T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

PERSON RE-IDENTIFICATION USING NOVEL REGULARIZATION METHODS IN DEEP NETWORKS

AYŞE ŞERBETÇİ TURAN
A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
DEPARTMENT OF COMPUTER ENGINEERING

GEBZE
2020

T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

PERSON RE-IDENTIFICATION USING NOVEL REGULARIZATION METHODS IN DEEP NETWORKS

AYŞE ŞERBETÇİ TURAN

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER ENGINEERING

THESIS SUPERVISOR

PROF. DR. YUSUF SİNAN AKGÜL

GEBZE

2020

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DERİN ÖĞRENME AĞLARINDA ÖZGÜN REGULARİZASYON YÖNTEMLERİ KULLANARAK KİŞİ YENİDEN KİMLİKLENDİRME

AYŞE ŞERBETÇİ TURAN
DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMANI
PROF. DR. YUSUF SİNAN AKGÜL

GEBZE
2020

GEBZE TEKNİK ÜNİVERSİTESİ	DOKTORA JÜRİ ONAY FORMU

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 24/07/2020 tarih ve 2020/36 sayılı kararıyla oluşturulan jüri tarafından 24/08/2020 tarihinde tez savunma sınavı yapılan Ayşe Şerbetçi Turan'ın tez çalışması Bilgisayar Mühendisliği Anabilim Dalında DOKTORA tezi olarak kabul edilmiştir.

JÜRİ

ÜYE
(TEZ DANIŞMANI] : Prof. Yusuf Sinan Akgül

ÜYE : Prof. Dr. Songül Varlı Albayrak

ÜYE : Dr. Öğr. Üyesi Yakup Genç

ÜYE : Doç. Dr. Erchan Aptoula

ÜYE : Dr. Öğr. Üyesi Ayşe Betül Oktay

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun
…..…/…..…/..…..… tarih ve ……/…….. sayılı kararı.

İMZA/MÜHÜR

SUMMARY

Person Re-Identification (ReID) aims at retrieving the images of a query person from a large set of gallery images. It has been an attractive research field in computer vision due to the ever-increasing demand for camera networks in public spaces. In recent years, significant improvements have been observed in person ReID task in parallel with the developments in deep learning. However, due to the large discrepancy between the training/test distributions, the ReID models generally lack in generalizing to the test data, which is the phenomenon known as overfitting.

In this thesis, we propose an ensemble method to increase the generalization capability of the ReID models. Ensemble models, which consist of multiple base learners whose decisions are combined in test time, deal with the overfitting problem effectively and increase the generalization capability. However, training an ensemble of deep networks is computationally inefficient. To overcome this difficulty, we create diverse and accurate base learners in a single network by designing a multi-branch architecture. Detailed analysis of the experiments on three benchmark datasets demonstrates the effectiveness of our approach, which outperforms the state-of-the-art approaches. We adapt the proposed approach to Binary Neural Networks. Our experiments show that the proposed approach improves the Binary Neural Networks in terms accuracy and training stability in image classification task and outperforms the conventional ensemble model by a large margin in person ReID, which indicates that our model is not only an ensemble model, but also an effective regularizer for deep networks.

Keywords: Person Re-Identification, Ensemble, Overfitting, Deep Learning

ÖZET

Kişi yeniden kimliklendirme (KYK), geniş bir veritabanından bir sorgu kişisine ait görüntüleri getirme problemidir. Halka açık alanlarda güvenlik kamera ağlarına talebin artmasından dolayı KYK problemini iyileştirmeyi amaçlayan çalışmalar da daha fazla rağbet görmeye başlamıştır. KYK problemi için geliştirilen modeler eğitim/test verisi arasındaki kişi farklılıkları ve değişen görüntüleme koşullarından dolayı test verisi üzerinde düşük performans gösterir. Bu şekilde test üzerinde düşük performans gösteren modellerin genelleştirme kapasitesi düşüktür ve bu probleme "overfitting" denir.

Bu tez çalışmasında, KYK modellerinin genelleştirme kapasitesini artırmak amacıyla bir topluluk öğrenme yöntemi önerilmektedir. Topluluk öğrenme yöntemleri birçok temel öğrenicinin kararlarının birleştirilmesiyle oluşmaktadır ve bu modellerin genelleştirme kapasitesi temel öğrenicilere göre daha yüksektir. Ancak derin ağlardan oluşan bir topluluk öğrenme modeli geliştirmek pahalı bir yöntemdir. Bu zorluğu aşmak için bu çalışmada çok dallı bir derin ağ tasarımı yapılarak birçok temel öğrenicinin aynı derin ağ üzerinde eğitilmesi önerilmektedir. Literatürde çokça kullanılan üç adet veri kümesi üzerinde gerçekleştirilen deneyler ve bunların detaylı analizleri önerilen yöntemin var olan yöntemlerden daha iyi performans gösterdiğini ortaya koymaktadır. Önerilen yöntem ayrıca İkili Yapay Sinir Ağları'na uyarlanmıştır. Yapılan deneyler önerilen yöntemin görüntü sınıflandırma probleminde İkili Sinir Ağları'nı hem doğruluk hem de eğitim sırasında model kararlılığı açısından iyileştirdiğini, KYK'da ise topluluk öğrenicisi yöntemine göre çok daha iyi performans gösterdiğini ortaya koymaktadır. Bu sonuçlar önerilen yöntemin sadece verimli bir topluluk öğrenme yöntemi olmadığını, aynı zamanda derin ağlar için etkili bir regularizasyon yöntemi olduğunu da göstermektedir.

Anahtar Kelimeler: Kişi Yeniden Kimliklendirme, Topluluk Öğrenme Modeli, Derin Öğrenme

ACKNOWLEDGEMENTS

I would first like to thank my supervisor, Prof. Dr. Yusuf Sinan Akgül, for his endless support and guidance throughout this study. He always shared his experience generously when I needed in my academic and personal life.

I also would like to thank my committee members, Dr. Yakup Genç, who always spared time when I asked for his advice and Professor Songül Varlı Albayrak, who shared her insightful comments gently. I would like to acknowledge Professor Erchan Aptoula and Dr. Ayşe Betül Oktay for their kind participation.

My parents, Kevser and Fatih Şerbetçi, I would like to thank you for your endless love and emotional support. I appreciate your patience and guidance throughout my life.

Finally, I would like to thank my husband, Hasan Turan, who has always been my side, for his patience and never-ending spiritual support. This thesis cannot be completed without his efforts.

TABLE OF CONTENTS

	Page
SUMMARY	v
ÖZET	vi
ACKNOWLEDGEMENTS	vii
TABLE OF CONTENTS	viii
LIST OF ABBREVIATIONS AND ACRONYMS	xi
LIST OF FIGURES	xii
LIST OF TABLES	xiii
1. INTRODUCTION	1
1.1. Person ReID and the Challenges	1
1.2. Deep Learning for Person ReID	4
1.3. Contributions	9
2. REGULARIZATION AND ITS APPLICATIONS IN PERSON REID	11
2.1. Regularization	11
2.2. Regularization for Person ReID	13
2.2.1. Regularization via Data	13
2.2.2. Regularization via Multi-tasking	17
2.2.3. Regularization via Loss Constraints	21
3. ENSEMBLE LEARNING	25
3.1. Overview	25
3.2. Bootstrap Aggregation (Bagging)	26
3.3. Boosting	28
3.4. Mixture of Experts	30
3.5. Stacked Generalization	31
3.6. Cascading	32
3.7. Deep ensemble learning	33
3.8. Ensemble (or Multi-loss) Learning for Person ReID	34
4. DATASETS, EVALUATION AND THE BASELINE MODEL	37
4.1. Datasets	37
4.2. Evaluation Protocol	38

4.2.1. Cumulative Matching Characteristics 38
4.2.2. Mean Average Precision 38
4.2.3. Architecture 41
4.3. Baseline Model 42
4.3.1. Data Pre-Processing and Augmentation 42
4.3.2. Implementation Details 43
4.3.3. Results 43
5. END-TO-END CNN ENSEMBLES 45
5.1. Motivation 45
5.2. Problem Definition 45
5.3. The Proposed Method 47
5.4. Binary Hash Code Generation 52
5.5. Ranking with the Proposed Model 53
5.6. Experiments 54
5.6.1. Implementation Details 54
5.6.2. Comparison with the State-of-the-art 55
5.6.3. Further Analysis 59
5.6.4. Comparison with conventional ensemble 61
5.6.5. Hyper-Parameter Analysis and Ensemble of Ensembles 67
5.6.6. ResNet50 as the Backbone 68
6. END-TO-END BNN ENSEMBLES 74
6.1. Motivation 74
6.2. Experiments on Image Classification 75
6.2.1. Classification Performance 76
6.2.2. Stability Analysis 77
6.2.3. Robustness Analysis 78
6.3. Experiments on Person ReID 79
7. CONCLUSION AND FUTURE WORK 82
7.1. Conclusion 82
7.2. FUTURE WORK 83
7.2.1. Task-Specific Problems 83
7.2.2. Further Directions 85

REFERENCES 88

BIOGRAPHY 104
APPENDIX 105

LIST of ABBREVIATIONS and ACRONYMS

Abbreviations and Acronyms		Explanations
ReID	:	Re-Identification
CNN	:	Convolutional Neural Network
DCNN	:	Deep Convolutional Neural Network
GAN	:	Generative Adversarial Network
MoE	:	Mixture of Experts
UDA	:	Unsupervised Domain Adaptation
AP	:	Average Precision
mAP	:	Mean Average Precision
CMC	:	Cumulative Matching Characteristics
IDE	:	Identity Discriminative Embedding
FLOP	:	Floating Point Operation
BNN	:	Binary Neural Network
AB	:	All-Binary
SB	:	Semi-Binary

LIST of FIGURES

Figure No:		Page
1.1:	Person re-identification pipeline.	1
1.2:	Major challenges in person re-identification.	3
1.3:	The two main training objectives used in deep person ReID models.	5
1.4:	Training and testing stages in person ReID.	7
3.1:	Training process in Bagging.	26
3.2:	Inference process in Bagging.	27
3.3:	Training process in AdaBoost algorithm.	28
3.4:	Inference process in AdaBoost algorithm.	29
3.5:	Mixture of Experts model.	30
3.6:	Stacked Generalization.	31
3.7:	Cascading classifiers.	33
4.1:	Baseline IDE model.	41
5.1:	The architecture of the proposed system.	49
5.2:	Structure of the sub-networks.	51
5.3:	Normalized convolution feature maps as input to different learners.	52
5.4:	Base learner performances and cumulative ensemble performance.	64
5.5:	Comparison with conventional ensemble.	66
5.6:	Hyper-parameter analysis on Market-1501.	67
5.7:	Performance of ensemble of ensemble model.	68
5.8:	Adaptation of the proposed model on ResNet50.	71
5.9:	The model performance in different training stages.	73
6.1:	Adaptation of the proposed model to image classification.	76
6.2:	Stability comparison of our method with BNN baseline.	77
6.3:	Robustness comparison of our method with BNN baseline.	78
6.4:	Cumulative performance of classical ensemble and our ensemble.	81
6.5:	Convergence in performance when increasing ensemble size.	81

LIST of TABLES

Table No:	Page
4.1: Some person ReID datasets, ordered in release time.	40
4.2: Details on the datasets used in the experiments throughout this thesis.	42
4.3: Comparison of the baseline IDE model with the state-of-the-art.	44
5.1: Comparison with the state-of-the-art approaches in Market-1501.	56
5.2: Comparison with the state-of-the-art approaches in DukeMTMC-reid.	57
5.3: Comparison with the state-of-the-art on CUHK03 datasets.	58
5.4: Comparison with state-of-the-art methods on MSMT17 dataset.	59
5.5: Ablation studies of components of the proposed model.	63
5.6: Comparison with the state-of-the-art when ResNet50 is the backbone.	72
6.1: Comparison of our method with BNN baseline.	77
6.2: Comparison with classical ensemble of baseline BNN.	80

1. INTRODUCTION

1.1. Person ReID and the Challenges

The number of camera networks in public places is increasing dramatically because they are used for various purposes, such as security, surveillance, crowd analysis, and applications that involve human-robot interactions. In crime centers, the surveillance camera network is leveraged for tracking a suspect, while in airports and stations, it can be used for controlling and verifying the passenger transactions. Customer behavior can be monitored and analyzed in shopping centers. However, with human operators, detecting, tracking, or monitoring individuals in crowded areas for long durations is very laboring and inefficient. Also, the usage of these networks with a great variety of new purposes is emerging, which brings the requirement of automatic detection and identification of people.

Automatically recognizing an individual who has previously been observed over a camera network is Person Re-Identification (ReID). Figure 1.1 presents the process of a practical person ReID system [Web-1, 2020]. Given a query person image, the process involves extracting person bounding boxes from raw video frames captured from a camera network and retrieving the images of the query person.

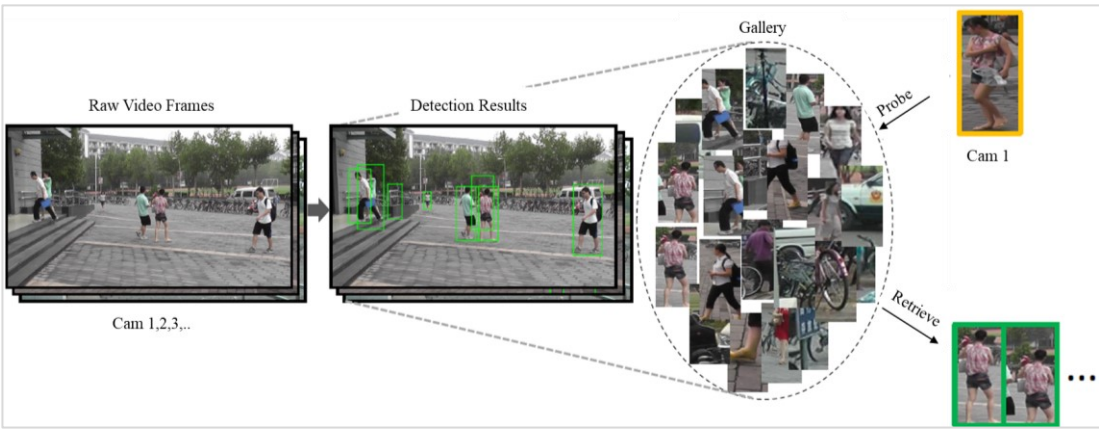

Figure 1.1: Person re-identification pipeline.

Differently from previous work, which aim to identify a person using biometric information such as face and gait [Wang et al., 2003], [Boulgouris et al., 2005]. Gheissari et al. [Gheissari et al., 2006] have first defined the appearance-based person ReID. The objective of appearance-based person ReID is recognizing a person who

has been observed within the same day, i.e., with the assumption of no changes in clothing. Since then, a considerable research effort has been made to improve the accuracy of this challenging task.

In the real scenario, a person ReID system determines whether a query person is in the gallery set or not, which is a critical and complicated decision process. Therefore, the researchers have mostly focused on the less complicated closed-world scenario, which presumes that the query person is guaranteed to appear in the gallery set [Zhu et al., 2018]. Then, the ReID problem reduces to finding the most similar person in the gallery image.

Other challenges in person ReID are mainly due to severe illumination, viewpoint, and pose variances caused by diverse imaging conditions and misalignments in person bounding boxes due to detection errors. In case of occlusion, low camera resolution, and similar clothing styles of distinct individuals, it becomes even more difficult to distinguish different persons from the images. Figure 1.2 illustrates some examples to these challenges from the widely used datasets. Despite these challenges, significant advances in the closed-world person ReID datasets have been obtained. The recently proposed techniques have surpassed human-level performance [Leng et al., 2019] on the relatively easy datasets, which are limited in their number of cameras and person identities. However, there are still many issues to be addressed for developing real-world person ReID systems.

The pioneering work for person ReID leveraged hand-crafted features such as SURF [Hamdoun et al., 2008], HSV histograms [Farenzena et al., 2010] and SIFT [Pedagadi et al., 2013]. After Deep Convolutional Neural Networks (DCNN) won the large-scale image classification task [Krizhevsky et al., 2012], the approaches for person ReID task have fundamentally changed as other pattern recognition tasks. Significant progress in person ReID has been gained via deep learning, where the research efforts mainly vary in their design of deep CNN architecture and the objective function [Ahmed et al., 2015], [Chen et al., 2017].

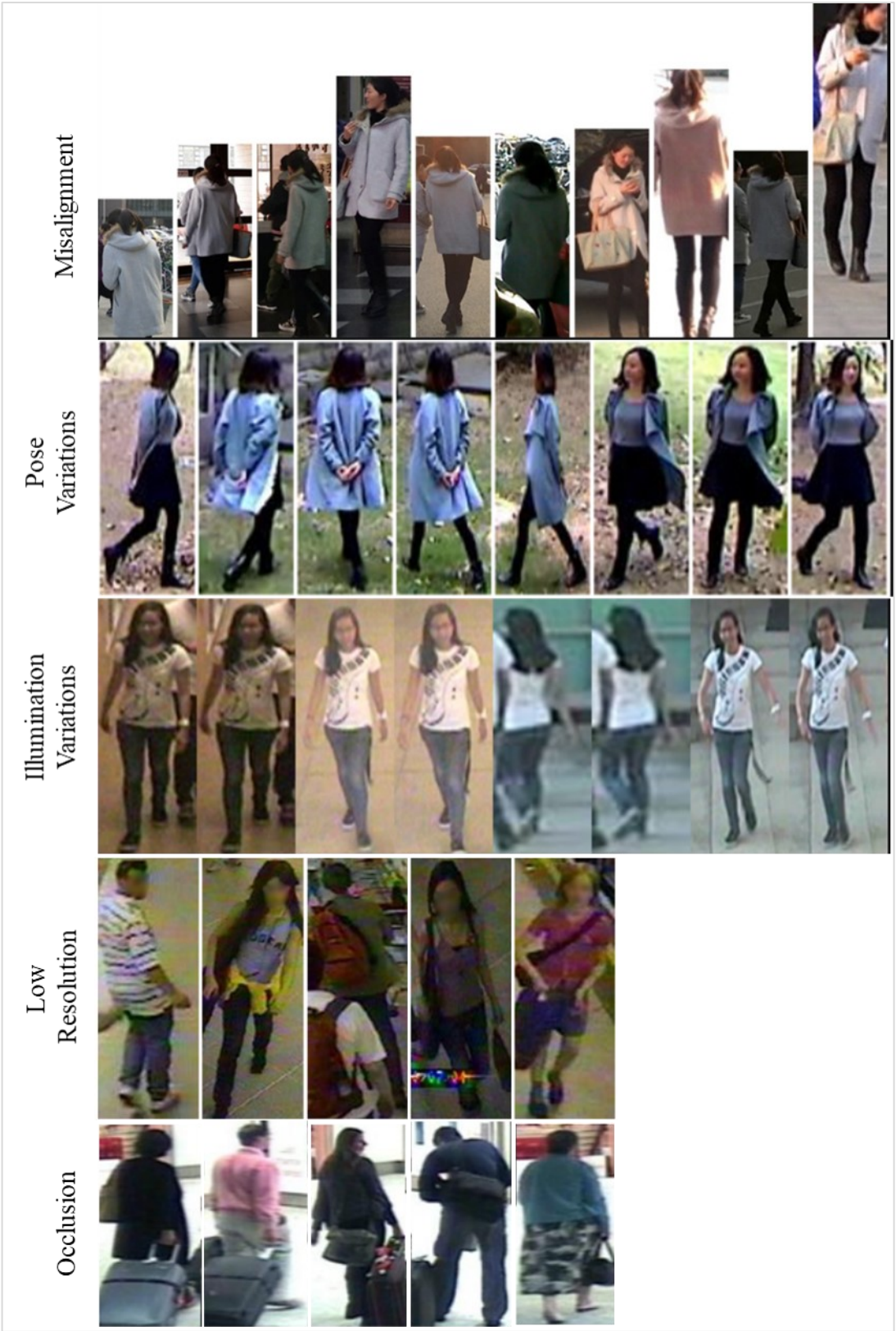

Figure 1.2: Major challenges in person re-identification.

