background shift for person re-identification. *In ICCV*. 2019*a*.

130. **Huang, Y.**, **Z.-J. Zha**, **X. Fu**, and **W. Zhang**, Illumination-invariant person re-identification. *In ACMMM*. 2019*b*.

131. **Huang, Z.** and **T. Zhang**, Black-box adversarial attack with transferable model-based embedding. *In ICLR*. 2020.

132. **Huang, Z.**, **Z. Zhang**, **C. Lan**, **W. Zeng**, **P. Chu**, **Q. You**, **J. Wang**, **Z. Liu**, and **Z.-j. Zha**, Lifelong unsupervised domain adaptive person re-identification with coordinated anti-forgetting and adaptation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

133. **Hussain, M.**, **D. Koundal**, and **J. Manhas** (2023). Deep learning-based diagnosis of disc degenerative diseases using mri: A comprehensive review. *Computers and Electrical Engineering*, **105**, 108524.

134. **Ilyas, A.**, **L. Engstrom**, **A. Athalye**, and **J. Lin**, Black-box adversarial attacks with limited queries and information. *In ICML*. PMLR, 2018*a*.

135. **Ilyas, A.**, **L. Engstrom**, **A. Athalye**, **J. Lin**, and **S. Feizi** (2019). Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*.

136. **Ilyas, A.**, **L. Engstrom**, and **A. Madry** (2018*b*). Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*.

137. **Ioffe, S.** and **C. Szegedy**, Batch normalization: Accelerating deep network training by reducing internal covariate shift. *In ICML*. PMLR, 2015.

138. **Isola, P.**, **J.-Y. Zhu**, **T. Zhou**, and **A. A. Efros**, Image-to-image translation with conditional adversarial networks. *In* . 2017.

139. **Jang, Y.**, **T. Zhao**, **S. Hong**, and **H. Lee**, Adversarial defense via learning to generate diverse attacks. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

140. **Jeddi, A.**, **M. J. Shafiee**, **M. Karg**, **C. Scharfenberger**, and **A. Wong**, Learn2perturb: an end-to-end feature perturbation learning to improve adversarial robustness. *In CVPR*. 2020.

141. **Jia, X.**, **Y. Zhang**, **B. Wu**, **K. Ma**, **J. Wang**, and **X. Cao**, Las-at: adversarial training with learnable attack strategy. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

142. **Jian, T.**, **Z. Wang**, **Y. Wang**, **J. Dy**, and **S. Ioannidis** (2022). Pruning adversarially robust neural networks without adversarial examples. *arXiv preprint arXiv:2210.04311*.

143. **Jin, G.**, **X. Yi**, **W. Huang**, **S. Schewe**, and **X. Huang**, Enhancing adversarial training with second-order statistics of weights. *In CVPR*. 2022.

144. **Joshi, C. K.**, **C. Yang**, and **S. Wang**, Certified robustness of graph neural networks against adversarial attacks. *In Proceedings of the AIES*, volume 36. 2022.

145. **Karras, T.**, **T. Aila**, **S. Laine**, and **J. Lehtinen** (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

146. **Karras, T.**, **S. Laine**, **M. Aittala**, **J. Hellsten**, **J. Lehtinen**, and **T. Aila**, Analyzing and improving the image quality of stylegan. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

147. **Katz, G.**, **C. Barrett**, **D. L. Dill**, **K. Julian**, and **M. J. Kochenderfer**, Reluplex: An efficient smt solver for verifying deep neural networks. *In CAV*. Springer, 2017.

148. **Katz, G.**, **C. Barrett**, **D. L. Dill**, **K. Julian**, and **M. J. Kochenderfer** (2022). Formal verification of neural networks: From verification of robustness to certificates. *IEEE Trans Neural Netw Learn Syst*.

149. **Khan, S. U.**, **T. Hussain**, **A. Ullah**, and **S. W. Baik** (2021). Deep-reid: Deep features and autoencoder assisted image patching strategy for person re-identification in smart cities surveillance. *Multimedia Tools and Applications*, 1–22.

150. **Khatun, A.**, **S. Denman**, **S. Sridharan**, and **C. Fookes** (2020). End-to-end domain adaptive attention network for cross-domain person re-identification. *arXiv preprint arXiv:2005.03222*.

151. **Kim, J.**, **B.-S. Hua**, **T. Nguyen**, and **S.-K. Yeung**, Minimal adversarial examples for deep learning on 3d point clouds. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

152. **Kinfu, K. A.** and **R. Vidal**, Analysis and extensions of adversarial training for video classification. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

153. **Kingma, D.**, **T. Salimans**, **B. Poole**, and **J. Ho** (2021). Variational diffusion models. *Advances in neural information processing systems*, **34**, 21696–21707.

154. **Kingma, D. P.** and **P. Dhariwal**, Glow: Generative flow with invertible 1x1 convolutions. *In Advances in Neural Information Processing Systems*. 2018.

155. **Kingma, D. P.** and **M. Welling** (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

156. **Kolesnikov, A.**, **X. Zhai**, **L. Beyer**, **J. Puigcerver**, **J. Yung**, and **S. Gelly**, Revisiting self-supervised visual representation learning. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

157. **Krizhevsky, A.** and **G. Hinton** (2009*a*). Cifar-10 (canadian institute for advanced research). *https://www.cs.toronto.edu/ kriz/cifar.html*.

158. **Krizhevsky, A.** and **G. Hinton** (2009*b*). Cifar-100 (canadian institute for advanced research). *https://www.cs.toronto.edu/ kriz/cifar.html*.

159. **Krizhevsky, A.**, **I. Sutskever**, and **G. E. Hinton**, Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems*. 2012.

160. **Kurakin, A.**, **I. Goodfellow**, and **S. Bengio** (2016). Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

161. **Kurakin, A.**, **I. J. Goodfellow**, and **S. Bengio**, Adversarial examples in the physical world. *In Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, 99–112.

162. **Lamdan, Y.**, **J. T. Schwartz**, and **H. J. Wolfson**, Object recognition by affine invariant matching. *In Proceedings CVPR'88: The Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 1988.

163. **LeCun, Y.**, **Y. Bengio**, and **G. Hinton** (2015). Deep learning. *Nature*, **521**(7553), 436–444.

164. **LeCun, Y.**, **L. Bottou**, **Y. Bengio**, and **P. Haffner** (1998*a*). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**(11), 2278–2324.

165. **LeCun, Y.**, **C. Cortes**, and **C. Burges** (1998*b*). The mnist database. *http://yann.lecun.com/exdb/mnist/*.