1.2. Deep Learning for Person ReID

The pioneering work for deep person ReID designed custom and relatively shallow networks for metric learning [Li et al., 2014], [Yi et al., 2014], [Ahmed et al., 2015]. After training very deep networks has become feasible [Simonyan Zisserman, 2014], [He et al., 2016], [Ioffe and Szegedy, 2015] and large scale person ReID datasets have been released, the trend has changed to adapting deeper models and using classification objective to fully utilize the label information [Xiao et al., 2016a], [Zheng et al., 2016]. Although training such deep models requires a long time, pre-training on large scale image classification datasets enables faster training and relatively easy convergence [Geng et al., 2016], [Zheng et al., 2016], [Chen et al., 2017].

Person ReID literature can be reviewed from the perspective of the problematic part proposed to be solved. [Li et al., 2017a], [Zhao et al., 2017], [Zheng et al., 2018] aim to solve the misalignments in person bounding boxes. In order to capture the small details [Y. Chen et al., 2017], [Qian et al., 2017], [Qian et al., 2019] adapts multi-scale feature learning. Camera discrepancy presents another challenge for person ReID. [Zhong et al., 2018] proposes a data augmentation method and [Chen et al., 2018] handles camera correlations explicitly, to overcome this problem. In [Ye et al., 2016], the final ranking is refined by using similarity and dissimilarity ranking aggregations of two baseline models. A graph-matching framework is utilized to estimate cross-camera labels in [Y. Huang et al., 2019]. [Shen et al., 2018] proposed using conditional random fields for obtaining a consistent similarity metric.

Generally, deep learning research which leverages DCNN's for person ReID presents mainly two main approaches to the problem: metric learning [Cheng et al., 2016], [Chen et al., 2017b] and discriminative feature learning learning [W. Li et al., 2017], [Chen et al., 2017a]. Both methods aim at obtaining a feature extraction network, which embeds the input images to an appropriate space where the test time ranking is performed in. However, the training objectives of these two approaches are fundamentally different, which are shown in Figure 1.3.

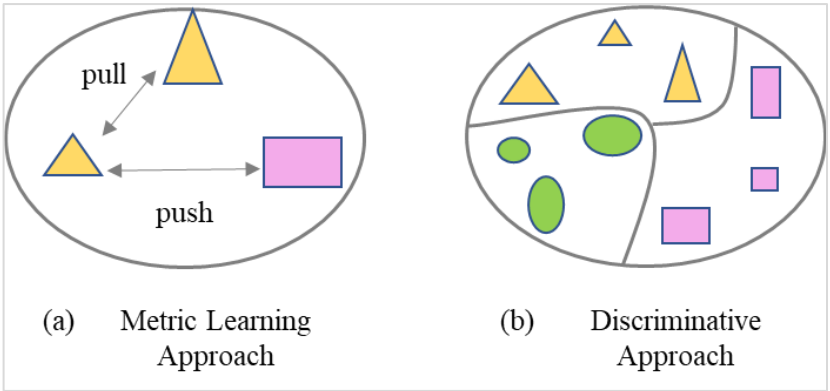

Figure 1.3: The two main training objectives used in deep person ReID models.

The objective of the metric learning approach is to map the input images to an embedding space, where images of the same person are close to each other, while the distances between the images of different identities are away. The discriminative approach, on the other hand, aims at finding decision boundaries between person classes without concerning the distances between the instances.

Figure 1.4 presents the schematic flow of these two approaches during the training and their typical test stage. Metric learning and discriminative learning are illustrated in the upper and lower branches, respectively. Additional steps for each approach are shown in dashed boxes. In the training stage of metric learning, each training sample is prepared generally as a triplet, which consists of an anchor, a positive (same person), and a negative (different person) sample. Then, a margin-based loss function is employed on the output of the feature extraction network as

$$\mathcal{L}(I_A, I_P, I_N) = \max\left(0, \|f(I_A) - f(I_P)\|^2 - \|f(I_A) - f(I_N)\|^2 + \alpha\right) \tag{1.1}$$

where $f(\cdot)$ is the feature extraction network, $\|\cdot\|^2$ is the Euclidean norm, α is the desired margin between the positive and negative image representations, and I_A, I_P, and I_N represents the images of the anchor, positive and negative persons, respectively. The metric learning approach is suitable for the test time ranking where the goal is to rank the gallery images with respect to their similarities to the query image. It is also advantageous in small scale datasets, since it increases the dataset size through triplet generation. However, there are some caveats for the training phase: the number of

triplets grows dramatically as the training data gets larger and most of the triplets add very little information as the training proceeds. Therefore, a considerable amount of

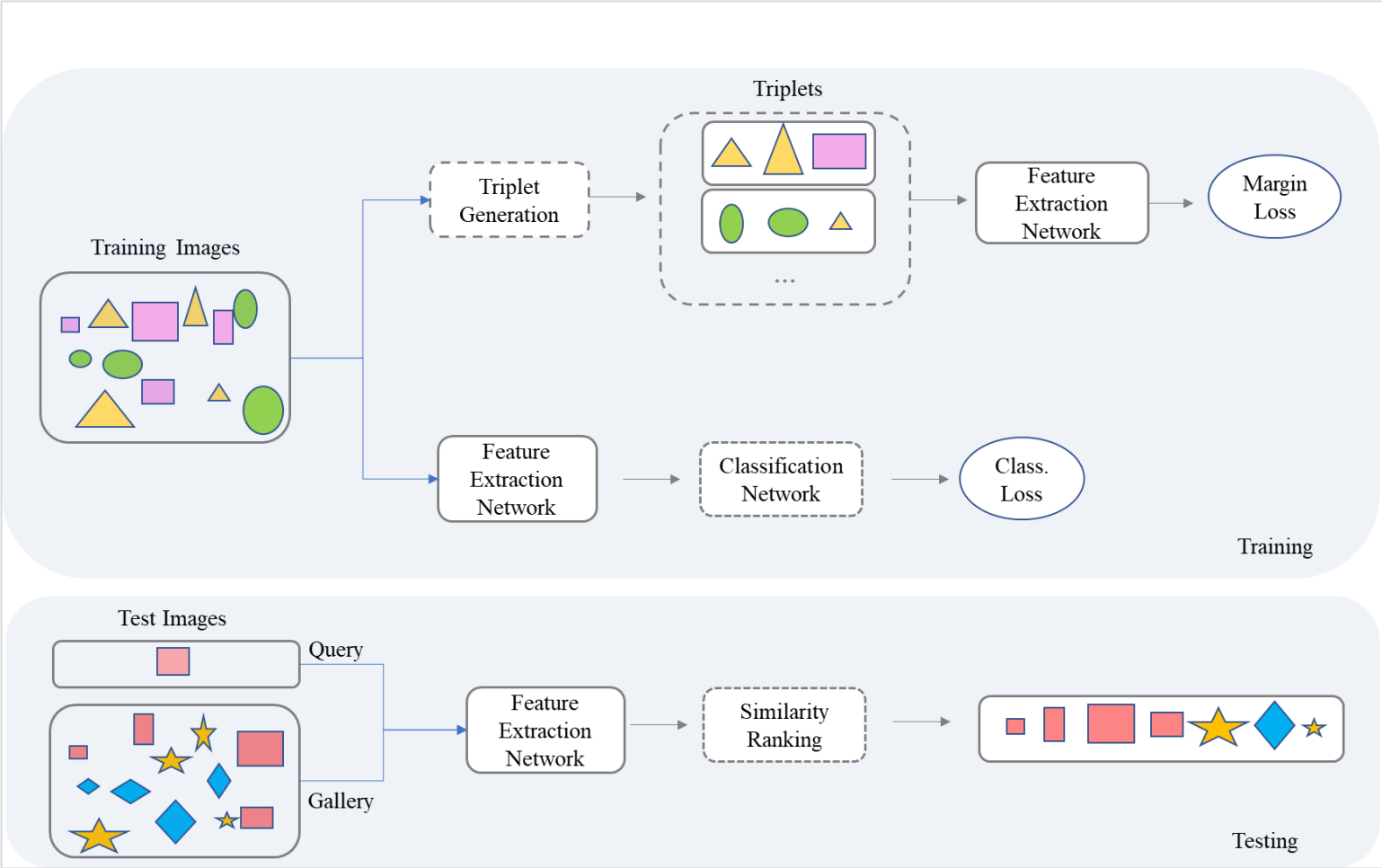

Figure 1.4: Training and testing stages in person ReID.

research has been dedicated to address the triplet generation problem [Balntas et al., 2016], [Oh Song et al., 2016], while some others suggest modifying margin-based loss functions [Cheng et al., 2016], [Chen et al., 2017b], [Hermans et al., 2017].

In the discriminative approach, on the other hand, an additional (usually one-layer) classification network is leveraged after the feature extraction network. Then, the network is trained to minimize the Cross-Entropy loss $\mathcal{L}$ between the predictions of the classification layer and the actual class labels as

$$\mathcal{L}(I_i, y_i] = -\sum_{k}^{C} y_i^k \log\left(f(I_i]^k \right) \tag{1.2}$$

where $f(\cdot)$ is the feature extraction network, I_i is the image of the ith training image, and y_i represents the actual probability that this image to be an instance of the kth person class. In contrast to metric learning, the training phase for this method is simple since the model does not require any laboring input preparation and converges relatively quickly. Besides, as opposed to the metric learning approach, it makes use of the full label information. The major drawback of discriminative models is that the training objective is not compatible with test time ranking. The classification network is discarded in test time, and the network is used as a feature extractor. More importantly, discriminative methods are biased to the training samples [Ranjan et al., 2017] and may induce high model variance on the test data, which is known as overfitting.

The critical point of overfitting is the well-known bias-variance dilemma pointed out by [Geman et al., 1992], in the early '90s. Briefly stated, the estimation error of a learning model can be decomposed into estimation bias and estimation variance. Bias represents how well the model predicts the actual values, while variance reflects how much the model is sensitive to specific sets of training data. There is a trade-off between the two terms. Therefore, the best model, which performs well on future data, should balance bias and variance. [Geman et al., 1992] also state that models that have a large number of parameters produce high variance (i.e., are more data-dependent). Bias and variance of the learning models are major subjects of statistics and especially of estimation theory. Differently from the statistics terminology, the machine learning

community frequently uses the term "overfitting" (or under-regularization)[1] to indicate the problem of having high variance. Therefore, every effort for addressing the overfitting problem deals with reducing the variance while preserving the low bias.

In order to decrease the effect of overfitting in deep networks, many regularization methods have been proposed, which have a crucial role in the impressive performance of these models. The regularization techniques such as weight decay, dropout, data augmentation and pre-training on large-scale image classification datasets are indispensable for obtaining a reasonable performance not only for person ReID, but also for other tasks.

Person ReID is more prone to overfitting due to two reasons: First, person ReID datasets are in relatively modest sizes, which hinders model generalization. The number of identities and cameras are very limited in the existing datasets compared to other tasks. Therefore, there is a large domain gap among person ReID datasets. Second, the training and test identities do not overlap; that is, the trained model is used to identify unseen person identities, which causes drastic performance drop during the test. As a result, for practical person ReID systems, developing effective regularization methods is of crucial importance.

1.3. Contributions

This thesis aims to design a novel regularization method for Person ReID. To this end, we utilize from ensemble learning, a well-known technique for improving the generalization capacity. Basically, ensemble learning is combining the decisions of multiple different and accurate base learners trained for the same task to make a final inference (e.g., classification). Our aim is to aggregate the performance of deep networks with the generalization ability of ensemble learning to build a novel regularization method for deep person ReID models.. However, training multiple deep networks as the base learners is inefficient. The present study proposes a simple yet effective ensemble model that is end-to-end trainable and adds negligible computational cost during both the training and inference times.

---

[1] We use overfitting and under-regularization interchangeably throughout the thesis.

Our contributions are:

- We provide a comprehensive survey of existing person ReID models categorized by their regularization schemes;
- We propose an end-to-end ensemble learning method for discriminative person ReID models to reduce the effects of overfitting;
- We obtain accurate and diverse base learners so that when their individual feature representations are combined in test time, they improve Rank-1 and mean average precision (mAP) scores by a large margin. We achieve state-of-the-art results on several large-scale benchmark datasets;
- Our approach is very efficient in both training and inference times compared to the conventional ensemble methods;
- We avoid the custom design of network architecture specialized to ReID. The proposed method requires minimal changes in DenseNet architecture and is not task-specific;
- Our method is applicable to other tasks as a general regularization method. We experimentally show that application of the proposed approach to Binary Neural Networks improves the test accuracy, training stability, and robustness to input perturbations.

The rest of this thesis is structured as follows: In Section 2, we provide a background for regularization in learning systems and review its applications in person ReID literature. In Section 3, we discuss the ensemble learning methods that are widely used. Person ReID datasets, the evaluation metrics, and a baseline model are given in Section 4. Also, we propose the evaluation of the baseline model on the three widely used benchmark datasets. In Section 5, the proposed end-to-end ensemble learning method is introduced, which addresses the overfitting in discriminative person ReID models effectively. To demonstrate its generalizability, we adopt the proposed model to Binary Neural Networks and share some initial findings in Section 6. Lastly, conclusions and possible further research directions are given in Section 7.

2. REGULARIZATION AND ITS APPLICATIONS IN PERSON REID

Regularization has always been an essential constituent in learning models to improve the test time performance. Besides the general techniques which aim to control the model complexity, there are many other task-specific regularization methods based on data augmentation, multi-tasking, and objective function design. In this section, we first provide a background for regularization, and then give a broad review of regularization techniques applied in deep person ReID models.

2.1. Regularization

In statistical decision theory, a learning network approximates a function from a set of representative observations [Barron and Barron, 1988], [Vapnik, 1999]. Given a set of observations X from an unknown distribution p and an approximating function $f(\cdot)$, the learning process corresponds to minimizing the empirical risk $\mathcal{R}(\cdot)$, which is the expectation of the loss function $\mathcal{L}$ specified for the task:

$$\mathcal{R}\big(f(\cdot)\big) = \mathbb{E}_{X\sim p}[\mathcal{L}(f(\cdot), \mathrm{X})]\,. \tag{2.1}$$

Given a neural network architecture, $f(\cdot)$, the empirical risk is minimized via optimizing the network parameters w:

$$\mathrm{w}^* = \arg\min_{\mathrm{w}\in\mathrm{W}} \mathcal{R}\big(f(\mathrm{w})\big) \tag{2.2}$$

where corresponds to set of all possible values for network parameters. However, it is not guaranteed that the function $f(\mathrm{w}^*)$ minimize the generalization error, which is the expected risk over p, the real distribution [Guyon et al., 1992]. Moreover, if $f(\mathrm{w}^*)$ has a high complexity, it may overfit to the observed examples and perform poorly on the unseen data points, a phenomenon called bias-variance dilemma [Geman et al., 1992]. Therefore, finding the optimal trade-off between the high bias of a model which is too inflexible and the high variance of a model with too much freedom is of crucial

importance for model generalization generalization [Geman et al., 1992], [Bishop, 1995].

Regularization techniques are used for imposing further smoothness constraints on the approximating functions to balance this trade-off [Girosi and Poggio, 1995]. Generally, the model is penalized using a regularization term, which is a function of model complexity. Then, the regularized risk is defined as

$$\mathcal{R}_{reg} = \mathcal{R} + \lambda\phi(f) \tag{2.3}$$

where λ is the regularization weight.

The most widely used regularization technique is weight decay, which constraints the growth of weights with L2 norm penalty term as $\phi(f) = \|\mathrm{w}\|_2^2$. Lasso is another technique that uses L1 norm penalty term $\phi(f) = \|\mathrm{w}\|$. Weight decay avoids weights to become too large, thus results in a smoother decision boundary. On the other hand, Lasso performs feature selection by promoting sparse representations. However, regularization techniques are not limited with these general techniques of controlling model complexity. Goodfellow et al. [Goodfellow, Ian and Bengio, Yoshua and Courville, 2016] defines the term regularization as "any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error". Dropout, batch normalization, and data augmentation are only a few of the general regularization methods which have become vital components in training deep neural networks. A taxonomy of regularization methods in deep networks and a comprehensive review can be found in [Kukačka et al., 2017].

Besides the general regularization techniques which are incorporated in the optimization or the training procedure, many regularization methods can be designed depending on the task in hand. In fact, based on the above-mentioned definition of Goodfellow et al. [Goodfellow et al., 2016], every method improving test time performance can be considered as a regularization technique. When it comes to deep person ReID models, considering the fact that almost all of the current state-of-the-art methods employs similar backbone architectures and follow a training procedure with the same optimizers and hyper-parameters, it is the regularization scheme which differentiates the research attempts from each other. Although researchers propose

solutions from different perspectives, most of the recent methods can be categorized into a set of regularization schemes. In the next section, we review the regularization techniques used in deep person ReID models, which constitute the mainstream of the current approaches.

2.2. Regularization for Person ReID

2.2.1. Regularization via Data

Data augmentation is an essential step for improving generalization in deep models. Besides basic data augmentation used in most ReID models such as random horizontal flipping and random cropping, there exists more advanced techniques for data-based regularization such as transfer learning from auxiliary tasks or datasets, data augmentation by randomly erasing part of the images, combining multiple datasets for the same task and data augmentation via unlabeled images. Below, we briefly review the usage of these methods for person ReID.

2.2.1.1.Transfer Learning

In machine learning, transferring learned information from a source dataset or task to another is called transfer learning. In deep neural networks this is generally accomplished by re-using part of a network pre-trained for an auxiliary task and finetuning its weights by retraining it for the new task. Transfer learning is the most widely used technique to increase the generalization ability of deep architectures not only for person ReID but also for other vision [Sermanet et al., 2014] and machine learning tasks [Yang et al., 2017].

In most state-of-the-art person ReID methods, instead of designing custom networks and training them from scratch, the general-purpose architectures pre-trained on large scale image dataset (i.e., ImageNet [J. Deng et al., 2009]) are used, such as ResNets [He et al., 2016] and DenseNets [Huang et al., 2017]. Generally, deep ReID methods re-train these networks by changing the task-specific classification layers with an appropriately designed layer which is randomly initialized [Geng et al., 2016]. In order

not to lose the learned information (i.e., weight values) and allow smooth adaptation to the new task, a much smaller learning rate is used in these pretrained layers. Small datasets, however, cannot generalize well even with knowledge transfer from ImageNet dataset. In order to deal with the large domain gap, [Geng et al., 2016], a two-step finetuning with an additional auxiliary dataset is suggested. Specifically, the ImageNet pretrained network is first finetuned with a large scale ReID dataset and the resulting network is finetuned on the small-scale target dataset.

A potential means for transferring knowledge to the ReID task is employing off-the-shelf architectures trained for some related tasks, such as human body part detectors or semantic parsers. One of the first attempts which incorporate human body part detection for feature extraction is [H. Zhao et al., 2017], which employs Region Proposal Network trained on MPII human pose dataset [Andriluka et al., 2014]. The designed network localizes the seven human body parts and competitively fuses the respective feature vectors. [J. Guo et al., 2019] detects human body parts using a human parsing model trained on Look-Into-Person [Liang et al.,, 2018] dataset. [Zheng et al., 2019] proposes PoseBox for pedestrian alignment, which performs pose estimation using Convolutional Pose Machines followed by affine transformation. To minimize the effects of pose estimation errors, they design a CNN which discriminates identities based on the original feature, PoseBox output and the confidence score of the pose estimation.

2.2.1.2.Data Augmentation via Intentional Occlusion

One of the major difficulties in person ReID arises from occlusion. Surveillance systems are generally used in crowded places such as airports, shopping centers, and public squares, which causes the detected person images to be highly occluded by other persons or static objects [Zhuo et al., 2018]. Therefore, many methods have been offered which augment the training data with intentionally occluded images in order to prevent overfitting and to enable invariance to occlusion.

One of such methods is Random Erasing data augmentation, which randomly selects a rectangle region and erases its pixels with random values [Zhong et al., 2017]. Its effectiveness is shown on object detection, classification, and person ReID. Random erasing data augmentation is used almost all of the current state-of-the-art methods [Wang et al., 2018], [Shen et al., 2018].

Another method aims to discover the critical image regions where a trained model presents sensitivity to occlusion [Huang et al., 2018]. Then, they generate adversarially occluded images and retrain the model with both the original and the occluded images to obtain a more generalizable model.

Instead of directly augmenting data using the raw images independently, [Dai et al., 2019] proposes another augmentation technique, which is performed on the feature maps extracted in the deeper layers of the network. This simple method drops the same regions of the feature maps of all images in a single batch, which enforces the network to search for alternative discriminative parts.

2.2.1.3.GAN-based Methods

Data augmentation using GANs is used with different purposes in person ReID: for example improving the performance in the supervised learning setting, creating an artificial, labeled dataset in the unsupervised setting, and incrementing the available data in semi-supervised data. Below we review the GAN-based methods in person ReID under these three learning settings.

Supervised Setting: One of the pioneering work which leverage GAN's in person ReID is [Zheng et al., 2017], which uses the GAN generated images “in vitro”. Specifically, they generate new unlabeled images from target dataset using GAN and later label these images with a uniform label distribution to obtain the augmented training data. Zhong et al. [Zhong et al., 2018] proposed to augment the training data by translating images to other camera styles within the target dataset using CycleGAN [Zhu et al., 2017]. They aim to eliminate the noise introduced by the artificial images using label smoothing regularization [Szegedy et al., 2016] while labeling the generated images. In the supervised setting, GAN's are also leveraged for pose invariance within the same dataset. In [Qian et al., 2018], GAN is conditioned on both the input image and one of eight canonical poses to pose-normalize the training images. [Ge et al., 2018] designs a siamese architecture which receives an image pair with different poses and employs a GAN for creating an image pair with the identical input target pose. The objective is to represent only the person identity in the generated images. Another method [Liu et al., 2018] extracts various human poses from MARS dataset and generates pose translated images for the target dataset. To preserve the person identities, they further guide the generator

by integrating an off-the-shelf person identity discriminator. Low resolution of input images is another problem in ReID. To address this problem, [Wang et al., 2018] proposes employing Super-Resolution GAN (SRGAN) in a cascaded manner to obtain a scale-adaptive model. Zheng et al. [Z. Zheng et al., 2019] propose a unified model for image generation and ReID learning. Specifically, they separately encode the appearance and structure of the person and generate cross identity images using these codes. This alleviates the need for auxiliary datasets or target pose conditioning. Wei et al. [Wang et al., 2019] focus on multi-modal person ReID problem and claim that discrepancy between modalities can be reduced using a unified space. To this end, they exploit CycleGAN to convert the images from one modality to another, thus obtain multi-spectral images.

Unsupervised Setting: A trained person ReID model must be adaptable to new datasets with minimal cost. However, there is a huge domain gap between the ReID datasets due to varying imaging conditions and person identity classes. Therefore, severe performance degradation is inevitable when the trained models are directly deployed in new domains. To bridge the domain gap between person ReID datasets, unsupervised domain adaptation (UDA) techniques leverage GAN's to translate images from the source domain to the target domain. Wei et al. [Wei et al., 2018] follow this approach and further design an identity loss function to preserve the person identities during the translation. To calculate the identity loss, they evaluate the foreground variations of the images before and after the translation. [Deng et al., 2018] utilizes CycleGAN [Zhu et al. 2017] for transferring images from the source to target dataset. They argue that the transferred images should be dissimilar to any of the target identities, which they call domain dissimilarity. They embed an appropriate contrastive loss to the objective function to meet the self-similarity and domain dissimilarity requirements.

[Liu et al., 2019] argues that different factors should be handled explicitly in image translation. They follow a divide-and-conquer approach and perform style-transferring of illumination, resolution and camera-view separately and design a selection network for adaptive fusion of the transferred images. [Zhou et al., 2019] proposes to style transfer images to each camera of the target domain.

Semi-supervised Setting: The vast amount of available data together with the difficulty of obtaining label information has led many research studies to investigate the ways of utilizing from unlabeled data [Odena, 2016]. Unlabeled data is generally used to enhance the supervised learning process, which is called semi-supervised learning. Since deep

neural networks always improve with more data, a considerable amount of research has been dedicated to investigating the ways of improving the model performance employing a semi-supervised setting.

In recent approaches for person ReID, the unlabeled dataset is generally obtained via Generative Adversarial Networks. The pioneering work is [Zheng et al., 2017], which produces unlabeled images to augment the training data for supervised learning. In order to label the generated images, they follow a similar approach to the well-known label smoothing regularization method and assign a uniform label distribution over the training classes. It is shown that unlabeled images improve the discriminative performance of the baseline model even if they are not assigned any specific class label. However, [Huang et al., 2019] claims that assigning a uniform label distribution to all of the generated images is misleading, since only certain real data from some classes are used in GAN to generate artificial data. Inspiring from pseudo-labeling [Lee, 2013], which assigns the dynamically predicted class probabilities to the unlabeled samples, [Huang et al., 2019] proposes a weighted pseudo labeling scheme for incorporating unlabeled images in the supervised training. [Ainam et al., 2019] argues that assigning uniformly distributed labels causes over-smoothing. They first cluster the training data and define a separate GAN for each cluster. Then, the generated images from these separate GANs are uniformly labeled with person identities that belong to the corresponding cluster.

2.2.2. Regularization via Multi-tasking

2.2.2.1. Part-based models

The visual analysis of deep ReID models demonstrates that they mostly activate the most discriminative part of a person image [Fu et al., 2019], which is generally the upper body part [Yao et al., 2019], [Huang et al., 2018]. This behavior hinders the model from extracting visual cues from other parts of the images, such as head and feet. Obviously, this can easily degrade the test time performance in case of occlusion or uncertainty on this part of the images. In order to reinforce the deep model to consider diverse attentive features, part-based models generally design separate branches for different parts and use explicit supervision for each. The total part loss is calculated as:

$$\mathcal{L}_P = \sum_{i}^{M} \mathcal{L}_i \tag{2.4}$$

where M is the number of parts and $\mathcal{L}_i$ is the loss for the ith part. Some methods use an additional global branch for further regularization, which is shown to be effective [F. Yang et al., 2019]. Then the total is calculated as

$$\mathcal{L}_T = \mathcal{L}_G + \lambda \mathcal{L}_P \tag{2.5}$$

where $\mathcal{L}_G$ is the loss of global branch and λ is the weight for the part loss. These models vary mostly in the strategy of partitioning the image into smaller regions. We review the part-based models regarding to their partitioning strategy.

Pre-defined Partitioning: The simple method to extract local features is to partition the output feature maps of the backbone network into horizontal stripes and define multiple loss functions based on this partition. In [Li et al., 2017], feature vectors obtained from each part are fused for the classification layer. A separately supervised global branch together with the local part branch constitutes the multi-task framework. [Sun et al., 2018] supervise each branch separately instead of the fused feature vector, thus reinforce the network to extract discriminative features from each part. In [Yao et al., 2019] discriminative parts are automatically extracted and identification loss is adapted on both global and part feature vectors. [Wang et al., 2018] and [Fu et al., 2019] follow a very similar approach and divide the person image into horizontal stripes with multiple granularity and extract discriminative features from each region. [Wang et al., 2018] also employs triplet loss, while [Fu et al., 2019] proposes using both average and max pooling layers to enhance the performance. [Sun et al., 2019], employs the same splitting approach for partial-ReID problem, where the images may contain partial observation of a pedestrian. The main objective is to learn visibility-aware features to consider only the shared regions across images. They utilize self-supervision mechanism for producing visibility scores for predefined regions, which are uniform splits of the holistic image.

Although some ReID datasets have human annotated bounding boxes [Douglas and Tao, 2008] and [Hirzer et al., 2011a], in some recently released datasets the pedestrian bounding boxes are automatically detected [Zheng et al., 2016], [Wei et al.,

2018], which is closer to the real scenario. The errors in the auto-detected bounding boxes and viewpoint/pose changes result in misalignment between person images [Suh et al., 2018]. As a result, the rough image partitioning method described above may result in inferior performance in case of severe occlusion.

Dynamic Partitioning: In order to deal with the misalignment issue while extracting the part-based features, another line of research proposes to localize the human body part regions automatically. In [Miao et al., 2019], a human pose estimator is employed for detecting human landmarks. The architecture is composed of two parallel branches: a global branch which is guided by the detected landmarks and a fixed part-based branch. The feature vectors extracted for each landmark are combined via average pooling in the global branch and identification loss is used for both the global branch and predefined part branches. [Zhang et al., 2019] utilizes DensePose [Güler, et al., 2018] model to segment the pedestrian images to surface-based body part regions, thus aligns the body part regions across images. However, the dense semantic representation of the pedestrian images can be erroneous due to lack of labeled data for person ReID. Therefore, the dense semantic alignment stream is used only during the training as a regulator for the main stream. The two streams are jointly optimized via combination of their local and global feature vectors.

2.2.2.2.Multi-scale Models

Adopting multi-scale structures to deep neural networks dates back to 2014 for many vision tasks, such as face verification [Sun et al., 2014], fine-grained image similarity learning [Wang et al., 2014] and depth estimation [Eigen et al., 2014] to name a few. One of the first multi-scale approaches is proposed by Chen et al. [Chen et al., 2017], who designs an end-to-end architecture for feature learning. More specifically, they propose a consensus learning strategy to co-train two networks, each working on different input scales. Inspired by knowledge distillation from neural nets [Hinton et al., 2015], the networks are guided by a larger one, which, in fact, is composed of their combination. This enables the combined network to compensate for the errors of individual networks. The requirement of separate networks for each scale avoids the model to be extended to more scales. Another method is proposed in [Qian et al., 2017], where each convolution layer consists of 4 streams of convolutional filters

each having different scales. The outputs of these streams are concatenated to be fed as input for the next layer. In the last layer, a soft attention mechanism is applied to weight the importance of different scales. A concurrent study is published by Li et al. [Li et al., 2017]. The most distinguishing aspect of this method is adopting Spatial Transformer Networks [Jaderberg et al., 2015] with some prior constraints to appropriately localize and transform the deformable human body parts. Then, they extract both local and global features via specially designed multi-scale context-aware network. They call their network context-aware because they stack multi-scale convolution outputs in each layer. Instead of using larger filter size, they employ dilated convolution to capture the context information. Except from [Chen et al., 2017], the above-mentioned methods generally design custom networks which are trained from scratch in a long time.

More recent studies, on the other hand, can finetune pre-trained very deep networks on relatively large-scale person ReID datasets. [Wang et al., 2018] proposes to utilize from global feature vectors extracted from multiple resolutions along the convolutional blocks of the backbone network. The network is trained with a metric learning objective on both the feature vectors alone and their weighted combination, resulting in a multi-tasking architecture. In this way a scalable model is obtained which is capable of cutting the feature extraction process in early layers in case of resource constraints. [Chang et al., 2018] proposes a factorization network to model the latent discriminative person attributes at multiple network levels. In each level, a factor selection module determines the activation of certain factor modules, each representing the latent attributes. The final layer combines the global feature vector and the factor signature vectors from all levels to discriminate person identity. In this way, the decision is made by considering both the high-level semantic feature vector and latent attributes from all levels. The overall network is partially supervised for each training image since some of the factor modules are deactivated. Therefore, although the model is supervised on a single fused feature vector, we consider the proposed method under multi-tasking architectures. [Qian et al., 2019], improves their previously proposed model [Qian et al., 2017] in multiple ways. First, they employ an attention learning layer to weight discriminative features from different scales. For this purpose, they concatenate the filter maps from different scales and integrate a self-attention module based on this multi-scale feature map.

2.2.2.3. Ensemble Models

Ensemble models consist of multiple base learners which are accurate but diverse from each other. The combined decisions of diverse learners is known to improve the test time performance. In deep person ReID, many multi-branch part-based methods are proposed (See 2.2.2.1), which mostly works in the same way as ensemble models. Apart from these part based approaches, one of the first studies in ensemble learning for person ReID is proposed by [Chen et al., 2017]. Although the main purpose is to extract scale-specific features, the proposed architecture is an ensemble of two base learners and is trained in an end-to-end manner. [Wang et al., 2019] proposes a multi-branch architecture to obtain an efficient ensembling model from ResNet backbone. Specifically, after obtaining the high-level feature maps from the backbone network, they partition the feature map into horizontal stripes in multiple times each time with different number of parts, and supervise each branch separately. The test time ranking is performed on the fused features from different parts. Instead of splitting the images spatially, [Zhai et al., 2019] proposes channel-wise split of the feature maps. They integrate separate identity classifiers resulting a multi-branch architecture.

2.2.3. Regularization via Loss Constraints

The objective function is one of the most appropriate constituents in deep networks, which enables explicit regularization. In addition to the generic regularization terms, such as weight decay and Lasso, the nature of the task-based loss function itself can properly regularize the network. In this section, we review deep ReID models which propose new objective functions or improving the existing ones.

2.2.3.1. Improved Triplet Loss

Considering it as an image retrieval task, several attempts have been made to design objective functions for metric learning in person ReID. Since its introduction by Schroff et al. for face recognition [Schroff and Philbin, 2015], triplet loss is used

frequently for obtaining a metric space in person ReID [Su et al., 2016]. By adjusting the relative distances of images, the triplet loss aims to produce an embedding space where the images of the same identity class are closer to each other than the images of any other identity classes. To this end, the training data is prepared as triplets, where each triplet contains an anchor, a positive and a negative sample.

$$\mathcal{L}(I_A, I_P, I_N) = \max\left(\|f(I_A) - f(I_P)\|^2 - \|f(I_A) - f(I_N)\|^2 + \alpha, 0\right) \qquad (2.6)$$

The preparation of the training data as triplets results in a cubic increase in the training samples, which leads to a problem to be addressed: most of the triplets already satisfy the correct ordering and are not informative, resulting long training time with sub-optimal solutions. The limitations of the original triplet loss have promoted the researchers to improve its effectiveness and making it more applicable. In their prominent work, Hermans et al. [Hermans et al., 2017] proposed batch hard mining strategy. Unlike previous approaches which perform hard mining from the whole possible triplets, batch hard mining considers only the current mini-batch for the negative and positive pairs. In this way, it avoids risk of overfitting to the outliers, thus increase the generalization ability and overcome the computational overhead of conventional triplet mining strategy.

A quadruplet loss is designed in [Chen et al., 2017b] to reduce the intra-class variation as a complementary approach to the triplet loss, which aims to increase the inter-class variations. To accomplish this, they sample quadruplets instead of triplets, and add a new constraint to the objective function which pushes away negative pairs from positive pairs w.r.t different probes. Therefore, the model learns the correct ordering with respect to multiple probes. [Xiao et al., 2017] embeds the hard batch sampling strategy of [Hermans et al., 2017] to the quadruplet loss, and mine hardest positive and hardest negative pair in each batch, without the restriction of using the same anchor for the negative pair. In this way, in each iteration, the bounds of two classes are pushed away. [Yuan et al., 2019] proposes an improved triplet loss function which makes use of all relationships in a triplet instead of calculating the loss w.r.t a single anchor. Zhang et al. [Zhang et al., 2019] proposed a training strategy for incremental margin loss. After training a base network with conventional triplet loss, they introduce feature shifts on top of mid-level blocks of the backbone network,

recursively. The mid-level feature shifts are trained with an increased margin loss, allowing a refinement on the feature vector.

2.2.3.2.Improvements on Cross-Entropy Loss

Cross-entropy loss is employed in person ReID in the discriminative approach, where the network aims at classifying the images to the correct person identities. Given an input image x_i and its expected label distribution t_i the cross-entropy loss calculates the divergence between the predicted and the real class distributions:

$$\mathcal{L}(x_i, t) = -\sum_{j=1}^{K} t_i^j \log\left(p(y_j|x_i)\right) \tag{2.7}$$

where K is the number of classes and $p(y_j|x_i)$ represents the network output which is converted to a probability by the softmax function:

$$p(y_i|x_i) = \frac{\exp\left(y_i|x_i\right)}{\sum_j \exp\left(y_j|x_i\right)} \tag{2.8}$$

The discriminative approach makes use of full labels through this objective and generally performs well on large datasets.

One of the modifications on the discriminative objective is called label smoothing regularization (LSR), which was proposed by Szegedy et al. [Szegedy et al., 2016] for large scale image classification. LSR avoids network from being too confident on its outputs by smoothing the ground-truth labels. The smoothed distribution is

$$\hat{t}(k|x) = (1-\epsilon)t(k|x) + \epsilon\frac{1}{|K|} \tag{2.9}$$

where ϵ is the smoothing strength. For person ReID, a variant of LSR is first utilized in incorporating the GAN-generated images into training in [Zhong et al., 2018]. The cross-entropy loss for unlabeled samples are calculated based on a uniform target probability distribution, i.e, $t_i^j = \frac{1}{|K|}, \forall j \in K$. LSR is now employed by many state-of-

the-art ReID models as a generic regularization method [Liu et al.,2018], [Luo et al., 2019], [Chung et al., 2019].

[Fan et al., 2019] point out softmax loss is biased towards weights and features having large norms. That is, 1) a sample with the same angular distance to two different classes tends to be classified into the class with the larger norm, and 2) the samples having a larger norm outputs a larger score, which they call weight and feature bias, respectively. Therefore, they propose defining the softmax loss on a sphere by normalizing both the weights and features, which reduces the uncertainty of decision process. They use a scale factor for controlling the temperature of the softmax function. A similar work is proposed by [Wojke and Bewley, 2018], which also normalizes the weights and features. [Wu et al., 2019] extends the angular loss to a view-point aware loss, where the distance calculation is done between the sample features and their identity and viewpoint centers, in a unified hyper-sphere.

In [Wu et al., 2019], a softmax-like loss function is designed to learn a discriminative embedding for unlabeled images in a semi-supervised setting. Since the sample classes are not known, the learning objective is to increase the cosine similarity between distinct images. The proposed exclusive loss approximates this objective by calculating the softmax probabilities based on a projection matrix consisting of the normalized feature vectors, which is updated as the training proceeds. In [Lin et al., 2020] the same approach is used for fully unsupervised training. In [Lin et al., 2020] a similar softmax-like function, called *repelled loss*, is designed which uses the sample-to-cluster distance to measure the cluster membership probabilities in unsupervised learning of camera invariant features.

[Ye et al., 2020] proposes a hard-aware instance re-weighting strategy for improving discrimination ability when the data contains noisy label. The re-weighting term uses the distance of the sample to its class center as a measure of hardness level.

3. ENSEMBLE LEARNING

In this section, we first briefly explain ensemble learning as a regularization method. Then we review the most common ensembling techniques as well as give some recent applications to deep networks.

3.1. Overview

The estimation error of a learning model can be decomposed into *bias* and *variance* terms, as discussed in Section 1. In order to increase the generalization capability of the learning model, variance should be decreased. However, there is a trade-off between the two terms and reducing variance generally results in increased bias, which degrades the accuracy of the learning models. The problem of having high variance is called overfitting, which means the model has too much degrees of freedom, and over-fits to the data. The trade-off between the bias and variance terms make overfitting (under-regularization) problem is a challenging task.

Ensembling has been one of the most effective approaches to tackle the under-regularization issue [Meir, 1995], [Breiman, 1996b]. Ensembling is the process of training multiple base learners for the same task and combining their decisions in inference time. The base learners are trained either in parallel or sequential. If the base learners are accurate and diverse from each other, they make errors on separate parts of the test data, which provides error compensation. Therefore, the performance of the ensemble models strongly depends on the individual accuracy of the base learners and their diversity. As a result, many studies have been dedicated to increase model diversity without compromising accuracy [Opitz and Shavlik, 1996], [Granitto et al., 2005], [Melville and Mooney, 2005].

Many methods have been offered for enforcing diversity between base learners as well as for combining their decisions. Generally, the base learners vary in their algorithms, hyperparameters, input representations (modalities), training sets and sub-problems to fulfill this requirement.

There are multiple approaches in designing ensembles, namely bagging, boosting, mixture of experts, stacked generalization and cascading. The approaches mainly vary in their algorithm for combining models, training base learners and data

usage. In this section, we review these main approaches and give some example algorithms for each that are widely used.

3.2. Bootstrap Aggregation (Bagging)

Bootstrapping is a technique in statistics used for improving the standard error and confidence interval estimation. Basically, it produces new samples from a given sample by replacement and performs estimation based on the bootstrap samples. Bootstrap aggregation, or bagging in short, is an ensemble method which aggregates the decisions of base learners trained with different bootstrap replicates of a dataset [Breiman, 1996a]. Figure 3.1 presents the training procedure in Bagging, for training each base learner L, a different bootstrap sample is used.