166. **Lecuyer, M.**, **V. Atlidakis**, **R. Geambasu**, **D. Hsu**, and **S. Jana**, Certified robustness to adversarial examples with differential privacy. *In (SP)*. IEEE, 2019.

167. **Lee, H.-Y.**, **H.-Y. Tseng**, **J.-B. Huang**, **M. Singh**, and **M.-H. Yang**, Diverse image-to-image translation via disentangled representations. *In ECCV)*. 2018.

168. **Li, B.**, **C. Chen**, **W. Wang**, and **L. Carin** (2019*a*). Certified adversarial robustness with additive noise. *Adv. neural inf. process. syst*, **32**.

169. **Li, H.**, **N. Dong**, **Z. Yu**, **D. Tao**, and **G. Qi** (2021*a*). Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, **32**(5), 2814–2830.

170. **Li, J.**, **R. Ji**, **H. Liu**, **X. Hong**, **Y. Gao**, and **Q. Tian**, Universal perturbation attack against image retrieval. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019*b*.

171. **Li, K.**, **Z. Ding**, **K. Li**, **Y. Zhang**, and **Y. Fu**, Support neighbor loss for person re-identification. *In ACMMM*. 2018*a*.

172. **Li, M.**, **C. Deng**, **T. Li**, **J. Yan**, **X. Gao**, and **H. Huang**, Towards transferable targeted attack. *In CVPR*. 2020.

173. **Li, M.**, **X. Zhu**, and **S. Gong**, Unsupervised person re-identification by deep learning tracklet association. *In ECCV*. 2018*b*.

174. **Li, T.**, **Y. Wu**, **S. Chen**, **K. Fang**, and **X. Huang**, Subspace adversarial training. *In CVPR*. 2022.

175. **Li, W.**, **R. Zhao**, **T. Xiao**, and **X. Wang**, Deepreid: Deep filter pairing neural network for person re-identification. *In CVPR*. 2014.

176. **Li, W.**, **X. Zhu**, and **S. Gong**, Harmonious attention network for person re-identification. *In CVPR*, volume 1. 2018*c*.

177. **Li, X.**, **J. Li**, **Y. Chen**, **S. Ye**, **Y. He**, **S. Wang**, **H. Su**, and **H. Xue**, Qair: Practical query-efficient black-box attacks for image retrieval. *In CVPR*. 2021*b*.

178. **Li, X.**, **S. Wang**, **X. Liu**, **W. Hou**, **T. Liu**, **L. Sigal**, and **A. C. Sankaranarayanan**, Adversarial pruning for reducing overfitting in deep neural networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

179. **Li, X.**, **X. Zhang**, **T. Liu**, **Z. Gao**, and **C. Zhang** (2021*c*). Dual virtual adversarial training for semi-supervised learning. *IEEE Transactions on Cybernetics*.

180. **Li, Y.**, **W. Jin**, **H. Xu**, and **J. Tang**, Deeprobust: a platform for adversarial attacks and defenses. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35. 2021*d*.

181. **Li, Y.**, **S. Ruan**, **H. Qin**, **S. Deng**, and **M. A. El-Yacoubi** (2023). Transformer based defense gan against palm-vein adversarial attacks. *IEEE Transactions on Information Forensics and Security*, **18**, 1509–1523.

182. **Li, Y.**, **Y. Yang**, **W. Zhou**, and **T. Hospedales**, Feature-critic networks for heterogeneous domain generalization. *In ICML*. PMLR, 2019*c*.

183. **Li, Y.**, **H. Yao**, **L. Duan**, **H. Yao**, and **C. Xu**, Adaptive feature fusion via graph neural network for person re-identification. *In ACMMM*. 2019*d*.

184. **Li, Y.-J.**, **C.-S. Lin**, **Y.-B. Lin**, and **Y.-C. F. Wang**, Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. *In ICCV*. 2019*e*.

185. **Lian, X.**, **H. Zhang**, **C.-J. Hsieh**, **Y. Huang**, and **J. Liu** (2016). A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. *Adv. neural inf. process. syst*, **29**.

186. **Liang, H.**, **E. He**, **Y. Zhao**, **Z. Jia**, and **H. Li** (2022). Adversarial attack and defense: A survey. *Electronics*, **11**(8), 1283.

187. **Liao, S.**, **Y. Hu**, **X. Zhu**, and **S. Z. Li**, Person re-identification by local maximal occurrence representation and metric learning. *In CVPR*. 2015.

188. **Lin, S.**, **H. Li**, **C.-T. Li**, and **A. C. Kot** (2018). Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv preprint arXiv:1807.01440*.

189. **Lin, W.-A.**, **C. P. Lau**, **A. Levine**, **R. Chellappa**, and **S. Feizi** (2020). Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks. *Advances in Neural Information Processing Systems*, **33**, 3487–3498.

190. **Lin, Y.**, **X. Dong**, **L. Zheng**, **Y. Yan**, and **Y. Yang**, A bottom-up clustering approach to unsupervised person re-identification. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33. 2019.

191. **Lipton, Z. C.**, **M. Tripathi**, and **C. Elkan** (2018). Detecting and correcting for label shift with black box predictors. *ICML*, **18**, 3156–3165.

192. **Liu, D.** and **W. Hu** (2022). Imperceptible transfer attack and defense on 3d point cloud classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

193. **Liu, D.**, **L. Wu**, **R. Hong**, **Z. Ge**, **J. Shen**, **F. Boussaid**, and **M. Bennamoun** (2023). Generative metric learning for adversarially robust open-world person re-identification. *ACM Transactions on Multimedia Computing, Communications and Applications*, **19**(1), 1–19.

194. **Liu, D.**, **R. Yu**, and **H. Su**, Extending adversarial attacks and defenses to deep 3d point cloud classifiers. *In 2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019*a*.

195. **Liu, H.** and **Y. Li**, Query-uap: Query-efficient universal adversarial perturbation for large-scale person re-identification attack. *In Pattern Recognition and Computer Vision: 5th Chinese Conference, PRCV 2022, Shenzhen, China, November 4–7, 2022, Proceedings, Part III*. Springer, 2022.

196. **Liu, J.**, **A. Levine**, **C. P. Lau**, **R. Chellappa**, and **S. Feizi**, Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

197. **Liu, J.**, **Z.-J. Zha**, **D. Chen**, **R. Hong**, and **M. Wang**, Adaptive transfer network for cross-domain person re-identification. *In CVPR*. 2019*b*.

198. **Liu, S.**, **P.-Y. Chen**, **X. Chen**, and **M. Hong**, signsgd via zeroth-order oracle. *In ICLR*. 2019*c*.