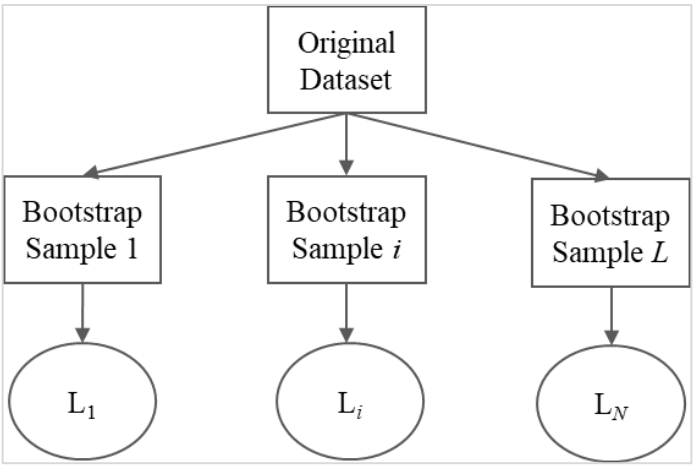

Figure 3.1: Training process in Bagging.

Since the bootstrap samples are performed with replacement, some of the observations will occur more than once in the drawn sample while some others will not occur [McCue, 2006]. Specifically, each observation has the following probability of being selected for a bootstrap sample

$$1 - (1 - \frac{1}{n})^n \tag{3.1}$$

where n is the sample size. For sufficiently large n, this converges to

$$1 - \frac{1}{e} \cong 0.63 \tag{3.2}$$

which means 37% of the observations will be missing in each sample.

The combination of the base learners during inference can be performed in different ways. The most widely used methods for regression and classification are aggregating the averages over the outcomes of the base learners and majority vote, respectively.

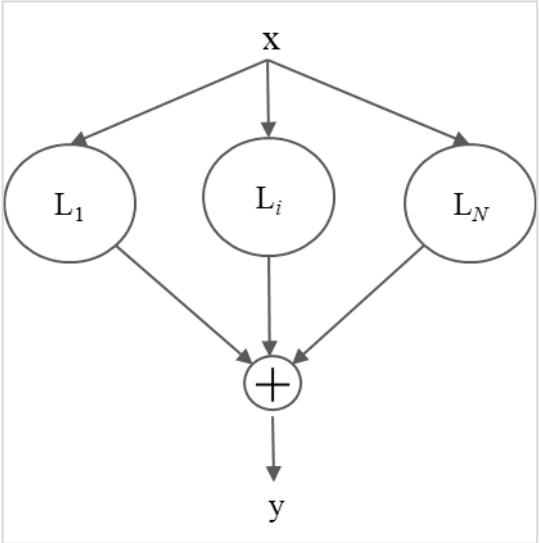

Figure 3.2: Inference process in Bagging.

The performance of the bagging method vitally depends on the instability of the base learners used. If the learning algorithm is stable, and produce similar results for similar datasets, the base learners cannot cancel out the errors made by other learners. As a result, one cannot utilize from bootstrap samples to improve the prediction accuracy. However, if the base learnes are unstable that is, small changes in the dataset results in large variance in the predictions of the base learners, then the bootstrap samples are adventagous, and fulfill the diversity requirement of ensembling.

Breiman points out that neural nets and decision trees are unstable predictors, while *k*-nearest neighbour models are stable [Breiman, 1996c]. Until the advances in deep neural networks, decision tree was a popular and effective model for bagging [Breiman, 1996a]. The most popular bagging model is Random Forests [Breiman, 2001], which is based on bagging with decision trees. The difference is that random forests aim at reducing the correlation between the decision trees (base learners) by

introducing more randomness during training. Specifically, during each split, a random set of features are selected to be a candidate for split variable, which makes the trees different from each other.

On the other hand, since deep neural networks are data hungry models, using 63% of the original dataset degrades the performance of the base learners, thus increase the model bias. Although bagging helps in reducing the variance and provides more gain than other strategies, the overall accuracy is limited [Renda et al., 2019], [Lakshminarayanan et al., 2017]. [Lee et al., 2015] show that for deep networks, random initialization of the weights provides sufficient diversity, making the bootstrapping unnecessary.

3.3. Boosting

Boosting is a sequential process for eventually obtaining a more accurate classifier compared to a single classifier trained conventionally. [Schapire , 1990] is the first who proposed a boosting algorithm to convert a weak learner to an accurate one. Basically, a weak learner model is trained many times with different distributions from domain X, to obtain several different hypotheses. These hypotheses are combined by to obtain a *single* more accurate hypothesis. Later, [Freund, 1995] proposed a more accurate boosting algorithm based on the Schapire's model.

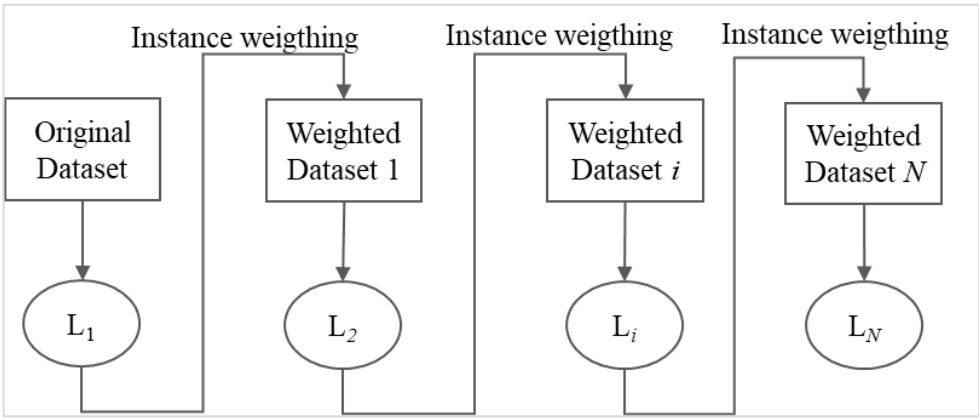

Figure 3.3: Training process in AdaBoost algorithm.

The most widely used boosting model, which is developed by Freund and Schapire [Freund and Schapire, 2005] is adaptive boosting, a.k.a. AdaBoost. Unlike

the previous algorithms, AdaBoost adjusts adaptively to the errors of the weak hypothesis. Specifically, in each iteration, the algorithm assigns new weights to the training examples based on the last hypothesis obtained (Figure 3.3). The weighting is done so that the observations misclassified by the preceding learner become more important. Figure 3.4 illustrates the inference process of AdaBoost. The resulting ensemble model is the weighted combination of all hypotheses where the weight of each hypothesis is directly proportional to its accuracy $\boldsymbol{\alpha}$.

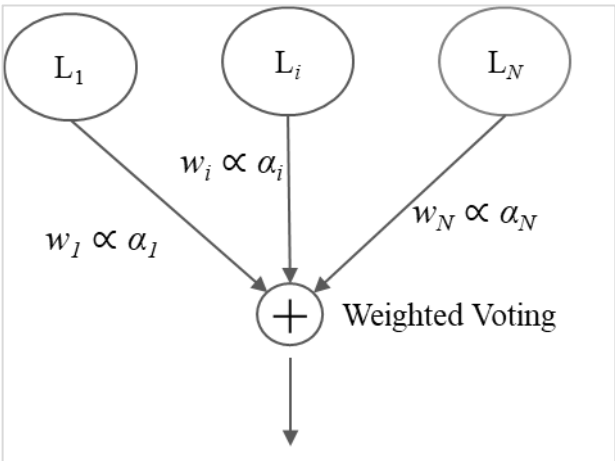

Figure 3.4: Inference process in AdaBoost algorithm.

Gradient Boosting [Friedman, 2001] is a more advanced boosting model, which aims to minimize a loss function while building the ensemble. Decision trees are used as the weak learners in gradient boosting. The procedure involves calculating the loss function based on the current ensemble model and adding new tress until convergence. Minimizing the loss function with the new tree is accomplished by parameterizing it and updating the parameters during the training. There are other algorithms which bring some improvement over the original one, such as Stochastic Gradient Boosting [Friedman, 2002] and Regularized Gradient Boosting. In the former, each tree is trained with a random sub-sample of the dataset to decrease the base learner correlation, while the latter regularizes the trees to avoid overfitting. Another algorithm called Extreme Gradient Boosting (XGBoost) [Chen and Guestrin, 2016] makes substantial improvements in the efficiency, and provides a scalable model which makes it a popular alternative [Nielsen, 2016] .

3.4. Mixture of Experts

Mixture of Experts (MoE) is an approach where each base learner is a neural network that specializes on a smaller region in the input space. To obtain this kind of specialization, a gating function makes soft partitions on the input space and defines these regions where the individual expert opinions are more confidential [Yuksel et al., 2012]. The gating function and the expert networks are trained jointly, as shown in Figure 3.5.

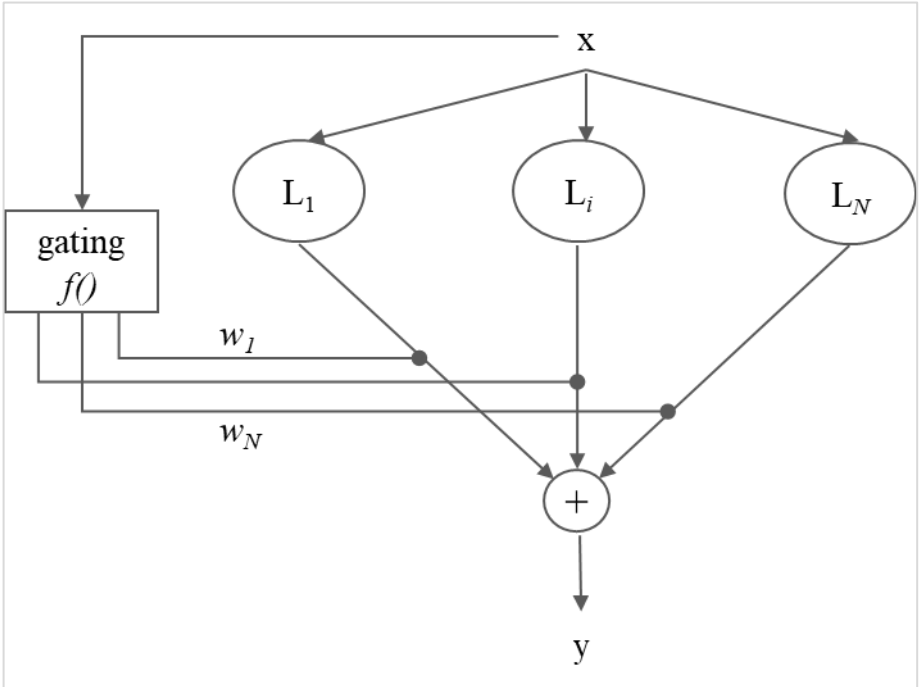

Figure 3.5: Mixture of Experts model.

Initial work defined a loss function that strongly couples the experts [Jacobs and Jordan, 1993], which results in cooperation of the base experts and causes using many experts for each training instance. Later Jacobs et al. [Jacobs et al., 1991] proposes Adaptive Mixture of Experts, which uses a loss function explicitly promotes localization and decreases the interdependence of the experts. As a result, when the gating network and the local experts are trained jointly, the model tends to devote a single expert for each instance.

Recently MoE model is adopted to deep neural networks in several ways. Shazeer [2019] proposed sparsely gated MoE layer for training large natural language processing data in recurrent networks. Their model increases the network capacity

drastically while introducing minor extra computational consumption. Fu et al. [2018] designs a convolutional mixture of expert layer to assess the importance of features from different levels for scene parsing. In [Miech et al., 2018] Mixture of Embedding Expert model is proposed, which enables learning from heterogeneous data. The proposed model is used for both text-to-video and video-to-text retrieval tasks.

3.5. Stacked Generalization

Stacked Generalization was originally proposed by [Wolpert, 1992] for either deducing the biases of single learners, or model combination in multiple learners. The approach is mainly based on by stacking two level of learners which are trained sequentially, which is is shown in Figure 3.6. In its form of model combination, a cross-validation process is followed to train the base models and the combiner model L_C. Specifically, for each cross-validation step, the base learners are trained on the training data and their outputs for the validation examples are collected for training later the combine model L_C. [Ting and Witten, 1999] have improved this method by using class probabilities of the base models instead of their single class prediction.

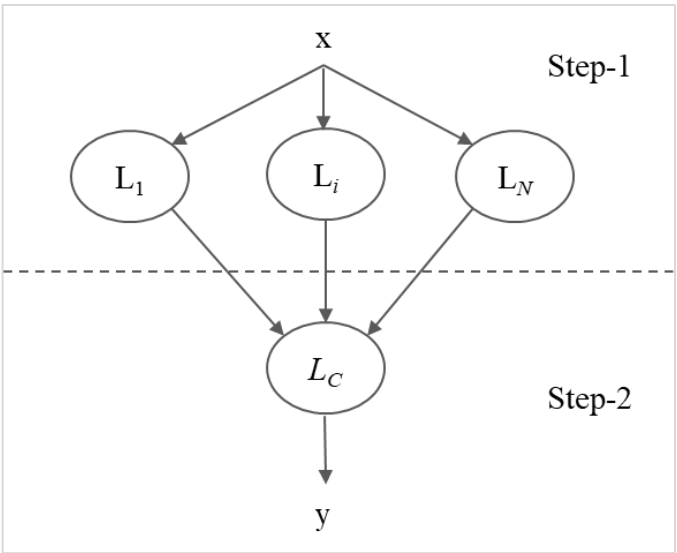

Figure 3.6: Stacked Generalization.

In deep learning, stacking is first used for greedy layer-wise training deep belief networks and stacked autoencoders [Hinton et al., 2006], [Bengio et al., 2007]. More recently, many methods have been proposed based on stacking idea for addressing

task-specific problems. [Yu et al., 2015] designed a deep architecture for representation learning, where extreme learning machine is used as the building block of stacking. [Hu et al., 2016] used the stacked generalization philosophy to build a deep network from randomly fixed single-hidden-layer networks for unsupervised learning. In [Palangi et al., 2017], Convolutional Deep Stacking Networks is proposed to reconstruct sparse vectors in distributive compressive sensing. Stacked generalization is used for combining various forest change detection algorithms in [Healey et al., 2018].

3.6. Cascading

Cascading is a sequential approach for creating classifier ensembles, where the base learners have increasing complexity. The earlier, efficient classifiers are used for handling most of the instances where they are certain, and the rest is handled by the succeeding ones depending on their difficulty. This process substantially speeds up the inference process.

Cascading has been introduced by [Alpaydin and Kaynak, 1998] to overcome the computational overhead of *k*-NN classifier in handwritten digit recognition. They cascade a parametric linear model and a *k*-NN classifier, where the large percentage of the digits are handled by the linear model and a the more expensive *k*-NN is used only for the remaining "exceptions". Later, Viola and Jones [Viola and Jones, 2001] proposed cascading of many boosted classifiers for object detection in their seminal work. The classifiers are trained in a way that when they detect a positive instance, the evaluation of the successive and more accurate classifier is triggered for spending more computation on the promising area. For increasing the complexity of the base learners, they adjust the threshold in AdaBoost algorithm. Wu et al. [Wu et al., 2004] improved the original method by using forward feature selection to construct the ensemble classifiers, which is more efficient.

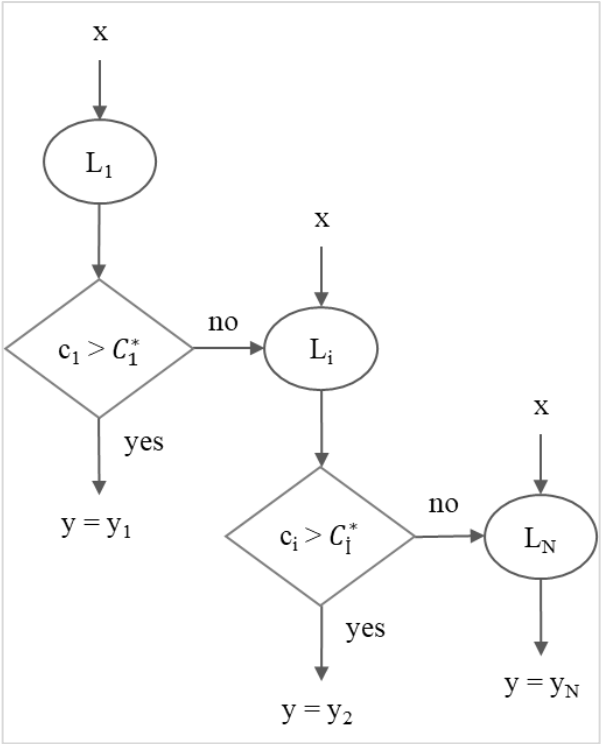

Figure 3.7: Cascading classifiers.

Cascading also helps reducing the computational time in deep networks, especially on real-time tasks. Li et al. [Li et al., 2015] proposes cascading deep networks which processes the input image in different scales for fast face detection. Sabokrou et al. [Sabokrou et al., 2017] cascades two deep networks for early identification of simple normal patches in anomlay detection. In [Cai and Vasconcelos, 2018], cascading is used in Region-based CNN for high-quality object detection,. The detectors are trained with increasing *intersection over union* thershold to be sequantially more selective against close false positives.

3.7. Deep ensemble learning

Neural networks enable a great way for training ensemble models thanks to randomness in both their initializations and training algorithms (i.e. stochastic gradient descent). Differences in hyper-parameters, random initialization, and random selection of minibatches during the training often provide sufficient diversity between the base learners.

Dropout technique [Srivastava et al., 2014] can be considered as one of the first attempts for deep ensemble learning. Although, there is no explicit networks trained

separately, by dropping some portion of the outputs, the technique allows to train a different architecture in each gradient step. Therefore, it is interpreted as an implicit model averaging method. This idea is generalized for other regularization techniques in [Singh et al., 2016]. In [Izmailov et al., 2018] multiple points along the trajectory of Stochastic Gradient Descent (SGD) is averaged to obtain a broader optima. These models exploit the ways of achieving more regularized models instead of combining diverse models during test time. Cyclic learning rate is another way of obtaining ensemble of deep networks. [Huang et al., 2017] and [Garipov et al., 2018] utilize this schedule for taking snapshots of the weights in different local minima during training. These models are very efficient in training time because they train only a single model. During test time, however, they require k times more computation to combine the predictions of k models [Izmailov et al., 2018].

Ensemble models require an increased computational cost compared to simple base learners. The computational cost of an ensemble model composed of deep neural networks hinders fats improvement on deep ensemble modeling. For this purpose, weight sharing is adapted in recently proposed methods. In [Opitz et al., 2017] multiple classifiers are trained on the non-overlapping splits of the last embedding layer to obtain a more representative feature vector. To promote the diversity between the base learners, online gradient boosting strategy is introduced. Hard example mining on a cascaded network is proposed at different levels in [Yuan et al., 2017]. [Kim et al., 2018] suggests utilizing from multiple attention masks to obtain an ensemble model. [Guo et al., 2018] proposes special grouping of training data for composing ensembles. Despite their efficiency, these models require additional training strategies in order to obtain diverse base learners.

3.8. Ensemble (or Multi-loss) Learning for Person ReID

In order to improve the generalization capacity of deep networks, many methods have suggested jointly training multiple loss functions on a single network for person ReID [Chen et al., 2017a], [Li et al., 2017], [Li et al., 2017] and also for other tasks [Shi et al., 2018], [Zheng et al., 2018], [Xuan et al., 2018], [Wang et al., 2019]. These approaches are analogous to the ensemble methods because they extract multiple

feature representations from different embedding spaces and combine them in test time.

The local feature learning is a natural way for defining multiple loss functions for a deep person ReID model. In [Sun et al., 2018], the input image is split into horizontal stripes and different feature extraction and classification layers are appended on each branch. In [Zhao et al., 2017], separate loss functions are defined for human body part regions and their representations are concatenated during the test time. [Li et al., 2018] designs a harmonious attention module to use on both local and global branches to overcome the misalignments problem in person bounding boxes.

More recently, [Zhai et al., 2019] adapted a multi branch network from non-overlapping channel splits of feature maps to extract diverse global features. There are also methods that propose to extract multiple features from multiple scales without defining explicit loss function for different features [Chang et al., 2018]. Instead, they optimize a single global loss on the fused feature vector.

The alignment problem in person ReID is to find the matches between the body parts of two human images. The aim of [Li et al., 2018] is addressing the misalignment problem by jointly learning attention selection and feature representation. The approach of Li et al. [Li et al., 2018], uses a multi branch network for global and local feature representation. In contrast to our work, this approach requires a special network architecture design to incorporate soft pixel and hard regional attention mechanisms. Another work that focuses on the misalignment problem is [Zhao et al., 2017], which automatically detects human part regions. The feature vectors extracted from each part are concatenated, and a global triplet loss is defined over the concatenated feature vectors. [Wang et al., 2018] extract multiple features from different levels of the backbone network to enable scalable person ReID.