199. **Liu, S.**, **P.-Y. Chen**, **B. Kailkhura**, **G. Zhang**, **A. O. Hero III**, and **P. K. Varshney** (2020). A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, **37**(5), 43–54.

200. **Liu, S.**, **W. Chen**, **W. Liu**, **Y. Lee**, and **P. Ng**, Learning to defend against adversarial attacks. *In International Conference on Learning Representations*. 2018*a*.

201. **Liu, S.**, **B. Kailkhura**, **P.-Y. Chen**, **P. Ting**, **S. Chang**, and **L. Amini** (2018*b*). Zeroth-order stochastic variance reduction for nonconvex optimization. *Adv. neural inf. process. syst*, **31**.

202. **Liu, S.**, **S. Lu**, **X. Chen**, **Y. Feng**, **K. Xu**, **A. Al-Dujaili**, **M. Hong**, and **U.-M. O'Reilly**, Min-max optimization without gradients: Convergence and applications to adversarial ml. *In ICML*. 2020a.

203. **Liu, X.**, **M. Cheng**, **H. Zhang**, and **C.-J. Hsieh**, Towards robust neural networks via random self-ensemble. *In (ECCV)*. 2018*c*.

204. **Liu, X.**, **W. Liu**, **T. Mei**, and **H. Ma**, A deep learning-based approach to progressive vehicle re-identification for urban surveillance. *In ECCV*. Springer, 2016.

205. **Liu, X.**, **Y. Wang**, **T. Liu**, and **Z. Zhang**, Tiny imagenet visual recognition challenge. *In ACM Multimedia*. ACM, 2017.

206. **Liu, Z.**, **X. Song**, **C. Liu**, **C. Shen**, and **J. Feng**, Large-scale deep coupled convolutional networks for person re-identification. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018*d*.

207. **Liu, Z.**, **M. Sun**, **T. Zhou**, **G. Huang**, and **T. Wang** (2019*d*). Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*.

208. **Long, M.**, **Y. Cao**, **J. Wang**, and **M. Jordan**, Learning transferable features with deep adaptation networks. *In ICML*. 2015.

209. **Lowe, D. G.** (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial intelligence*, **31**(3), 355–395.

210. **Lowe, G.** (2004). Sift-the scale invariant feature transform. *Int. J*, **2**(91-110), 2.

211. **Luo, M.**, **X. Huang**, **W. Zhang**, **H. Wang**, and **L. Yang**, Multi-scale diffusion convolutional recurrent network for human motion prediction. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36. 2022.

212. **Lv, J.**, **W. Chen**, **Q. Li**, and **C. Yang**, Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. *In CVPR*. 2018.

213. **Ma, L.**, **X. Jia**, **Q. Sun**, **B. Schiele**, **T. Tuytelaars**, and **L. Van Gool**, Pose guided person image generation. *In NeurIPS*. 2017.

214. **Ma, L.**, **Q. Sun**, **S. Georgoulis**, **L. Van Gool**, **B. Schiele**, and **M. Fritz**, Disentangled person image generation. *In CVPR*. 2018.

215. **Ma, X.**, **Y. Liu**, **J. Bailey**, and **H. Shi** (2021). Towards certified robustness for deep neural networks with lipschitz continuous activations. *IEEE Trans Neural Netw Learn Syst*, **32**(2), 458–470.

216. **Madry, A.**, **A. Makelov**, **L. Schmidt**, **D. Tsipras**, and **A. Vladu** (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

217. **Madry, A.**, **A. Makelov**, **L. Schmidt**, **D. Tsipras**, and **A. Vladu**, Towards deep learning models resistant to adversarial attacks. *In International Conference on Learning Representations (ICLR)*. 2018.

218. **Mao, X.**, **Q. Li**, **H. Xie**, **R. Y. Lau**, **Z. Wang**, and **S. Paul Smolley**, Least squares generative adversarial networks. *In ICCV*. 2017.

219. **Mekhazni, D.**, **A. Bhuiyan**, **G. Ekladious**, and **E. Granger**, Unsupervised domain adaptation in the dissimilarity space for person re-identification. *In ECCV*. Springer, 2020.

220. **Mekhazni, D.**, **M. Dufau**, **C. Desrosiers**, **M. Pedersoli**, and **E. Granger**, Camera alignment and weighted contrastive learning for domain adaptation in video person reid. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023.

221. **Mirza, M.** and **S. Osindero** (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

222. **Miyato, T.**, **S.-i. Maeda**, **M. Ishii**, and **M. Koyama**, Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *In Proceedings of the International Conference on Learning Representations (ICLR)*. 2018.

223. **Mohanty, A.**, **B. Banerjee**, and **R. Velmurugan** (2022). Ssmtreid-net: Multi-target unsupervised domain adaptation for person re-identification. *Pattern Recognition Letters*, **163**, 40–46.

224. **Moosavi-Dezfooli, S.-M.**, **A. Fawzi**, and **P. Frossard**, Deepfool: a simple and accurate method to fool deep neural networks. *In CVPR*. 2016.

225. **Nayak, G. K.**, **R. Rawal**, and **A. Chakraborty**, De-crop: Data-efficient certified robustness for pretrained classifiers. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023.

226. **Ni, H.**, **J. Song**, **X. Luo**, **F. Zheng**, **W. Li**, and **H. T. Shen**, Meta distribution alignment for generalizable person re-identification. *In CVPR*. 2022.

227. **Nichol, A.** and **J. Schulman** (2018). Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, **2**(3), 4.

228. **Odena, A.**, **C. Olah**, and **J. Shlens**, Conditional image synthesis with auxiliary classifier gans. *In Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

229. **Oh, T.-H.**, **S. Jung**, **J.-Y. Lee**, and **I. S. Kim**, Adversarial perturbations to manipulate the appearance of 3d objects. *In Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

230. **Papernot, N.** and **P. McDaniel** (2016). On the effectiveness of defensive distillation. *arXiv preprint arXiv:1607.05113*.

231. **Papernot, N.**, **P. McDaniel**, **S. Jha**, **M. Fredrikson**, **Z. B. Celik**, and **A. Swami**, The limitations of deep learning in adversarial settings. *In 2016 EuroS&P*. IEEE, 2016*a*.

232. **Papernot, N.**, **P. McDaniel**, **X. Wu**, **S. Jha**, and **A. Swami**, Distillation as a defense to adversarial perturbations against deep neural networks. *In 2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016*b*.

233. **Peng, J.**, **G. Jiang**, and **H. Wang** (2023). Adaptive memorization with group labels for unsupervised person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

234. **Peng, P.**, **T. Xiang**, **Y. Wang**, **M. Pontil**, **S. Gong**, **T. Huang**, and **Y. Tian**, Unsupervised cross-dataset transfer learning for person re-identification. *In CVPR*. 2016.

235. **Peyré, G.**, **M. Cuturi**, *et al.* (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, **11**(5-6), 355–607.

236. **Plantinga, A.** (1961). Things and persons. *The Review of Metaphysics*, 493–519.

237. **Poursaeed, O.**, **I. Katsman**, **B. Gao**, and **S. Belongie**, Generative adversarial perturbations. *In CVPR*. 2018.

238. **Pu, N.**, **W. Chen**, **Y. Liu**, **E. M. Bakker**, and **M. S. Lew**, Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. *In Proceedings of the 28th ACM International Conference on Multimedia*. 2020.

239. **Qi, H.**, **X. Wang**, **J. Zhang**, **J. Liu**, and **W.-L. Chao**, Adversarial person re-identification. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

240. **Qi, L.**, **J. Liu**, **L. Wang**, **Y. Shi**, and **X. Geng** (2023). Unsupervised generalizable multi-source person re-identification: A domain-specific adaptive framework. *Pattern Recognition*, 109546.

241. **Qian, X.**, **Y. Fu**, **Y.-G. Jiang**, **T. Xiang**, and **X. Xue**, Multi-scale deep learning architectures for person re-identification. *In CVPR*. 2017.

242. **Qin, C.**, **R. Liu**, **L. Pan**, **X. Wang**, **C. Liu**, **E. Ding**, and **J. Cheng** (2019). Adversarial robustness: From self-supervised pre-training to fine-tuning. *arXiv preprint arXiv:1907.00282*.

243. **Qiu, S.**, **Q. Liu**, **S. Zhou**, and **C. Wu** (2019). Review of artificial intelligence adversarial attack and defense technologies. *Applied Sciences*, **9**(5), 909.

244. **Qu, X.**, **L. Liu**, **L. Zhu**, and **H. Zhang** (2023). Attribute-aware style adaptation for person re-identification. *Multimedia Systems*, **29**(2), 469–485.

245. **Quan, R.**, **B. Xu**, and **D. Liang** (2023). Discriminatively unsupervised learning person re-identification via considering complicated images. *Sensors*, **23**(6), 3259.

246. **Radford, A.**, **L. Metz**, and **S. Chintala**, Unsupervised representation learning with deep convolutional generative adversarial networks. *In Proceedings of the 4th International Conference on Learning Representations (ICLR)*. 2016*a*.

247. **Radford, A.**, **L. Metz**, and **S. Chintala** (2016*b*). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

248. **Raghunathan, A.**, **J. Steinhardt**, and **P. Liang** (2018). Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*.

249. **Raj, A.**, **Y. Bresler**, and **B. Li**, Improving robustness of deep-learning-based image reconstruction. *In ICML*. PMLR, 2020.

250. **Rami, H.**, **M. Ospici**, and **S. Lathuilière**, Online unsupervised domain adaptation for person re-identification. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

251. **Ren, C.-X.**, **B. Liang**, **P. Ge**, **Y. Zhai**, and **Z. Lei** (2019). Domain adaptive person re-identification via camera style generation and label propagation. *TIFS*, **15**, 1290–1302.

252. **Ren, K.**, **T. Zheng**, **Z. Qin**, and **X. Liu** (2020). Adversarial attacks and defenses in deep learning. *Engineering*, **6**(3), 346–360.

253. **Rezende, D.** and **S. Mohamed**, Variational inference with normalizing flows. *In International conference on machine learning*. PMLR, 2015.

254. **Rice, L.**, **E. Wong**, and **Z. Kolter**, Overfitting in adversarially robust deep learning. *In ICML*. PMLR, 2020.

255. **Ristani, E.**, **F. Solera**, **R. Zou**, **R. Cucchiara**, and **C. Tomasi**, Performance measures and a data set for multi-target, multi-camera tracking. *In ECCV*. Springer, 2016.

256. **Rombach, R.**, **A. Blattmann**, **D. Lorenz**, **P. Esser**, and **B. Ommer**, High-resolution image synthesis with latent diffusion models. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

257. **Ronneberger, O.**, **P. Fischer**, and **T. Brox**, U-net: Convolutional networks for biomedical image segmentation. *In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015.

258. **Rorty, A. O.** (1973). The transformations of persons. *Philosophy*, **48**(185), 261–275.

259. **Rumezhak, T.**, **F. G. Eiras**, **P. H. Torr**, and **A. Bibi**, Rancer: Non-axis aligned anisotropic certification with randomized smoothing. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023.

260. **Saha, A.**, **A. Subramanya**, **K. Patil**, and **H. Pirsiavash**, Role of spatial context in adversarial robustness for object detection. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020.

261. **Salimans, T.**, **I. Goodfellow**, **W. Zaremba**, **V. Cheung**, **A. Radford**, and **X. Chen**, Improved techniques for training gans. *In Advances in Neural Information Processing Systems (NIPS)*. 2016.

262. **Salman, H.**, **S. Jain**, **E. Wong**, and **A. Madry**, Certified patch robustness via smoothed vision transformers. *In Proceedings of the CVPR*. 2022.

263. **Salman, H.**, **J. Li**, **I. Razenshteyn**, **P. Zhang**, **H. Zhang**, **S. Bubeck**, and **G. Yang** (2019). Provably robust deep learning via adversarially trained smoothed classifiers. *NIPS*, **32**.

264. **Salman, H.**, **M. Sun**, **G. Yang**, **A. Kapoor**, and **J. Z. Kolter** (2020). Denoised smoothing: A provable defense for pretrained classifiers. *Adv. neural inf. process. syst*, **33**, 21945–21957.

265. **Sehwag, V.**, **S. Mahloujifar**, **T. Handina**, **S. Dai**, **C. Xiang**, **M. Chiang**, and **P. Mittal** (2021). Robust learning meets generative models: Can proxy distributions improve adversarial robustness? *arXiv preprint arXiv:2104.09425*.

266. **Seo, S.**, **Y. Lee**, and **P. Kang** (2023). Cost-free adversarial defense: Distance-based optimization for model robustness without adversarial training. *Computer Vision and Image Understanding*, **227**, 103599.

267. **Sheikh, H.** *et al.* (2004). Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, **13**(4), 600G612.

268. **Siarohin, A.**, **E. Sangineto**, **S. Lathuilière**, and **N. Sebe**, Deformable gans for pose-based human image generation. *In CVPR*. 2018.