The objective of [Chang et al., 2018] is to factorize the visual appearance of the person image into latent discriminative factors, which works similarly to the attention mechanism. They employ a factor selection module in each layer that activates or deactivates the extracted features, which are fused at the end and are globally optimized. The methods optimizing a single loss on the fused feature vectors has the risk of promoting only the high-level discriminative features. Our multi-loss approach avoids this problem by optimizing each base learner explicitly.

The aim of [Wang et al., 2018] is to represent both local and global features in the image representation by partitioning the input image into multiple regions in a coarse-to-fine manner. The method requires ReID specific network design which considers human body structures such as head and shoulders, main body and the legs. It benefits from joint training of metric learning and discriminative learning since triplet loss is employed for each granularity and cross entropy loss is used for each part.

4. DATASETS, EVALUATION AND THE BASELINE MODEL

In this section, we take a look at some of the existing datasets. Then, we revisit the metrics used for evaluating the person ReID models. Lastly, we define a baseline model, which the most recent study is based on.

4.1. Datasets

Collecting data for Person ReID is a laboring task. The procedure involves raw data collection, bounding box detection, and labeling [Leng et al., 2019]. Raw data collection is capturing the video data from a camera network, where cameras are placed in different places and have varying imaging conditions. Bounding box annotation is the task of identifying the image area where the person appears. Lastly, the detected bounding boxes are annotated with the correct identity, which should be consistent with the other images of the same person captured from different cameras.

There are many datasets collected for image and video-based person ReID. Table 4.1 lists the details of some ReID datasets according to their release time [Web-2, 2020]. The number of identities and cameras represents the number of distinct persons and cameras, respectively. The number of images represents the total number of images for training and evaluation. The datasets having multi-shot property provides multiple images for the probe image, which can enhance the ReID performance. Those having the tracking traclets contain videos of both the probe and gallery sets. The initial datasets, which are limited in their number of cameras and identities, have bounding boxes annotated manually, while for some recent and more massive datasets bounding boxes were obtained via person detection and tracking algorithms [Leng et al., 2019].

Person ReID datasets have some difficulties, some of which are especially designed for simulating the real-world problem. For example, QMUL iLIDS [Zheng et al., 2009] is collected in an airport, so the images suffer from severe occlusion. MARS [Zheng et al., 2016] is a large scale video-based dataset and due to automatic bounding box detection, there are some distractors, which is more close to the real scenario. Similarly, MSMT17 [Wei et al., 2018] dataset aims to simulate the real scenario. Therefore, the videos were taken under various conditions, which presents complex scenes and backgrounds. There are some other datasets that are not listed in

the table, such as Partial iLIDS, which is specifically designed for the partial ReID problem from iLIDS dataset.

The pioneering work generally used small-scale ReID datasets such as VIPeR [Gray et al., 2007], QMUL iLIDS [Zheng et al., 2009], GRID [Loy et al., 2010], 3DPeS [Baltieri et al., 2011] and PRID2011 [Hirzer et al., 2011] with handcrafted features. In the second stage of person ReID literature, relatively shallow networks were carefully designed for person ReID. In this stage, the metric learning approach was advantageous, since it copes with the data scarcity by preparing the training data in pairs or triplets [Li et al., 2014], [Yi et al., 2014].

After very deep networks became trainable, large scale person ReID datasets were released and impressive results have been obtained on these datasets using only the classification loss, with the help of knowledge transfer from large scale image classification datasets [Zheng et al., 2016], [Zhai et al., 2019]. In Table 4.2, the number of person identities and images for query and gallery images are given for these larger datasets, which are also used for evaluating the proposed person ReID system in this thesis.

4.2. Evaluation Protocol

4.2.1. Cumulative Matching Characteristics

The performance of person ReID methods are generally measured by Cumulative Matching Characteristics (CMC) curves. The CMC curve represents the number of correctly detected queries within the first n ranks.

4.2.2. Mean Average Precision

Mean Average Precision (mAP) is a metric used for evaluating information retrieval systems. Considering the large-scale person ReID as an image retrieval task, mean Average Precision (mAP) is used as another evaluation metric. In order to define Mean Average Precision, we first review *precision at K* and *average precision* scores.

Precision at K: Corresponds to the number of relevant results among the top K documents. For person ReID, this should be considered as the number of positive identities among the top K persons retrieved.

Average Precision (AP): Represents the average of precision at different K values for a single query person.

Mean Average Precision (mAP): The mean of the average precision values of a set of query images.

Table 4.1: Some person ReID datasets, ordered in release time.

Dataset	Release time	# identities	# cameras	# images	Label method	Crop size	Multi-shot	Tracking sequences
VIPeR	2007	632	2	1K	Hand	128X48		
QMUL iLIDS	2009	119	2	0.5K	Hand	Vary	✔	
GRID	2010	1025	8	1K	Hand	Vary		
3DPeS	2011	192	8	1K	Hand	Vary	✔	
PRID2011	2011	934	2	24K	Hand	128X64	✔	✔
CUHK01	2012	971	2	3K	Hand	160X60	✔	
CUHK02	2013	1816	10(5 pairs)	7K	Hand	160X60	✔	
CUHK03	2014	1467	10(5 pairs)	13K	Hand/DPM	Vary	✔	
iLIDS-VID	2014	300	2	42K	Hand	Vary	✔	✔
Shinpuhkan	2014	24	16	22K	Hand	128X48	✔	✔
Market1501	2015	1501	6	32K	Hand/DPM	128X64	✔	
MARS	2016	1261	6	1100K	DPM + GMMCP	256X128	✔	✔
DukeMTMC-reID	2017	1812	8	36K	Hand	Vary	✔	
DukeMTMC4ReID	2017	1852	8	46K	Doppia	Vary	✔	
MSMT17	2018	4101	15	126K	Faster RCNN	Vary	✔	
LPW	2018	2731	3, 4, 4	592K	Detector + NN + Hand	-	✔	✔

4.2.3. Architecture

In [Zheng et al., 2016], it has been shown that very deep networks pre-trained on Imagenet [Deng et al., 2009] produces state-of-the-art results when fine-tuned on large scale person ReID datasets with classification loss, which is called identity discriminative embedding (IDE) model. Considering that fine-tuning a deep network is much more efficient than training one from scratch, most researchers have adopted this approach in recent ReID models. Preliminary studies have suggested using GoogLeNet [Szegedy et al., 2015] as the backbone [Geng et al., 2016], while more recent studies generally employ versions of ResNet [He et al., 2016] or DenseNet [Huang et al., 2017]. We choose IDE model as the baseline in this thesis due to its remarkable performance.

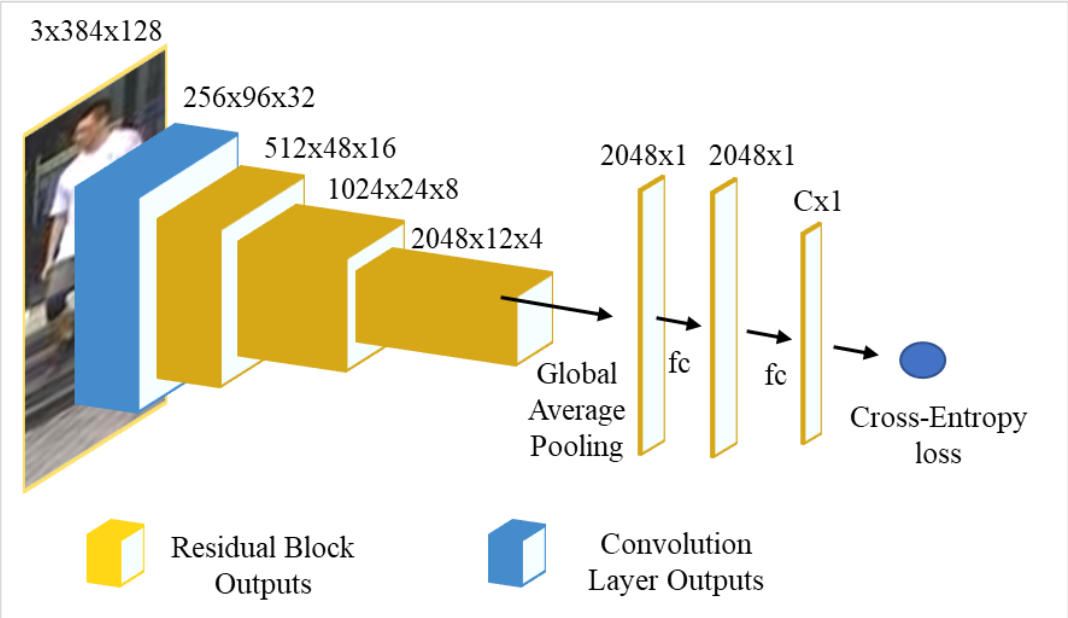

Figure 4.1: Baseline IDE model.

In Figure 4.1, the baseline IDE model, which uses ResNet50 as the backbone is illustrated. Generally, the images are resized to 394128 and global average pooling is applied to the (2048×H×W) feature maps obtained at the end. Then, an optional generally 2048 length fully connected layer is appended. Lastly, another fully connected layer with length C is used, where C is the number of person identities. The network is trained with softmax loss.

4.3. Baseline Model

In this section, we define a baseline model for person ReID in detail. First, we give the architecture design, which is based on the recent general-purpose deep networks. Then we review basic pre-processing and data augmentation methods used in most studies. Lastly, we report the scores of this baseline model on the widely used large-scale datasets.

Table 4.2: Details on the datasets used in the experiments throughout this thesis.

	Number of ID's		Number of Images		
Dataset	Training	Test	Training	Query	Gallery
Market-1501	750	751	12,396	3,368	19,372
DukeMTMTC-reid	702	702	17,661	2,228	16,522
CUHK03-Labeled	767	700	7,368	1,400	5,328
CUHK03-Detected	767	700	7,365	1,400	5,332
MSMT17	1041	3060	32,621	11659	82,161

4.3.1. Data Pre-Processing and Augmentation

In order to utilize from a pre-trained network, the input images of the new dataset are normalized in consistent with the dataset used in pre-training. Therefore, the input images are normalized between [0,1] and mean subtracted with the ImageNet mean.

The basic data augmentation for person ReID involves randomly cropping the input image and resizing to a certain image width and height. Also, the images are randomly flipped horizontally with 0.5 probability during training, which is found to be very effective. Another augmentation method is Random Erasing, which replaces a randomly selected rectangular area in the image with noise [Zhong et al., 2017]. The random erasing data augmentation is also applied with 0.5 probability to an area whose size is the half of the input image area.

During testing, the query and gallery images are normalized and mean subtracted as for the training stage. After obtaining the embedding vectors, the images are

horizontally flipped and passed through the network again to obtain the embedding feature vector for the flipped images. The final ranking is performed based on the concatenation of these two embedding vectors, which compensates some of the errors due to vertical misalignments between the query and gallery images.

4.3.2. Implementation Details

In general, Adam [Kingma and Ba, 2015] is used as the optimizer. In order to finetune the network, the classification layer of the original network is thrown and fresh embedding and classification layers are appended. The learning rate is set to 0.1 and 0.01 for the new and existing layers, respectively. Mini batch size is set to 32 and the network is trained 50 epochs in total, where the learning rates are decreased by a factor of 0.1 after 40 epochs. The training of IDE model on an NVIDIA GTX 1080 Ti requires 70 minutes on Market-1501 dataset.

4.3.3. Results

We evaluate the baseline model on the most widely used person ReID datasets given in Table 4.2. The baseline model is compared to some previous work which uses Inception architecture [Szegedy et al., 2015] as the backbone in Table 4. RE stands for random erasing data augmentation [Zhong et al., 2017]. It improves the mAP and Rank-1 scores on all datasets. As shown in the table, ResNet50 gives superior performance on Inception architecture. Note that, the previous works do not use the standard IDE model and improve the baseline Inception model in some way, such as using multi-scale features or pose-sensitive embedding. The results indicate that the IDE model using ResNet50 or DenseNet121 as the backbone is a valid baseline. Therefore, this model is used as the baseline model in this thesis.

Table 4.3: Comparison of the baseline IDE model with the state-of-the-art.

Method	Market-1501		DukeMTMC-reid		CUHK03-Labeled		CUHK03-Detected	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Deep Transfer (CVPR) [Geng et al., 2016]	83.7	65.5	-	-	-	-	-	-
PSE (CVPR) [Sarfraz et al. 2018]	84.4	64.9	71.7	50.4	-	-	-	-
DPFL (ICCV) [Chen et al., 2017]	88.9	73.1	79.2	60.6	43.0	40.5	40.7	37.0
ResNet50	88.0	70.9	77.9	58.9	48.1	43.6	45.2	41.1
DenseNet121	88.2	71.7	79.5	61.8	49.7	46.0	50.1	46.5
ResNet50+RE	90.7	76.7	82.5	66.6	56.4	52.2	55.9	49.9
DenseNet121+RE	**91.1**	**77.8**	**83.5**	**68.7**	**58.1**	**53.7**	**56.2**	**51.3**

5. END-TO-END CNN ENSEMBLES

5.1. Motivation

Neural networks provide a natural diversity between the models, mainly due to their intrinsic randomness at initialization. Therefore, the same networks trained in parallel optimize to different locations and become sufficiently diverse without any explicit effort [Alpaydin, 1993]. Nevertheless, many methods have been offered to improve the diversity of neural networks [Liu and Yao, 1999].

The present study suggests an ensemble model for reducing the model variance in person ReID using deep networks. Overfitting is even more severe in this challenging task due to the discrepancy between the training and test distributions caused by different person identities and varying imaging conditions. Therefore, it is inevitable to lose performance when a single, overfitted model is used for feature extraction in test time. To tackle this problem, we propose an ensemble learning model to extract diverse feature vectors to be combined in test time ranking. Each base learner extracts different and complementary information, improving the ranking accuracy. The proposed ensemble model is trained in an end-to-end manner, where the base learners share a considerable amount of costly convolution operations. As a result, an ensemble model that is efficient in both training and test time is obtained.

The proposed model is evaluated on four benchmark datasets via several experiments. Analysis based on the experiments confirmed that our model is favorable in terms of ranking performance and computational resource consumption while outperforming state-of-the-art results.

5.2. Problem Definition

Given a 2D query person image q, the objective of the person ReID model is to rank a large set of gallery images $\mathrm{G} = \{\mathrm{g_i}\}$ with respect to their similarities to the query image, where $|\mathrm{G}| = \mathrm{M} \in \mathbb{Z}^+$and $\mathrm{g_i}$ is the $\mathrm{i^{th}}$ gallery image. Ideally, the ReID system is not aware of the query and gallery images before, and it is expected that the top-ranked images from the gallery set belong to the same person with the query image q.

In order to calculate the similarity scores between the query and gallery images, a feature extraction model is required. In deep networks, the high-level semantic

feature extraction layer is used for this purpose. To this end, a training data $X = \{x_i\}$ consisting of N images of C distinct persons and their identity labels $Y = \{y_i\}$ are used to train a CNN, where $|X| = |Y| = N \in \mathbb{Z}^+$, $C \in \mathbb{Z}^+$, $y_i \in \mathbb{Z}^+, y_i \leq C$, and $C \leq N$ (usually $C << N$).

A classification network consists of two sequential networks. The first one is a feature extraction network that projects the given input image x of D pixels to an H dimensional embedding space, and we represent it by the function

$$h(x; \theta_h): \mathbb{R}^D \rightarrow \mathbb{R}^H \tag{5.1}$$

where θ_h are the parameters of this function. The second one is usually a one-layer network, which is a function of $h(\cdot)$ and produces C dimensional class probabilities from the *H* dimensional feature vectors. We represent this network by the function

$$f(h(\cdot); \theta_f): \mathbb{R}^H \rightarrow \mathbb{R}^C \tag{5.2}$$

where θ_f represents the parameters of this network. The empirical risk of $f(\cdot)$ and $h(\cdot)$ is defined as

$$\mathcal{R}(f, h) = \sum_{i}^{N} \mathcal{L}\,(x_i, y_i) \tag{5.3}$$

where $\mathcal{L}(x_i, y_i)$ is the classification loss for the ith training example. A CNN classifier is trained by minimizing the cross-entropy, i.e., the divergence of the estimated and the actual class probability distributions. For simplicity, let f_i^k and y_i^k represent the estimated and actual probability of the ith training image to be an instance of the kth person class, respectively. The cross-entropy loss for this example is

$$\mathcal{L}(x_i, y_i) = -\sum_{k}^{C} y_i^k \log\,(f_i^k) \tag{5.4}$$

The goal of the optimization is to find the network parameters that best fits to the objective

$$\theta_f^*, \theta_h^* = \arg \min_{\theta_f, \theta_h} \mathcal{R}(f, h) \tag{5.5}$$

The training data is composed of limited number of observations and is far from representing the real distribution. Aa a result, empirical risk minimization does not guarantee to minimize the generalization error, which is the expected risk over the real distribution [Guyon et al., 1992]. Moreover, if the model has a high complexity, it may overfit to the observed examples and perform poorly on test data. The discrepancy between the training/test distributions is more severe in person ReID task due to the differences in person identity classes. Therefore, reducing the model variance is of crucial importance in person ReID models. In the following section, we define our ensemble model which deals with this problem.

5.3. The Proposed Method

The traditional ensemble models train multiple base learners independently, which allows increasing diversity and accuracy of the individual learners. When it comes to deep networks, this approach suffers from inefficiency because deep networks require a massive amount of computational consumption during training. To overcome this difficulty, we propose a framework that trains multiple base learners in a single deep network in an end-to-end manner while most of the expensive convolution operations are shared.

Figure 5.1 presents the proposed ensemble model, which is based on DenseNet. The low and mid-level blocks (Block 1, 2, and 3) of DenseNet are not modified. To embed multiple base learners in this backbone architecture, we integrate sub-networks on top of various layers in the last dense block (Dense Block 4). Each sub-network, together with the shared backbone architecture, composes a base learner. The structure of the sub-networks are illustrated by Figure 5.2. In order not to lose the spatial information, the input feature maps are fed into an embedding layer without using any form of pooling. We call this property of the proposed model as spatial-awareness.

Finally, a classification layer outputs the logits for the cross-entropy loss for this base learner.

There are 2L base learners in the proposed model: the base learners 1 to L are obtained by integrating sub-networks after the channel-wise splits of the third block's output, while the base learners L+1 to 2L are obtained by integrating sub-networks

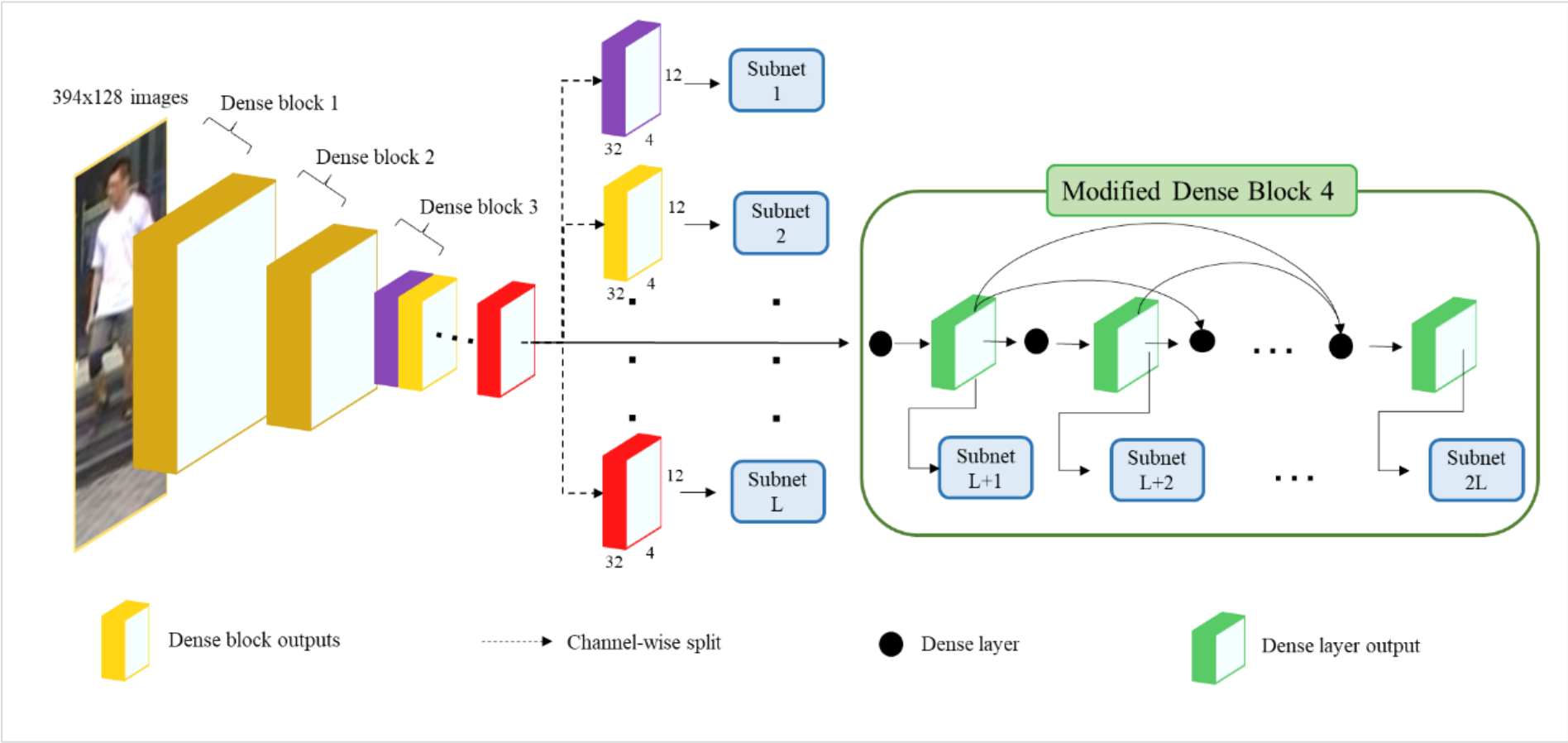

Figure 5.1: The architecture of the proposed system.

after each dense layer in the fourth dense block. Each base learner includes two sequential networks: the feature extraction network consists of the shared backbone and the embedding layer of its sub-network, and the classification network consists of only the classification layer.