269. **Sinha, A.**, **S. P. Joshi**, **P. S. Das**, **S. Jana**, and **R. Sarkar** (2022). An ml prediction model based on clinical parameters and automated ct scan features for covid-19 patients. *Scientific Reports*, **12**(1), 11255.

270. **Song, C.**, **K. He**, **L. Wang**, and **J. E. Hopcroft** (2018). Improving the generalization of adversarial training with domain adaptation. *arXiv*.

271. **Song, J.**, **A. Vahdat**, **M. Mardani**, and **J. Kautz**, Pseudoinverse-guided diffusion models for inverse problems. *In International Conference on Learning Representations*. 2023.

272. **Song, L.**, **C. Wang**, **L. Zhang**, **B. Du**, **Q. Zhang**, **C. Huang**, and **X. Wang** (2020). Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, **102**, 107173.

273. **Song, S.**, **W. Zhang**, **J. Liu**, and **T. Mei**, Unsupervised person image generation with semantic parsing transformation. *In CVPR*. 2019.

274. **Song, Y.** and **S. Ermon** (2019). Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, **32**.

275. **Soo, S.** (2014). Object detection using haar-cascade classifier. *Institute of Computer Science, University of Tartu*, **2**(3), 1–12.

276. **Su, J.**, **D. V. Vargas**, and **K. Sakurai** (2019). One pixel attack for fooling deep neural networks. *TEVC*, **23**(5), 828–841.

277. **Subramanyam, A.** (2023). Meta generative attack on person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*.

278. **Sun, B.** and **K. Saenko**, Deep coral: Correlation alignment for deep domain adaptation. *In ECCV*. 2016.

279. **Sun, Q.**, **Y. Liu**, **T.-S. Chua**, and **B. Schiele**, Meta-transfer learning for few-shot learning. *In CVPR*. 2019.

280. **Sun, X.**, **G. Cheng**, **H. Li**, **L. Pei**, and **J. Han**, Exploring effective data for surrogate training towards black-box attack. *In CVPR*. 2022.

281. **Sun, Y.**, **L. Zheng**, **Y. Yang**, **Q. Tian**, and **S. Wang**, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). *In ECCV*. 2018.

282. **Szegedy, C.**, **W. Zaremba**, **I. Sutskever**, **J. Bruna**, **D. Erhan**, **I. Goodfellow**, and **R. Fergus** (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

283. **Tai, Y.**, **J. Yang**, **X. Liu**, and **C. Xu**, Memnet: A persistent memory network for image restoration. *In Proceedings of the ICCV*. 2017.

284. **Tang, H.**, **D. Xu**, **G. Liu**, **W. Wang**, **N. Sebe**, and **Y. Yan**, Cycle in cycle generative adversarial networks for keypoint-guided image generation. *In ACMMM*. 2019.

285. **Tang, Y.**, **Y. Xi**, **N. Wang**, **B. Song**, and **X. Gao** (2020). Cgan-tm: A novel domain-to-domain transferring method for person re-identification. *TIP*, **29**, 5641–5651.

286. **Tashiro, Y.**, **Y. Song**, and **S. Ermon** (2020). Diversity can be transferred: Output diversification for white-and black-box attacks. *NIPS*, **33**, 4536–4548.

287. **Tjeng, V.**, **K. Xiao**, and **R. Tedrake** (2017). Evaluating robustness of neural networks with mixed integer programming. *arXiv preprint arXiv:1711.07356*.

288. **Tramer, F.**, **N. Carlini**, **W. Brendel**, and **A. Madry** (2020). On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, **33**, 1633–1645.

289. **Tramèr, F.**, **A. Kurakin**, **N. Papernot**, **I. Goodfellow**, **D. Boneh**, and **P. McDaniel** (2017). Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*.

290. **Tseng, C.-H.**, **H.-C. Liu**, **S.-J. Lee**, and **X. Zeng**, Perturbed gradients updating within unit space for deep learning. *In 2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022.

291. **Tsipras, D.**, **S. Santurkar**, **L. Engstrom**, **A. Turner**, and **A. Madry**, Robustness may be at odds with accuracy. *In ICLR*. 2019. URL `https://openreview.net/forum?id=SyxAb30cY7`.

292. **Tu, C.-C.**, **P. Ting**, **P.-Y. Chen**, **S. Liu**, **H. Zhang**, **J. Yi**, **C.-J. Hsieh**, and **S.-M. Cheng**, Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. *In Proceedings of the AIES*, volume 33. 2019.

293. **Tu, J.**, **M. Ren**, **S. Manivasagam**, **M. Liang**, **B. Yang**, **R. Du**, **F. Cheng**, and **R. Urtasun**, Physically realizable adversarial examples for lidar object detection. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

294. **Tzeng, E.**, **J. Hoffman**, **K. Saenko**, and **T. Darrell**, Adversarial discriminative domain adaptation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

295. **Tzeng, E.**, **J. Hoffman**, **N. Zhang**, **K. Saenko**, and **T. Darrell** (2014). Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

296. **Uesato, J.**, **B. O'Donoghue**, **P. Kohli**, and **A. van den Oord**, Adversarial risk and the dangers of evaluating against weak attacks. *In Proceedings of the PMLR*. 2018.

297. **van den Oord, A.**, **N. Kalchbrenner**, and **K. Kavukcuoglu** (2016). Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*.

298. **Wang, D.**, **H. Lu**, **H. Wang**, and **M. Yang** (2019*a*). Vehicle re-identification with learned representation and spatial verification. *IEEE Transactions on Image Processing*, **28**(6), 2775–2789.

299. **Wang, H.**, **G. Wang**, **Y. Li**, **D. Zhang**, and **L. Lin**, Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020*a*.

300. **Wang, H.** and **Y. Wang** (2022). Self-ensemble adversarial training for improved robustness. *ICLR*.

301. **Wang, J.**, **X. Zhu**, **S. Gong**, and **W. Li**, Transferable joint attribute-identity deep learning for unsupervised person re-identification. *In CVPR*. 2018*a*.

302. **Wang, K.**, **Z. He**, **Z. Lin**, **J. Liu**, **T. Wang**, and **J. Liu**, Adversarial pruning: Towards more robust deep neural networks. *In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020*b*.

303. **Wang, L.**, **H. Li**, **D. Zou**, **T. Zhang**, and **H. Chen** (2021*a*). Boosting black-box adversarial attacks via translation. *arXiv preprint arXiv:2104.01732*.

304. **Wang, L.**, **W. Zhang**, **D. Wu**, **F. Zhu**, and **B. Li**, Attack is the best defense: Towards preemptive-protection person re-identification. *In Proceedings of the 30th ACM International Conference on Multimedia*. 2022*a*.

305. **Wang, S.**, **R. Liu**, **H. Li**, **G. Qi**, and **Z. Yu** (2022*b*). Occluded person re-identification via defending against attacks from obstacles. *IEEE Transactions on Information Forensics and Security*, **18**, 147–161.

306. **Wang, S.**, **P. Yang**, **L. Li**, **S. Li**, and **K. Li** (2018*b*). A survey on deep learning for big data. *Information Fusion*, **42**, 146–157.

307. **Wang, S.**, **T. Zhang**, **Q. Huang**, and **X.-S. Hua** (2020*c*). A black-box adversarial attack on re-identification models. *IEEE Transactions on Information Forensics and Security*, **15**, 1674–1685.

308. **Wang, W.**, **Y. Hu**, **F. Wu**, and **Z. Zhang**, Isometric projection for manifold learning with diffusion geometry. *In International Conference on Neural Information Processing*. Springer, 2021*b*.

309. **Wang, W.**, **B. Yin**, **T. Yao**, **L. Zhang**, **Y. Fu**, **S. Ding**, **J. Li**, **F. Huang**, and **X. Xue**, Delving into data: Effectively substitute training for black-box attack. *In CVPR*. 2021*c*.

310. **Wang, X.** (2013). Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, **34**(1), 3–19.

311. **Wang, Z.**, **R. Hu**, **C. Liang**, **Y. Yu**, **J. Jiang**, **M. Ye**, **J. Chen**, and **Q. Leng** (2016). Zero-shot person re-identification via cross-view consistency. *TMM*, **18**(2), 260–272.

312. **Wang, Z.**, **R. Hu**, **Y. Yu**, **C. Liang**, and **W. Huang**, Multi-level fusion for person re-identification with incomplete marks. *In ACMMM*. 2015.

313. **Wang, Z.**, **J. Jiang**, **Y. Yu**, and **S. Satoh** (2019*b*). Incremental re-identification by cross-direction and cross-ranking adaption. *IEEE Transactions on Multimedia*, **21**(9), 2376–2386.

314. **Wang, Z.**, **S. Zheng**, **M. Song**, **Q. Wang**, **A. Rahimpour**, and **H. Qi**, advpattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns. *In CVPR*. 2019*c*.

315. **Wei, J.**, **L. Yao**, and **Q. Meng** (2023). Self-adaptive logit balancing for deep neural network robustness: Defence and detection of adversarial attacks. *Neurocomputing*.

316. **Wei, L.**, **S. Zhang**, **W. Gao**, and **Q. Tian**, Person transfer gan to bridge domain gap for person re-identification. *In CVPR*. 2018.

317. **Wei, W.**, **S. Yang**, **Y. Luo**, **Z. Wu**, and **J. Liu**, Vehicle re-identification using hierarchical spatial transformer network. *In IEEE International Conference on Computer Vision (ICCV)*. 2019.

318. **Wei, X.**, **Y. Guo**, **J. Yu**, and **B. Zhang** (2022). Simultaneously optimizing perturbations and positions for black-box adversarial patch attacks. *IEEE Trans. Pattern Anal. Mach. Intell.*

319. **Wolf, A.** (). Making medical image reconstruction adversarially robust.

320. **Wong, E.** and **Z. Kolter**, Provable defenses against adversarial examples via the convex outer adversarial polytope. *In ICML*. PMLR, 2018.

321. **Wong, E.**, **L. Rice**, and **J. Z. Kolter** (2020). Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*.

322. **Wong, E.**, **F. Schmidt**, and **Z. Kolter**, Wasserstein adversarial examples via projected sinkhorn iterations. *In ICML*. PMLR, 2019.

323. **Wu, A.**, **W.-S. Zheng**, and **J.-H. Lai**, Unsupervised person re-identification by camera-aware similarity consistency learning. *In ICCV*. 2019.

324. **Wu, J.**, **Y. Yang**, **Z. Lei**, **Y. Yang**, **S. Chen**, and **S. Z. Li** (2023). Camera-aware representation learning for person re-identification. *Neurocomputing*, **518**, 155–164.

325. **Wu, K.**, **A. Wang**, and **Y. Yu**, Stronger and faster wasserstein adversarial attacks. *In ICML*. PMLR, 2020*a*.

326. **Wu, W.**, **Y. Su**, **X. Chen**, **S. Zhao**, **I. King**, **M. R. Lyu**, and **Y.-W. Tai**, Boosting the transferability of adversarial samples via attention. *In CVPR*. 2020*b*.

327. **Wu, Y.**, **Y. Lin**, **X. Dong**, **Y. Yan**, **W. Ouyang**, and **Y. Yang**, Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018*a*.

328. **Wu, Y.**, **C. Zhao**, **Y. Zhang**, **S. Liu**, **W. Chen**, **L. Liu**, and **J. Lu**, Defense against adversarial attacks using high-level representation guided denoiser. *In Thirty-Second AAAI Conference on Artificial Intelligence*. 2018*b*.

329. **Xiao, C.**, **T. Ma**, **J. Liu**, **Y. Wang**, **C. Sun**, and **H. Huang**, Generating diverse adversarial examples with generative models. *In International Conference on Machine Learning*. PMLR, 2021.

330. **Xiao, T.**, **H. Li**, **W. Ouyang**, and **X. Wang**, Learning deep feature representations with domain guided dropout for person re-identification. *In CVPR*. 2016.

331. **Xie, C.**, **J. Wang**, **Z. Zhang**, **Y. Zhou**, **L. Xie**, and **A. Yuille**, Adversarial examples for semantic segmentation and object detection. *In CVPR*. 2017.

332. **Xie, C.**, **Y. Wu**, **L. v. d. Maaten**, **A. L. Yuille**, and **K. He**, Feature denoising for improving adversarial robustness. *In CVPR*. 2019.

333. **Xie, S.**, **Z. Zheng**, **L. Chen**, and **C. Chen**, Learning semantic representations for unsupervised domain adaptation. *In ICML*. 2018.

334. **Xu, H.**, **X. Liu**, **Y. Li**, **A. Jain**, and **J. Tang**, To be robust or to be fair: Towards fairness in adversarial training. *In International Conference on Machine Learning*. PMLR, 2021*a*.

335. **Xu, X.**, **W. Wang**, **Z. Li**, **Y. Zhang**, and **Y. Guo** (2021*b*). Unsupervised domain adaptation for video re-identification with similarity learning. *Information Sciences*, **576**, 214–224.