Similar to the classical method given in Section 5.2, the ℓth feature extraction network project the input images of the feature embedding space, where some of its parameters are now shared. Let the shared and non-shared parameters is represented by $\theta_h^{\ell,S}$ and $\theta_h^{\ell,N}$, respectively. Then, the feature extraction network is $h_\ell(x;\theta_h^{\ell,S},\theta_h^{\ell,N})$. On the other hand, the ℓth classification network has only non-shared parameters and can be represented by $f_\ell(h_\ell(x);\theta_f^{\ell})$. The empirical risk defined in Equation 1 is then extended by the inclusion of the set of all feature extraction and classification networks in the ensemble model

$$\mathcal{R}_e(\mathcal{F},\mathcal{H}) = \sum_{i}^{N} \mathcal{L}_e\,(x_i, y_i), \tag{5.6}$$

where $\mathcal{F} = \{f_1,\ldots,f_{2L}\}$, $\mathcal{H} = \{h_1,\ldots,h_{2L}\}$, and $\mathcal{L}_e$ is the ensemble loss for a single training sample. Let $f_{\ell,i}^k$ be the output of the ℓth learner for the ith sample x_i to be an instance of the kth person class. Then, the ensemble cross entropy for this instance is

$$\mathcal{L}_e(x_i, y_i) = -\sum_{\ell}^{2L}\sum_{k}^{C} y_i^k \log\left(f_{\ell,i}^k\right) \tag{5.7}$$

The objective of the training is to estimate the parameters of all feature extraction and classification networks that minimize the empirical risk

$$\theta_F^*, \theta_H^* = \arg\min_{\theta_F,\theta_H} \mathcal{R}_e(\mathcal{F},\mathcal{H})\ , \tag{5.8}$$

where $\theta_F = \{\theta_f^1, \theta_f^2, \ldots, \theta_f^{2L}\}$ and $\theta_H = \{\theta_h^1, \theta_h^2, \ldots, \theta_h^{2L}\}$ are the parameters of all feature extraction and classification networks, respectively. Joint optimization of these networks allows for end-to-end training of the base learners.

The efficiency of the ensemble models depends on the accuracy and diversity of the base learners. To fulfill these requirements, we benefit from DenseNet architecture. There are two primary sources of diversity in our model: random initialization of the non-shared parameters and the diverse input feature maps of the base learners. The random initialization of the parameters is an intrinsic property of the neural networks and different input feature maps add further variety. Figure 5.3 shows the normalized input feature maps for a baseline classification model and the base learners of our model for some sample cases. The baseline model (second column) receives all of the feature maps, which are concentrated on the most discriminative part of the image. Base learners, on the other hand, receive information that focuses on diverse and complementary parts, as expected.

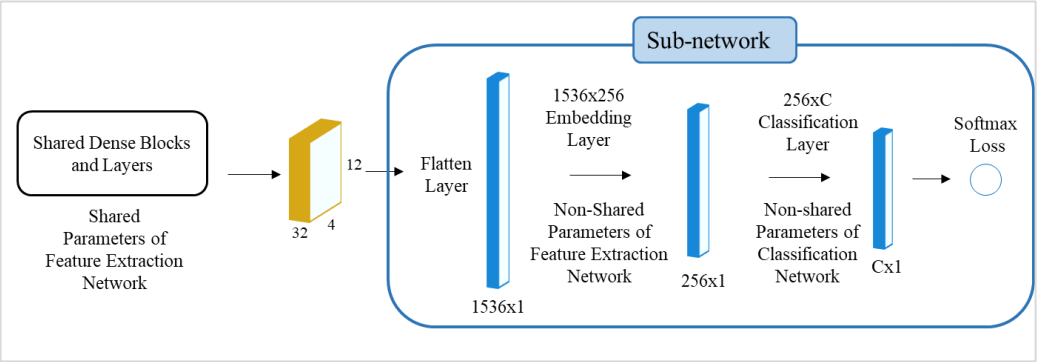

Figure 5.2: Structure of the sub-networks.

To accomplish the accuracy objective, we utilize spatially aware base learners, via adopting full connections between the feature maps and the embedding layer. In other words, unlike most of the work, we do not employ any pooling mechanism on the produced feature maps to keep the spatial information of the features in the final representation. Generally, fully connected layers increase the number of parameters dramatically if the input and output dimensions are in high order. DenseNet architecture is suitable to overcome this difficulty thanks to its tiny layers (32) filters for Densenet121). Also, we use a small number of nodes in the embedding layer (512). As a result, we acquire spatially-aware base learners with minimal cost, which makes our end-to-end model computationally efficient. In Section 5.6.3, our analysis shows that the spatially-aware base learners of our model outperform most existing methods, individually. Unlike conventional ensemble models, our models is very

computationally very efficient thanks to the shared dense blocks which include most of the costly convolutional layers. More specifically, we observed that 98% of the FLoating-point OPerations (FLOPs) occur in these shared blocks. Our analysis in Section 5.6.4. shows that the proposed model increases the number of FLOPs by a small amount while performing on par with a conventional ensemble model that requires at least two times more FLOPs.

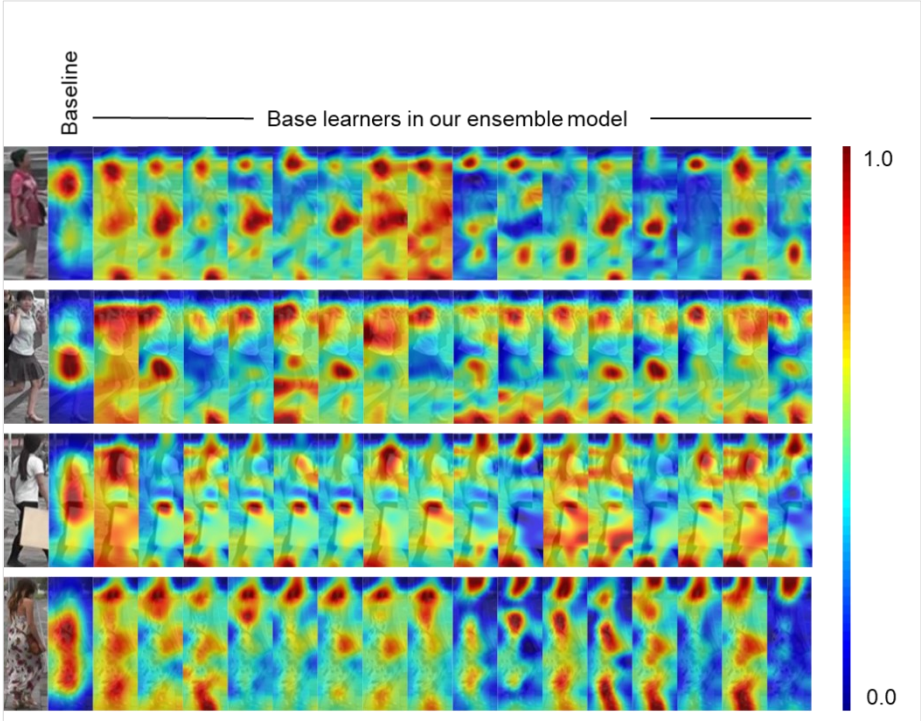

Figure 5.3: Normalized convolution feature maps as input to different learners.

5.4. Binary Hash Code Generation

Binary hash code generation is a technique used in Image Retrieval systems, where efficient distance calculation is crucial for rapid search in large databases. The feature vectors extracted from the query and the database images are quantized, and the distance calculation is done in Hamming space in the bit level.

Person ReID is a specific application of image retrieval, where efficiency is one of the critical issues. Therefore, real-world ReID systems should involve solutions for fast distance calculation also. However, many ReID models operate on real-valued feature representation and use Euclidean distance as the similarity measure. This issue

becomes even more problematic when feature vectors extracted from different stages are concatenated to form a one large feature vector.

The proposed end-to-end ensemble model produces feature vectors that are 8192 in length. Therefore, an efficient distance calculation is also crucial for our approach. To address this issue, following many image retrieval models, we propose using *tanh* function as the activation function in the last layer of the feature extraction network. In this way, the feature vectors can be quantized into binary vectors in the inference time, and the distance calculation can be performed in Hamming space, i.e., in the bit level.

5.5. Ranking with the Proposed Model

At the end of the training, we obtain several base learners which discriminate person classes in separate embedding spaces. To extract representative feature vectors from the images of the unseen query and gallery persons, the classification networks of the base learners are ignored and only the feature extraction parts are used.

Given a query image q, let the ℓth feature extractor function, $h_\ell(q, \theta_H^*)$, produce feature vectors from its embedding layer. We combine the feature vectors of all the feature extractors.

$$\mathcal{H}(q) = h_1(q)|h_2(q)|\cdots|h_{2L}(q) \tag{5.9}$$

where | represents the concatenation operation. We also calculate the feature vectors of each gallery image g_i as $\mathcal{H}(g_i)$ and sort the gallery images concerning their distances to the query image by a distance function $D(q, g_i)$. We use two types of distance calculation metrics in our experiments: Euclidean distance and Hamming distance.

Generally, the distance between two vectors $\mathrm{x}, \mathrm{y} \in \mathbb{R}^N$ is calculated in Euclidean space as

$$D_E(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{5.10}$$

which is inefficient for the *8192* length feature vectors of the proposed method. Therefore, we employ *tanh* as the activation function in the embedding layers to allow binary code generation for Hamming distance calculation. A real-valued feature vector $\mathrm{x} \in \mathbb{R}^N$ is quantized into a binary vector as $\mathrm{b} = sign(x) \in \mathbb{R}^N$ where $sign(\cdot)$ is an element-wise operation

$$sign(x) = \begin{cases} +1, & \text{if } x \geq 0 \\ -1, & \text{if } x < 0 \end{cases} \tag{5.11}$$

Hamming distance between two binary vectors represents the number of bit positions where the two bits are different. Mathematically, Hamming distance between $x, y \in \{0,1\}^N$ is calculated as

$$D_H(\mathrm{x,y}) = \sum_{i=1}^{N} \mathbf{1}\,(x_i \neq y_i) \tag{5.12}$$

where $\mathbf{1}[\cdot]$ is the indicator function.

It is worth noting that the designed ensemble model is scalable in terms of the number of base learners to be used in test time. Unlike many part-base models [Sun et al., 2018], which produces a tightly-coupled relationship between the feature extractors, the base learners in our model work independently as standalone feature extractors, and any of them can be ignored in return for a small performance loss.

5.6. Experiments

5.6.1. Implementation Details

Our method is based on DenseNet121. We initialize the parameters using a pre-trained model on Imagenet dataset. The images are resized to 384×128. For data augmentation, we use random horizontal flipping, random cropping and random erasing [Zhong et al., 2017]. In run time, the final feature representation is the sum of features extracted from both the original image and its flipped. The learning rate and batchsize is determined via experimenting on a baseline model, and set to 0.05 and 32,

respectively. We use the same settings for all experiments. The network is trained for 50 epochs, where the learning rate is decayed by a factor of 0.1 after 40 epochs. Embedding layer size depends on the number of base learners so that he final feature representation is 8192 in length, i.e., 1024 for 8 (L=4), 512 for 16 (L=8), and 256 for 32 (L=16) learners. The activation of the embedding layer is performed by non-linear *tanh* function.

5.6.2. Comparison with the State-of-the-art

The end-to-end ensemble model is compared with the state-of- the-art methods in this section. We use Market-1501 [Zheng et al. 2015], DukeMTMC-reid [Zheng et al. 2017], CUHK03 [Li et al. 2014] and MSMT17 datasets [Wei et al., 2018] to evaluate our approach.

5.6.2.1.Market-1501 Dataset

Market-1501 dataset consists of 32,668 images of 1501 identities collected from 6 cameras. 750 identities are reserved for evaluation and the rest is used for training. We use the same protocol with [Zheng et al., 2015] to evaluate our method.

The comparison with the state-of- the-art methods is given in Table 5.1. The proposed model is the best in terms of mAP score. In Rank-1 accuracy, our method produces competitive performance with PCB+RPP [Sun et al., 2018b], which is a part-based model including a ReID specific technique called refined part pooling.

Another important observation from Table 5.1 is the performance of the proposed method in Hamming space. The feature vectors in our method are still very informative when quantized into binary values and has the advantage of fast distance computation in exchange for minimal performance loss in Rank-1 accuracy. Although the proposed method produces large feature vectors (8192 in length), is more efficient in Hamming space and still superior than most of the previous approaches.

5.6.2.2.DukeMTMC-reid Dataset

This dataset is collected as a multi-target multi-camera pedestrian tracking dataset and consists of 85-minute videos from 8 cameras. A subset of this dataset is

prepared for image-based person ReID in [Zheng et al., 2017]. An evaluation protocol similar to Market-1501 is provided. Specifically, 17,661 images of 702 persons are used for training and 2228 images of 702 persons are used as query images which are searched among 16,522 gallery images.

Table 5.1: Comparison with the state-of-the-art approaches in Market-1501[2].

Method	R-1	mAP
SVDNet+RE [Zhong et al., 2017]	87.08	71.31
Pose-Transfer*[Liu et al., 2018]	87.65	68.92
DPFL [Chen et al., 2017]	88.90	73.10
DaRe+RE* [Wang et al., 2018]	89.00	76.00
Mid-level [Yu et al., 2017]	89.87	75.55
MLFN [Chang et al., 2018]	90.00	74.30
HA-CNN [Li et al., 2018]	91.20	75.70
DuATM [Si et al., 2018]	91.42	76.62
GP** [Almazan et al., 2018]	92.20	81.20
PCB [Sun, et al., 2018b]	92.40	77.30
MultiBranch [Zhai et al., 2018]	93.10	78.90
PCB+RPP [Sun, et al., 2018b]	**93.80**	81.60
Ours_Ensemble (Euc.)	93.19	82.10
Ours_Ensemble (Ham.)	93.13	**82.19**

In Table 5.2, our approach is compared to the most recent ReID methods evaluated on DukeMTMC-reid dataset. The end-to-end ensemble model show competitive performance with the previous work, including the multi-loss models. It is noticeable that Hamming space performance is better in mAP score compared to the Euclidean space.

[2] * and ** indicates DenseNet121/169 or ResNet101 is used as the backbone, respectively in all tables.

Table 5.2: Comparison with the state-of-the-art approaches in DukeMTMC-reid.

Method	R-1	mAP
Pose-Transfer* [Liu et al., 2018]	78.52	56.91
DPFL [Chen et al., 2017]	79.20	60.60
SVDNet+RE [Zhong et al., 2017]	79.80	62.00
Mid-level [Yu et al., 2017]	80.43	63.88
DaRe+RE* [Wang et al., 2018]	80.20	64.50
HA-CNN [Li et al., 2018]	80.50	63.80
MLFN [Chang et al., 2018]	81.00	62.80
DuATM [Si et al., 2018]*	81.82	64.58
PCB B [Sun et al., 2018]	81.90	65.30
PCB+RPP B [Sun et al., 2018]	83.30	69.20
MultiBranch [Zhai et al., 2019]	84.00	68.40
GP** [Almazan et al., 2018]	85.20	72.80
Ours_Ensemble (Euc.)	**86.26**	72.63
Ours_Ensemble (Ham.)	85.83	**73.16**

5.6.2.3.CUHK03 Dataset

CUHK03 dataset is a collection of 14,097 images of 1437 identities and has two versions based on the bounding box annotation procedure: manually labeled and automatically detected. In [Zhong et al., 2017], a new protocol similar to Market-1501 and DukeMTMTC-reid datasets is suggested for this dataset. According to this setting, there are 767 and 700 identities in training and test sets, respectively and the same evaluation procedure is employed with Market-1501 and DukeMTMC-reid.

The proposed method is compared to recent deep learning approaches in Table 5.3. End-to-end ensemble model presents superior performance than the previous techniques and on both labeled and auto-detected bounding boxes. Rank-1 accuracy and mAP scores are increased by 5% in the labeled set, while in the auto-detected set we observe 4% and 2% increase, respectively.

CUHK03 dataset presents the overfitting problem more obviously due to its relatively modest number of training images. Therefore, the proposed method's

advantage on the overfitting problem is shown more explicitly against other multi-loss approaches such as [Zhai et al., 2019] and [Sun et al., 2018].

Table 5.3: Comparison with the state-of-the-art on CUHK03 datasets.

Method	CUHK03-Labeled		CUHK03-Detected	
	R1	mAP	R1	mAP
Pose-Transfer* [Liu et al., 2018]	45.10	42.00	41.60	38.70
DPFL [Chen et al., 2017]	43.00	40.50	40.70	37.00
DaRe+RE* [Wang et al., 2018]	66.10	61.60	63.30	59.00
MLFN [Chang et al., 2018]	54.70	49.20	52.80	47.80
TriNet+RE [Zhong et al., 2017]	58.14	53.83	55.50	50.74
HA-CNN [Li et al., 2018]	44.00	41.00	41.70	38.60
PCB [Sun et al., 2018]	-	-	61.30	54.20
MultiBranch [Zhai et al., 2019]	-	-	61.70	55.30
PCB+RPP [Sun et al., 2018]	-	-	63.70	57.50
Ours_Ensemble (Euc.)*	**71.13**	66.23	**67.20**	**61.73**
Ours_Ensemble (Ham.)*	71.06	**66.30**	67.10	61.66

5.6.2.4.MSMT17 Dataset

MSMT17 is a recently released dataset that includes severe variation in imaging conditions, thus more realistic. The proposed method is compared with the existing models in Table 5.4. End-to-end ensemble model outperforms the previous studies GoogleNet [Szegedy et al., 2015], PDC [Su et al., 2017] and GLAD [Wei et al., 2017], which are reported by the publishers of the dataset. Our model also performs comparably with DG-Net [Zheng et al., 2019]. These results indicate that our method can produce competitive results on such challenging datasets.

Table 5.4: Comparison with state-of-the-art methods on MSMT17 dataset.

Method	R-1	mAP
GoogleNet [Szegedy et al., 2015]	47.6	23.0
PDC [Su et al., 2017]	58.0	29.7
GLAD [Wei et al., 2017]	61.4	34.0
DG-Net [Zheng et al., 2019]	**77.2**	**52.3**
ResNet50 IDE	66.5	38.9
DenseNet121 IDE	70.8	44.2
Ours (Euc.)	76.5	49.5
Ours (Ham.)	75.9	50.1

5.6.3. Further Analysis

In this section, we analyze our proposed model in detail to demonstrate its effectiveness. Specifically, we report the performance of the base learners individually, compare them with a baseline Identity Discriminative Embedding (IDE) model [Zheng et al. 2016] , and confirm their participation in the overall performance.