336. **Xu, Y.**, **B. Du**, and **L. Zhang** (2020). Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses. *IEEE Transactions on Geoscience and Remote Sensing*, **59**(2), 1604–1617.

337. **Yan, J.**, **H. Yin**, **X. Deng**, **Z. Zhao**, **W. Ge**, **H. Zhang**, and **G. Rigoll** (2022). Wavelet regularization benefits adversarial training. *arXiv preprint arXiv:2206.03727*.

338. **Yang, F.**, **J. Weng**, **Z. Zhong**, **H. Liu**, **Z. Wang**, **Z. Luo**, **D. Cao**, **S. Li**, **S. Satoh**, and **N. Sebe** (2022). Towards robust person re-identification by defending against universal attackers. *TPAMI*.

339. **Yang, F.**, **K. Yan**, **S. Lu**, **H. Jia**, **D. Xie**, **Z. Yu**, **X. Guo**, **F. Huang**, and **W. Gao** (2020). Part-aware progressive unsupervised domain adaptation for person re-identification. *TMM*.

340. **Yang, F.**, **Z. Zhong**, **H. Liu**, **Z. Wang**, **Z. Luo**, **S. Li**, **N. Sebe**, and **S. Satoh**, Learning to attack real-world models for person re-identification via virtual-guided meta-learning. *In AAAI*, volume 35. 2021*a*.

341. **Yang, F.**, **Z. Zhong**, **Z. Luo**, **Y. Cai**, **Y. Lin**, **S. Li**, and **N. Sebe**, Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. *In CVPR*. 2021*b*.

342. **Yang, M.**, **L. Wang**, **J. Zhang**, and **J. Yang**, Towards a scalable and effective deep architecture for vehicle re-identification. *In IEEE International Conference on Computer Vision (ICCV)*. 2017.

343. **Yang, Y.**, **Z. Cheng**, **K. Zhang**, **F. Duan**, and **X. Xi**, Learning to compare: Relation network for person re-identification. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

344. **Yao, Z.**, **B. Zhang**, **Z. Wang**, **W. Ouyang**, **D. Xu**, and **D. Feng**, Intersectgan: Learning domain intersection for generating images with multiple attributes. *In ACMMM*. 2019.

345. **Ye, M.**, **C. Liang**, **Y. Yu**, **Z. Wang**, **Q. Leng**, **C. Xiao**, **J. Chen**, and **R. Hu** (2016). Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *TMM*, **18**(12), 2553–2566.

346. **Ye, Y.**, **T. Pan**, **Q. Meng**, **J. Li**, and **H. T. Shen** (2022). Online unsupervised domain adaptation via reducing inter-and intra-domain discrepancies. *IEEE Transactions on Neural Networks and Learning Systems*.

347. **Yin, F.**, **Y. Zhang**, **B. Wu**, **Y. Feng**, **J. Zhang**, **Y. Fan**, and **Y. Yang** (2023). Generalizable black-box adversarial attack with meta learning. *IEEE Trans. Pattern Anal. Mach. Intell.*

348. **Yu, H.-X.**, **A. Wu**, and **W.-S. Zheng**, Cross-view asymmetric metric learning for unsupervised person re-identification. *In ICCV*. 2017.

349. **Yu, T.**, **Y. Yang**, **D. Li**, **T. Hospedales**, and **T. Xiang**, Simple and effective stochastic neural networks. *In AAAI*, volume 35. 2021.

350. **Yuan, Z.**, **J. Zhang**, **Y. Jia**, **C. Tan**, **T. Xue**, and **S. Shan**, Meta gradient adversarial attack. *In CVPR*. 2021.

351. **Zeng, Z.**, **Z. Wang**, **Z. Wang**, **Y. Zheng**, **Y.-Y. Chuang**, and **S. Satoh** (2020). Illumination-adaptive person re-identification. *TMM*.

352. **Zhang, C.**, **P. Benz**, **A. Karjauv**, **J. W. Cho**, **K. Zhang**, and **I. S. Kweon**, Investigating top-k white-box and transferable black-box attack. *In CVPR*. 2022*a*.

353. **Zhang, C.**, **C. Zhang**, **M. Zhang**, and **I. S. Kweon** (2023). Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.

354. **Zhang, D.**, **Z. Yang**, and **L. Zhang** (2019*a*). Clustering-guided self-paced learning for unsupervised person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, **29**(6), 1776–1785.

355. **Zhang, H.** and **J. Wang**, Towards adversarially robust object detection. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

356. **Zhang, H.**, **T.-W. Weng**, **P.-Y. Chen**, **C.-J. Hsieh**, and **L. Daniel** (2018*a*). Efficient neural network robustness certification with general activation functions. *Adv. neural inf. process. syst*, **31**.

357. **Zhang, H.**, **Y. Yu**, **J. Jiao**, **E. Xing**, **L. El Ghaoui**, and **M. Jordan**, Theoretically principled trade-off between robustness and accuracy. *In ICML*. PMLR, 2019*b*.

358. **Zhang, J.**, **J. Huang**, **Z. Tian**, and **S. Lu**, Spectral unsupervised domain adaptation for visual recognition. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022*b*.

359. **Zhang, J.**, **Y. Wu**, **J. Sun**, **X. Xue**, and **L. Zhang** (2020*a*). Adversarial open-world person re-identification. *IEEE Transactions on Image Processing*, **29**, 6024–6038.

360. **Zhang, K.**, **W. Luo**, **T. Xiang**, and **J. Xiao**, Unsupervised domain adaptation for person re-identification via simultaneous feature alignment and similarity diffusion. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020*b*.

361. **Zhang, K.**, **W. Zuo**, **Y. Chen**, **D. Meng**, and **L. Zhang** (2017*a*). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, **26**(7), 3142–3155.

362. **Zhang, M.**, **H. Wang**, **P. He**, **A. Malik**, and **H. Liu**, Improving gan-generated image detection generalization using unsupervised domain adaptation. *In 2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022*c*.

363. **Zhang, Q.**, **S. Bai**, and **P. H. Torr**, Occluded person re-identification. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018*b*.

364. **Zhang, R.**, **P. Isola**, **A. A. Efros**, **E. Shechtman**, and **O. Wang**, The unreasonable effectiveness of deep features as a perceptual metric. *In CVPR*. 2018*c*.

365. **Zhang, W.** and **D. Wu**, Discriminative joint probability maximum mean discrepancy (djp-mmd) for domain adaptation. *In 2020 IJCNN*. IEEE, 2020.