We use Densenet-121 pre-trained on ImageNet [Deng et al., 2009] as the base network for baseline IDE model to perform a fair comparison with the proposed end-to-end ensemble model. For the IDE model, we replace the ImageNet classification layer with two sequential layers: a 1024-length fully connected embedding layer and a classification layer is appended. Cross-entropy loss is used for the classification task. Dropout is employed after gAP and embedding layers to improve the generalization capability. In the inference time, the features extracted from gAP layer is used for similarity calculation, which produces better ranking scores than the embedding layer in our experiments.

The component-based experiments for the proposed method are conducted on the three widely used ReID datasets, namely Market-1501, DukeMTMC-reid and CUHK03. The performances of the baseline model, individual base learners, and the ensemble model are presented in Table 5.5 in different configurations. Specifically, the upper and lower part of the table demonstrates the results without and with using Random Erasing (RE) data augmentation, respectively. The used distance metric

(Euclidean or Hamming) is also indicated in the table. The inferences we make from the table concerning different criteria is as follows:

- Baseline performance: DenseNet121 pre-trained on ImageNet dataset presents a strong baseline for person ReID. Notably, it produces 88% Rank-1 and 73% mAP scores on Market-1501 when random erasing is not used. On the other hand, ResNet50 baseline is reported as 85% and 65% in the literature. Adopting random erasing data augmentation further improves the baseline model and results in competitive performance with state-of-the-art techniques. One can infer that having fewer parameters than ResNet counterparts, DenseNets, are suitable for constituting the backbone architecture for person ReID.
- Base learner performance: There are 2L=16 feature extraction networks in our best model. To evaluate each base learner's performance alone, we ignore the rest and calculate the similarity based on only the outputs of the current base learner. The average performance of the base learners (Avg. Base) is reported in both Euclidean and Hamming spaces. We observe that average base learner performance is significantly better than the baseline model, particularly on the relatively small scale CUHK03 dataset. We argue that the remarkable improvements of the base learners are due to their spatially-awareness. In most ReID models, the final feature maps produced by the backbone network are subject to pooling to provide feature transferability and reduce the feature vector size, which causes loss of information. On the other hand, our non-pooled version, which considers the spatial information through fully connected layers, outperforms the baseline model. To support this argument, we present the activation maps of test images that are input to embedding layers of the baseline model and our base learners in Figure **5.3**. The baseline model focuses only on the most discriminative part, while the base learners can concentrate on other clues, which result in diversity among them as discussed below. It is also noteworthy that quantization results in inferior performance in Hamming space when the base learners are used individually, but they still improve the real-valued baseline on all datasets.
- Base learner diversity: We observe further improvements over the individual models when the embedding features of different feature extraction networks

are combined. As shown in Table 5.5, our ensemble model (Ensemble) outperforms the individual feature extractors by a large margin, especially in the mAP score. These results imply that there is sufficient divergence between base learners, and they add complementary information to the feature representation, which is crucial for ranking. Note that ensembling the base learners provide much more improvement in Hamming space: more than 5% increase in mAP scores are obtained depending on the dataset. Although the individual feature extractors under-perform in Hamming space against Euclidean space, the ensemble performance is comparable with the performance in Euclidean space. To better demonstrate the effect of ensembling, we give the individual accuracy of each base learner and the cumulative ensemble performance on each dataset in Figure 5.4. From the figure, we can observe that deeper base learners (L=9 to 16) are generally likely to perform better. There exists diversity between learners, which causes both Rank-1 and mAP scores to increase when their features are combined.

- Random Erasing: Comparing the corresponding rows in the upper and lower parts of Table 5.5, we can infer that the proposed method improves with the Random Erasing (RE) data augmentation. The data augmentation improves the scores of both the baseline and the individual learners of the ensemble. The improvement on individual learners results in increased ensemble performance on all datasets, which indicates that the proposed method is complementary to the random erasing data augmentation.

5.6.4. Comparison with conventional ensemble

The proposed method creates an ensemble model from a single deep network in an efficient way by sharing a substantial amount of the model parameters among the base learners. A conventional approach, on the other hand, requires training multiple independent networks from scratch, which we call baseline ensemble. We compare the proposed model to the baseline ensemble model which consists of up to 9 IDE models trained separately, which produces $9 * 1024$=9216 length feature vectors in test time. Figure 5.5 presents the comparison of the proposed method with the baseline ensemble model on all datasets. In the top row, CMC curves of both ensemble models and their

base learners are given. The bottom row presents the Rank-1 accuracy as a function of model complexity, which is reflected by the number of FLOPs required for a single image in test time. As shown in the figure, the individual baseline model and our ensemble model requires 2.82 GFLOPs and 2.85 GFLOPs, respectively. On the other hand, the baseline ensemble model the number of FLOPs increases with the ensemble size (2.82 GFLOPs per base learner).

Table 5.5: Ablation studies of components of the proposed model.

Method	Metric	Market-1501		DukeMTMC-reid		CUHK03-Labeled		CUHK03-Detected	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP
Baseline	Euclidean	88.75	73.62	81.72	64.94	50.37	46.41	48.40	44.31
Avg. Base	Euclidean	90.79	77.45	83.73	68.18	63.94	57.94	60.06	54.31
Avg. Base	Hamming	89.41	75.23	82.10	65.67	60.50	54.33	57.27	51.05
Ensemble	Euclidean	**91.93**	79.50	**84.83**	70.33	**67.20**	**61.90**	**62.93**	**57.56**
Ensemble	Hamming	91.80	**79.63**	84.63	**70.50**	67.03	61.80	62.86	57.40
Baseline+RE	Euclidean	91.14	77.82	83.48	68.70	58.12	53.67	56.23	51.27
Avg. Base+RE	Euclidean	92.11	79.62	85.18	70.40	67.74	62.07	63.33	57.41
Avg. Base+RE	Hamming	90.90	77.22	83.51	68.10	64.64	58.56	60.28	53.88
Ensemble+RE	Euclidean	**93.19**	82.10	**86.26**	72.63	**71.13**	66.23	**67.20**	**61.73**
Ensemble+RE	Hamming	92.73	**82.19**	85.83	**73.16**	71.06	**66.30**	67.10	60.28

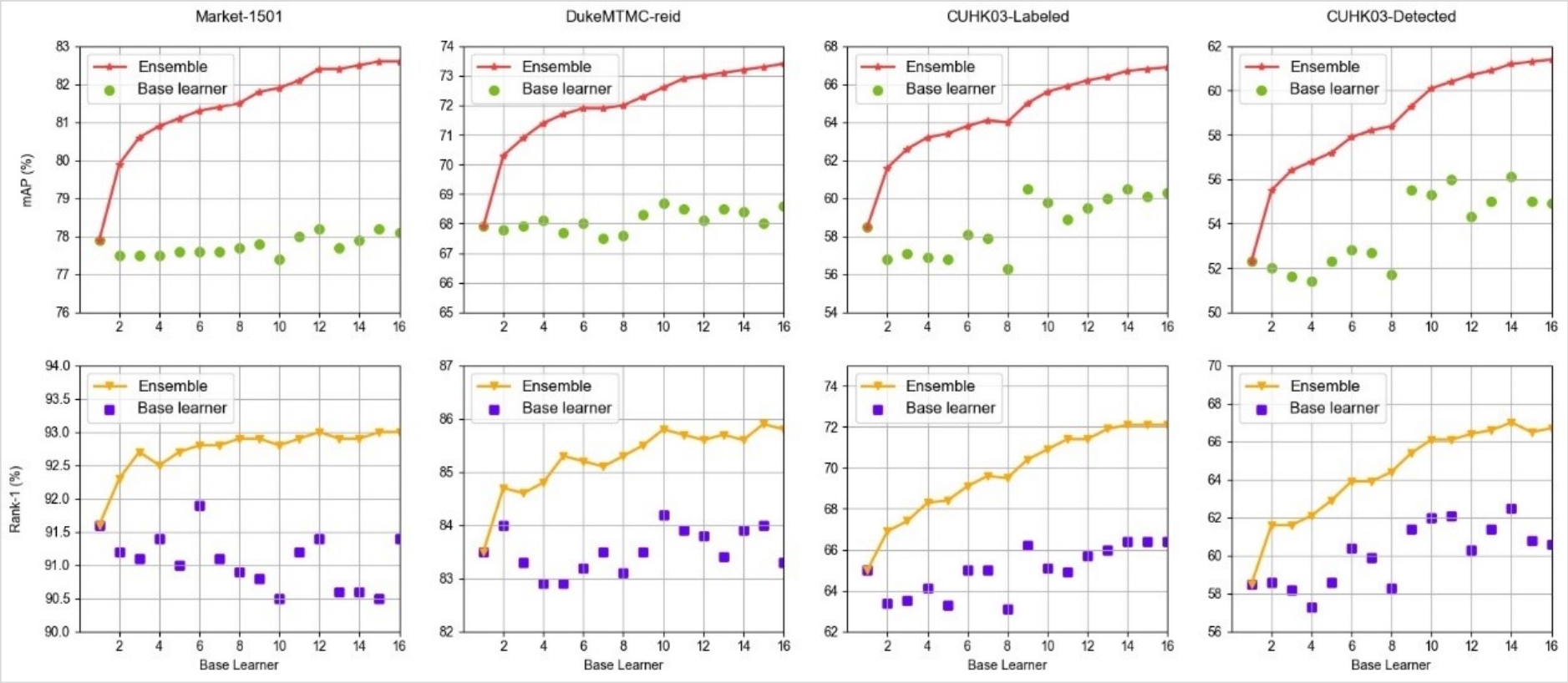

Figure 5.4: Base learner performances and cumulative ensemble performance.

As one can observe from the second row, for the large-scale datasets Market-1501 and DukeMTMC-reid, our method outperforms the baseline model significantly even though it requires nearly equal number of FLOPs. The baseline ensemble method requires several times more CPU time to perform slightly better (+1% Rank-1 score) than our model. For the relatively modest CUHK03 dataset, on the other hand, our approach yields significantly better results than the baseline ensemble (4% and 2% improvement for the labeled and detected bounding boxes, respectively) despite using smaller number of FLOPs. Furthermore, the average base learner performance on this dataset is competitive with the baseline ensemble model. We obtain base learner scores of 67% vs 66% for the labeled and 63% vs 65% for the detected bounding boxes, respectively. This result indicates that, end-to-end training of base learners is not only an effective ensembling approach but also a regularization method, which improves the individual performances of the base learners. As shown in the bottom row of the third and the fourth columns, the baseline ensemble model cannot achieve the performance of the proposed method, regardless of the number of base learners. This indicates the importance of weight sharing on small datasets. When a single network is trained on small datasets, it may suffer from overfitting even with model combination.

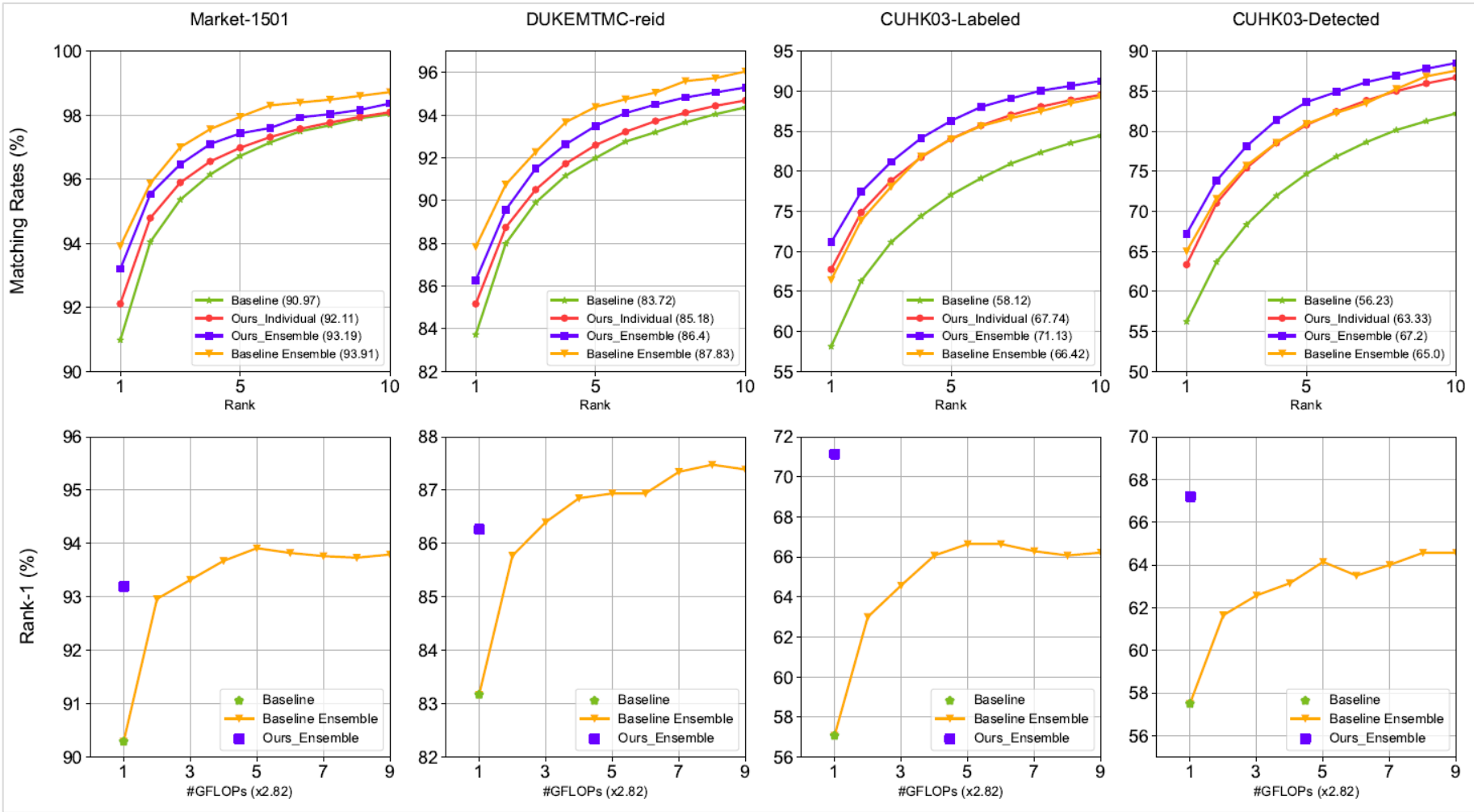

Figure 5.5: Comparison with conventional ensemble.

5.6.5. Hyper-Parameter Analysis and Ensemble of Ensembles

There are two hyper-parameters that are additional to the standard baseline hyper-parameters: the ensemble size and the size of feature vectors. We analyze the effects of these hyper-parameters on Market-1501 dataset in

Figure 5.6 left and right, respectively. The hyper-parameter ensemble size corresponds to the number of base learners and the feature vector size is the number of nodes in the embedding layer of each sub-networks. In order to fix the length of the final feature representation, we tune these hyper-parameters inter-dependently, e.g., for 8 learners the feature vector size is 1024. Note that, for feature vector size analysis, we fix our ensemble model to 16 learners and report the results accordingly.

Figure 5.6 represents that our approach performs better than the baseline model (Table 4.3) for varying hyper-parameters, which indicates that the model is not sensitive to the hyper-parameters. Therefore, in case of computational constraints, the model can be switched to a more compact version by setting the embedding size and ensemble size as 64 and 16, respectively (i.e. 64×16=1024).

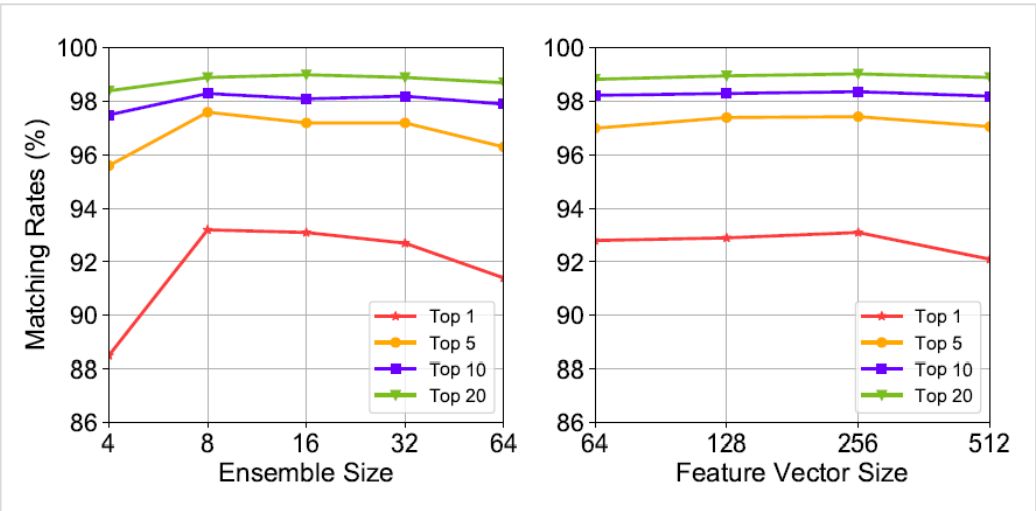

Figure 5.6: Hyper-parameter analysis on Market-1501.

To investigate whether our method further improves with model combination, we treat our ensemble model as a standalone base learner and create an ensemble of our model which we call ensemble of ensembles. It is noteworthy that due to the similar complexity of our model and that of the baseline model, ensemble of ensembles

is computationally comparable to the baseline ensemble model (2.82 vs 2.85 GFLOPs per learner). The comparison between our ensemble of ensembles model and the baseline ensemble is given in Figure 5.7 reporting mAP scores according to the ensemble size (thus the model complexity). As shown in the figure, our method still improves when used as a base learner and combined with other sister models. In addition, the ensemble of ensemble model performs superior to the conventional approach. This result shows that the proposed method is convenient as a base learner as itself compared to the baseline model and has more space to improve.

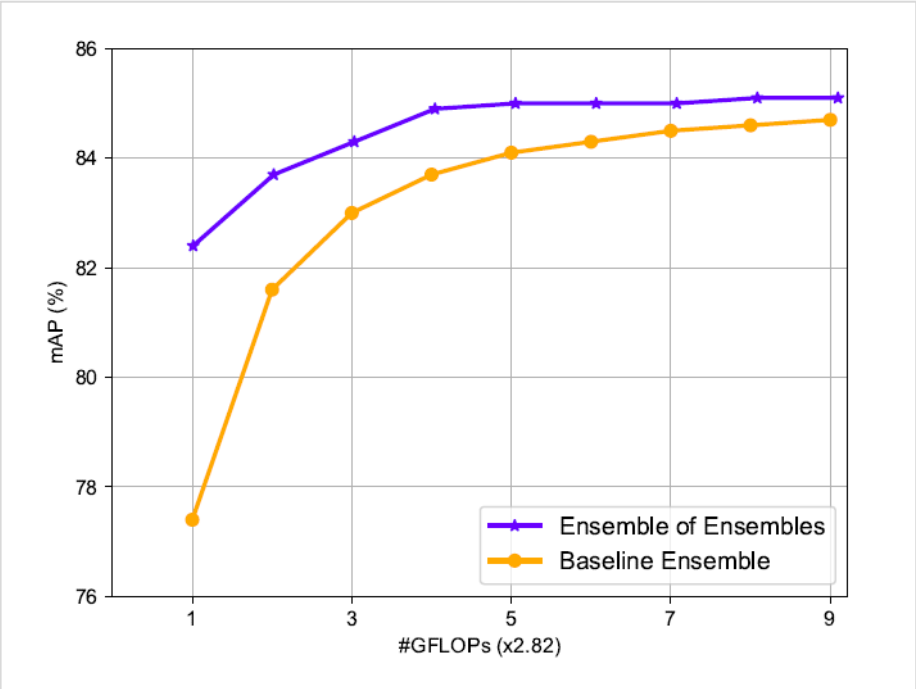

Figure 5.7: Performance of ensemble of ensemble model.

5.6.6. ResNet50 as the Backbone

In this section, we investigate the effectiveness of the proposed model on the other backbones. To this end, we adapt it to the widely used ResNet50 architecture. Our method is based on DenseNet architecture, which produces compact feature maps (32 filters) after each convolution layer. Since these compact feature maps represent the whole image, they are very informative and discriminative enough to be used as

the inputs to the base classifiers. The embedding features trained in this way results in diverse and accurate base learners as shown above.