366. **Zhang, X.**, **J. Cao**, **C. Shen**, and **M. You**, Self-training with progressive augmentation for unsupervised cross-domain person re-identification. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019*c*.

367. **Zhang, X.**, **Y. Ge**, **Y. Qiao**, and **H. Li**, Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

368. **Zhang, X.**, **H. Luo**, **X. Fan**, **W. Xiang**, **Y. Sun**, **Q. Xiao**, **W. Jiang**, **C. Zhang**, and **J. Sun** (2017*b*). Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*.

369. **Zhang, Y.**, **Y. Yao**, **J. Jia**, **J. Yi**, **M. Hong**, **S. Chang**, and **S. Liu** (2022*d*). How to robustify black-box ml models? a zeroth-order optimization perspective. *ICLR*.

370. **Zhang, Z.**, Improved adam optimizer for deep neural networks. *In 2018 IEEE/ACM 26th IWQoS*. Ieee, 2018.

371. **Zhao, C.**, **K. Chen**, **Z. Wei**, **Y. Chen**, **D. Miao**, and **W. Wang** (2019*a*). Multilevel triplet deep learning model for person re-identification. *Pattern Recognition Letters*, **117**, 161–168.

372. **Zhao, C.**, **X. Lv**, **S. Dou**, **S. Zhang**, **J. Wu**, and **L. Wang** (2021*a*). Incremental generative occlusion adversarial suppression network for person reid. *TIP*, **30**, 4212–4224.

373. **Zhao, C.**, **X. Lv**, **Z. Zhang**, **W. Zuo**, **J. Wu**, and **D. Miao** (2020*a*). Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification. *TMM*, **22**(12), 3180–3195.

374. **Zhao, C.**, **X. Wang**, **W. Zuo**, **F. Shen**, **L. Shao**, and **D. Miao** (2020*b*). Similarity learning with joint transfer constraints for person re-identification. *Pattern Recognition*, **97**, 107014.

375. **Zhao, G.**, **M. Zhang**, **J. Liu**, **Y. Li**, and **J.-R. Wen** (2022). Ap-gan: Adversarial patch attack on content-based image retrieval systems. *GeoInformatica*, 1–31.

376. **Zhao, G.**, **M. Zhang**, **J. Liu**, and **J.-R. Wen** (2019*b*). Unsupervised adversarial attacks on deep feature-based retrieval with gan. *arXiv preprint arXiv:1907.05793*.

377. **Zhao, R.**, **W. Ouyang**, and **X. Wang**, Learning mid-level filters for person re-identification. *In CVPR*. 2014.

378. **Zhao, Y.**, **Z. Zhong**, **F. Yang**, **Z. Luo**, **Y. Lin**, **S. Li**, and **N. Sebe**, Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. *In CVPR*. 2021*b*.

379. **Zheng, D.**, **J. Xiao**, **K. Chen**, **X. Huang**, **L. Chen**, and **Y. Zhao** (2022*a*). Soft pseudo-label shrinkage for unsupervised domain adaptive person re-identification. *Pattern Recognition*, **127**, 108615.

380. **Zheng, D.**, **J. Xiao**, **Y. Wei**, **Q. Wang**, **K. Huang**, and **Y. Zhao** (2022*b*). Unsupervised domain adaptation in homogeneous distance space for person re-identification. *Pattern Recognition*, **132**, 108941.

381. **Zheng, K.**, **C. Lan**, **W. Zeng**, **Z. Zhang**, and **Z.-J. Zha**, Exploiting sample uncertainty for domain adaptive person re-identification. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35. 2021.

382. **Zheng, L.**, **Z. Bie**, **Y. Sun**, **J. Wang**, **C. Su**, **S. Wang**, and **Q. Tian**, Mars: A video benchmark for large-scale person re-identification. *In ECCV*. 2016*a*.

383. **Zheng, L.**, **L. Shen**, **L. Tian**, **S. Wang**, **J. Wang**, and **Q. Tian**, Scalable person re-identification: A benchmark. *In ICCV*. 2015.

384. **Zheng, L.**, **Y. Yang**, and **A. G. Hauptmann** (2016*b*). Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

385. **Zheng, W.-S.**, **S. Gong**, and **T. Xiang**, Person re-identification by probabilistic relative distance comparison. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

386. **Zheng, Z.**, **L. Zheng**, **Z. Hu**, and **Y. Yang** (2018). Open set adversarial examples. *arXiv preprint arXiv:1809.02681*, **3**.

387. **Zheng, Z.**, **L. Zheng**, and **Y. Yang**, A discriminatively learned cnn embedding for person re-identification. *In Proceedings of the ACM International Conference on Multimedia (MM)*. 2016*c*.

388. **Zheng, Z.**, **L. Zheng**, and **Y. Yang** (2017). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, **3**.

389. **Zhong, Z.**, **L. Zheng**, **D. Cao**, and **S. Li**, Re-ranking person re-identification with k-reciprocal encoding. *In CVPR*. 2017.

390. **Zhong, Z.**, **L. Zheng**, **S. Li**, and **Y. Yang**, Generalizing a person retrieval model hetero- and homogeneously. *In ECCV*. 2018*a*.

391. **Zhong, Z.**, **L. Zheng**, **Z. Zheng**, **S. Li**, and **Y. Yang**, Camera style adaptation for person re-identification. *In CVPR*. 2018*b*.

392. **Zhou, K.**, **Y. Qiao**, and **T. Xiang**, Deep end-to-end relation network for person re-identification. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

393. **Zhou, K.**, **Y. Yang**, **A. Cavallaro**, and **T. Xiang**, Omni-scale feature learning for person re-identification. *In CVPR*. 2019.

394. **Zhou, M.**, **Z. Niu**, **L. Wang**, **Q. Zhang**, and **G. Hua**, Adversarial ranking attack and defense. *In ECCV*. Springer, 2020.

395. **Zhou, M.** and **V. M. Patel**, Enhancing adversarial robustness for deep metric learning. *In CVPR*. 2022.

396. **Zhou, Y.**, **F. Huang**, **W. Chen**, **S. Pu**, and **L. Zhang** (2023). Stochastic gradient perturbation: An implicit regularizer for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

397. **Zhu, J.-Y.**, **T. Park**, **P. Isola**, and **A. A. Efros**, Unpaired image-to-image translation using cycle-consistent adversarial networks. *In CVPR*. 2017.

398. **Zhu, X.**, **X. Zhan**, **L. He**, and **J. Sun**, Crafting gait recognition on a budget: Pedestrian re-identification with a single outdoor image. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.