Compared to the DenseNet, ResNet50 produces a huge number of feature maps (2048 filters) after the last bottleneck blocks. If we attempt to place base learners after these blocks, the number of required parameters would become too large. In order to apply the same model on the ResNet50 architecture, we partition the output channels of the bottleneck layers and place our base learners on top of these channels. The overall model is shown in Figure 5.8. As shown in the figure, ResNet50 has three bottleneck blocks whose output channels are split into multiple branches. At the end, spatially aware sub-networks (SN) are integrated on top of these channels. We partition each block's output into 8 and obtain 24 base learners in total. The sub-networks have the same structure as the DenseNet model, which is shown in Figure 5.2.

We compare this model to the baseline and some representative state-of-the-art models in the upper part of Table 5.6. We report the baseline ResNet50 model which is the IDE model introduced before. Then, we apply Random Erasing data augmentation and our method as additional tricks on the baseline model. The results indicate that our model is complementary to Random Erasing data augmentation and further improves the performance. As shown in the table, our method still produces state-of-the-art scores for Market-1501 and DukeMTMC-reid datasets when ResNet50 is used as the backbone. Moreover, it significantly outperforms the existing studies on the relatively small CUHK03 datasets and it achieves the second-best scores after our DenseNet model.

We perform another experiment to investigate whether our method further improves a strong baseline which applies some tricks proposed in the literature [Luo et al., 2019], which they call Bag of Tricks. Therefore, we integrate our end-to-end ensemble model as a trick to this baseline model. The lower part of Table 5.6 presents the results. In the first row, the scores of the strong baseline proposed by [Luo et al., 2019] is given. As shown in the bottom row of the table, our method further improves this baseline when applied as an additional trick. The improvement on CUHK03 dataset is more noticeable.

We also observe improvements in the training process. Specifically, our ensemble model converges faster and presents a smooth learning curve compared to

Bag of Tricks approach. Figure 5.9 presents the Rank-1 and mAP scores on all datasets during different training stages of both our model and the strong baseline model. Our model performs favorably and converges by 60K iterations, where the Bag of Tricks performs best after 120K iterations.

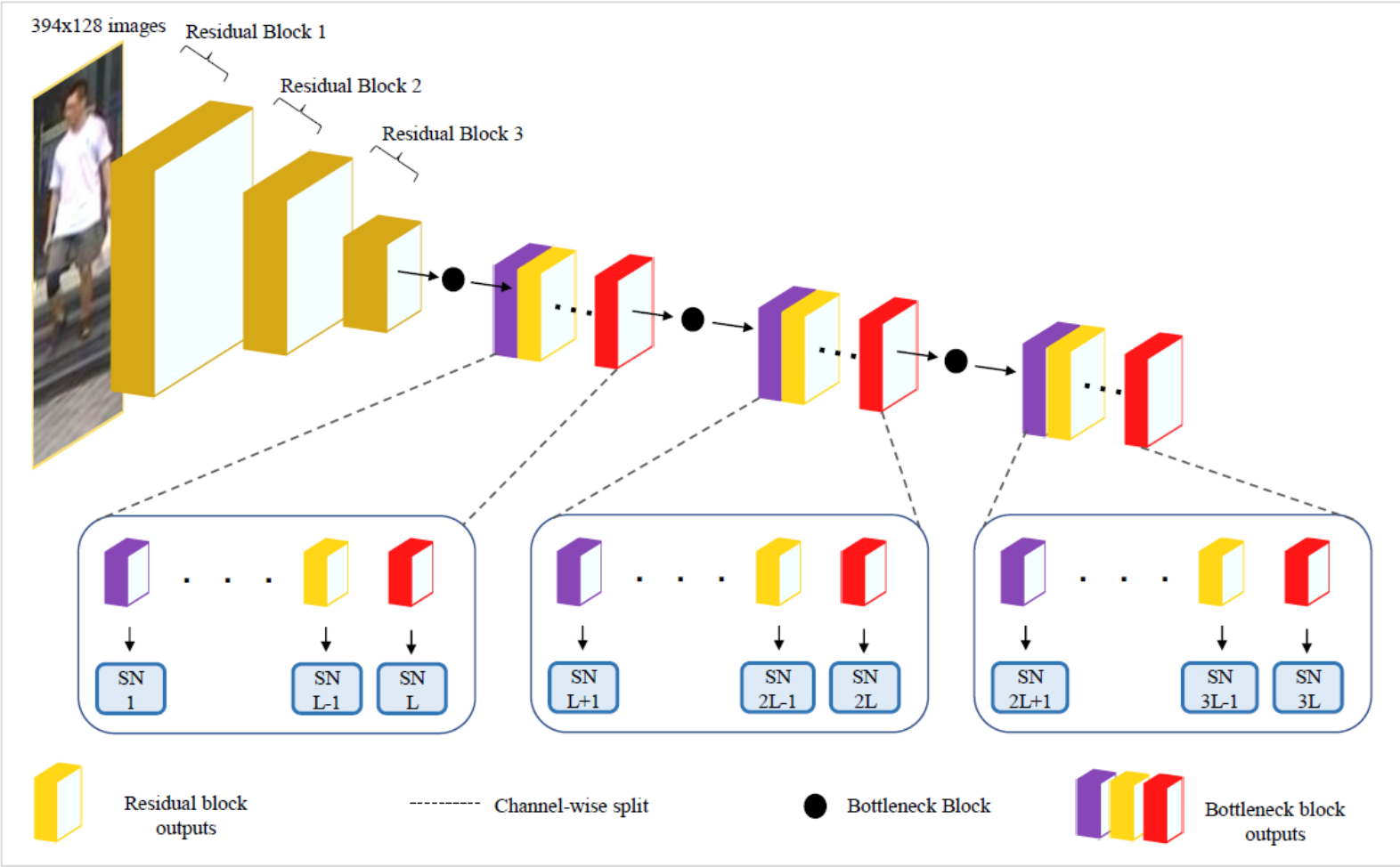

Figure 5.8: Adaptation of the proposed model on ResNet50.

Table 5.6: Comparison with the state-of-the-art when ResNet50 is the backbone.

Method	Market-1501		DukeMTMC-reid		CUHK03-Labeled		CUHK03-Detected	
	R-1	mAP	R1	mAP	R1	mAP	R1	mAP
DuATM* [Si et al., 2018]	91.42	76.62	81.82	64.58	-	-	-	-
GP** [Almazan et al., 2018]	92.20	81.20	85.20	72.80	-	-	-	-
PCB [Sun et al., 2018]	92.40	77.30	81.90	65.30	-	-	61.30	54.20
MultiBranch [Zhai et al., 2018]	93.10	78.90	84.00	68.40	-	-	61.70	55.30
PCB+RPP [Sun et al., 2018]	**93.80**	81.60	83.30	69.20	-	-	63.70	57.50
ResNet50	88.03	70.93	77.90	58.87	48.07	43.57	45.17	41.07
+RE	89.63	75.70	82.50	66.60	57.37	52.73	56.77	51.77
+Ours	91.76	80.16	84.86	70.43	68.70	64.83	65.13	60.16
Bag of Tricks [Luo et al., 2019]	**94.50**	86.05	86.69	76.49	71.23	69.59	68.21	66.41
+Ours	93.92	**86.11**	**87.44**	**77.23**	**75.63**	**73.95**	**72.50**	**70.49**

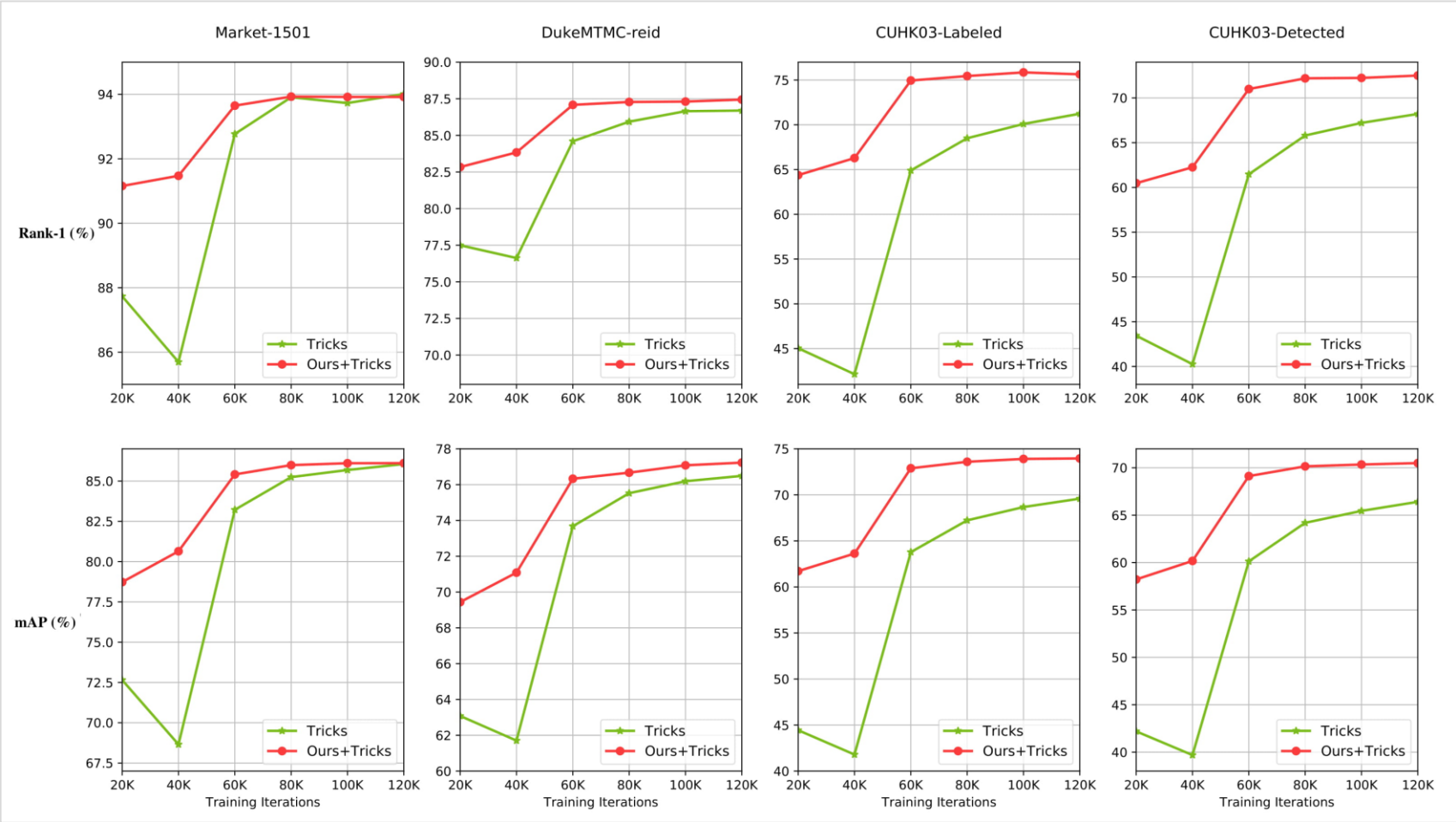

Figure 5.9: The model performance in different training stages.

6. END-TO-END BNN ENSEMBLES

In this section, we adapt our proposed ensemble model to Binary Neural Networks (BNNs). First, we give a brief background and overview of BNNs and explain our motivation for adapting the proposed model on this type of neural network. Then, we evaluate our model on image classification task using CIFAR10 dataset. We report our initial results and inferences, and analyze our model's performance on the classification task, its stability during training, and robustness to input perturbations. Then, we adopt our end-to-end BNN ensemble model for person ReID, and experimentally show that it outperforms conventional ensemble model by a large margin, which indicates that besides being an efficient ensemble model, the proposed approach has an intrinsic regularization effect.

6.1. Motivation

Deep neural networks are typically resource-intensive and require expensive GPU-based machines, which prevents their training and deployment on mobile and embedded devices [Guo et al., 2017], [Rastegari et al., 2016]. Network compression and quantization have been studied extensively to reduce the storage and memory space of deep networks [Ullrich et al., 2017], [Luo et al., 2017], [Belagiannis et al., 2019], [Tung and Mori, 2020]. Deep neural networks with low bitwidth [Zhou et al., 2016] or binary weights [Courbariaux et al., 2015] have been proposed to reduce the computational consumption, and they show near the state-of-the-art performance of full-precision counterparts.

Following the advances mentioned above, [Hubara et al., 2016] have successfully trained binary networks (BinaryNet), whose weights and activations both take only binary values during the forward pass. The binary neural networks have been evaluated on large scale image dataset in [Rastegari et al., 2016] for the first time. Besides the advantage of saving storage and memory space, BNN's perform only bitwise operations at run-time, and when computing the gradients at train-time. It is shown that binarizing only the weights reduces the memory consumption by 32x, and binarizing both the weights and activations allows 58x faster convolutional operations [Rastegari et al., 2016]. This makes BNNs one of the most promising techniques in training and deploying DNN's in low-end devices [Zhu et al., 2019]. However, there

is still much more room to improve the binary neural networks compared to full-precision networks.

In the present study, the improvement obtained in the Hamming space has encouraged us to evaluate the proposed ensemble model on the BNN's. In Section 5.6.3, we have shown that the proposed model outperforms the base learners more significantly when the feature extraction layer is trained with *tanh* activation function during training, and the features are quantized into binary values in run-time. The results indicate that quantization provides more diversity on the feature vectors, which enables error compensation in the run-time ranking. On the other hand, BNNs typically suffer from instability during training and have robustness issue in run-time due to quantization [Zhu et al., 2019]. Therefore, the proposed model is promising to be used as a regularizer in BNNs.

6.2. Experiments on Image Classification

In this section, the proposed model is adapted for image classification, and the initial findings are reported. We evaluate our method on CIFAR10 dataset using the Network in Network (NIN) architecture. We use the binarization method proposed by Rastegari et al. [Rastegari et al., 2016], [Web-3, 2020]. The adaptation of the proposed ensemble model is shown in Figure 6.1. We append multiple convolution and classification layers on top of the output of the Network in Network architecture, resulting in a multi-branch architecture, where each branch is a base learner. It is noteworthy that there is no weight-sharing between the branches to promote diversity, while the NIN backbone is shared. The outputs of the classification layers are combined by majority voting in test time.

Our proposed model is closely related to the work of [Zhu et al., 2019], where they train multiple independent BNNs to obtain a BNN ensemble; that is, the weights of NIN backbone are not shared between the base learners. They also provide several experimental analysis to demonstrate the main issues in BNNs. We mainly follow the same approach to investigate the effectiveness of our approach in BNNs. Since binarizing the first and last layers results in severe accuracy degradation, previous work adopted full-precision in the first and last layers. In this thesis study, we use two network experimental settings following [Zhu et al., 2019]: 1) All-Binary (AB), where

all layers are binarized, 2) Semi-Binary (SB), where all layers are binarized except the first and last layers.

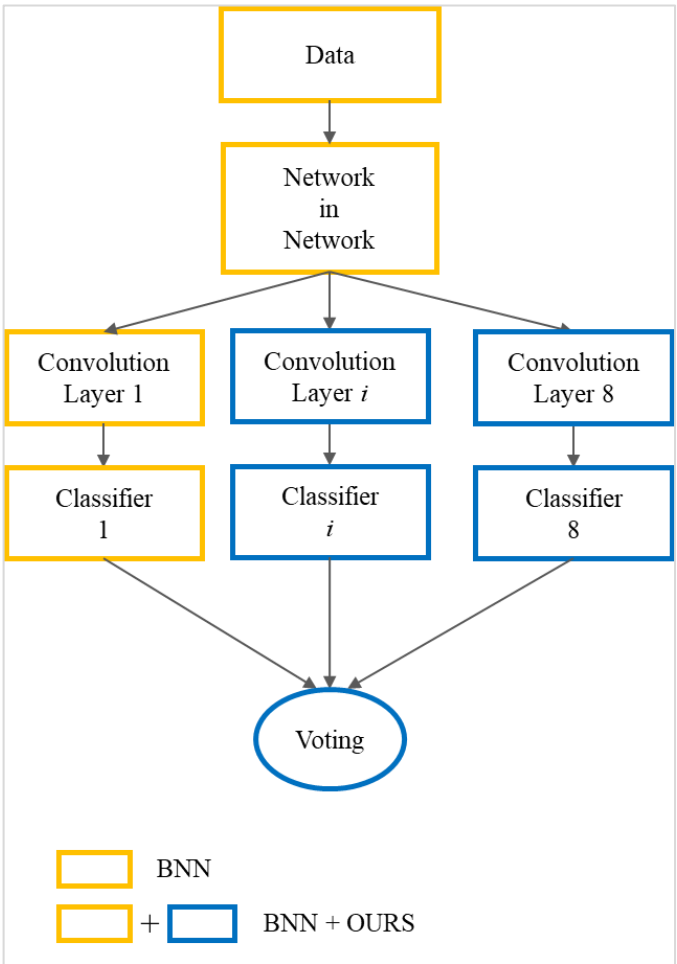

Figure 6.1: Adaptation of the proposed model to image classification.

6.2.1. Classification Performance

We compare our method with the BNN baseline for AB and SB setting in Table 6.1. Our approach produces comparable scores with the baseline model in SB setting. The baseline model slightly outperforms our model when the learning rate is high (i.e., 0.01), but they perform similarly for lower learning rates. On the other hand, our method improves the baseline model in AB setting consistently. In particular, it provides more than 2.5% improvement when the learning rate is high. These results are consistent with the inferences made in the Hamming space for ReID task. The proposed end-to-end ensemble model performs effectively in case of quantization.

Table 6.1: Comparison of our method with BNN baseline.

	learning rate		
	0.01	0.001	0.0001
AB	74.06	69.98	66.32
Ours AB	76.78	71.89	67.82
SB	83.46	81.51	75.87
Ours SB	82.28	80.86	75.42

6.2.2. Stability Analysis

Following [Zhu et al., 2019], we also compare the stability of the networks by measuring the fluctuation of the test accuracy after training the networks for 300 epochs. We continue to train for 20 epochs and measure the fluctuation on the test set after each epoch. However, there are many different experimental setups in the work of [Zhu et al., 2019], and it is not specified which one is used in the stability analysis. As a result, our baseline scores on the stability analysis are not compatible with their reported ones. Therefore, we compare our method with our baseline model.

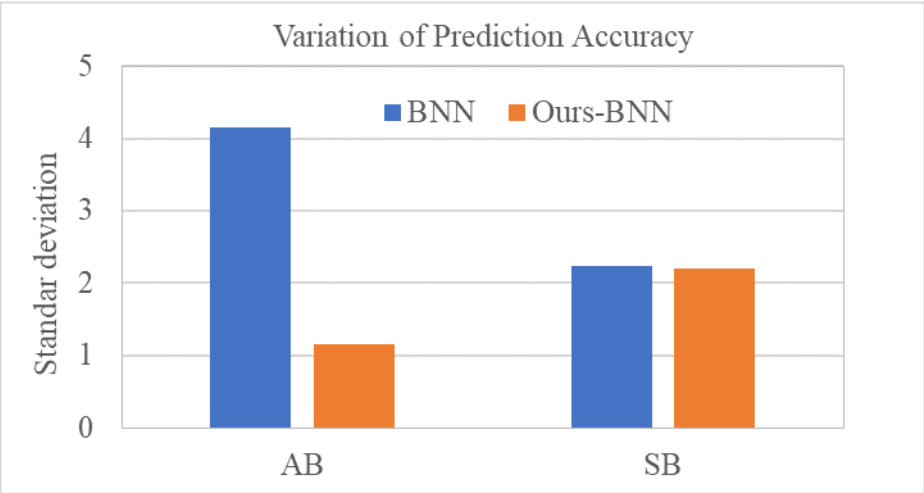

Figure 6.2 Stability comparison of our method with BNN baseline.

As shown in Figure 6.2, BNNs suffer from high fluctuation even after a large number of training iterations, especially when All-Binary setting is used. On the other hand, our method significantly improves the AB model in terms of training stability. Interestingly, the stability of the proposed model in AB setting is better than SB setting

on both baseline model and our model. These results imply that the proposed model is a strong regularizer for BNNs and provide sufficient diversity for the base learners in AB setting, which results in adequate error compensation. On the other hand, the proposed model has a negligible effect in SB setting, which indicates that it cannot fulfill the diversity requirement. Therefore, the proposed model needs further investigation on improving the base learner diversity in SB setting.

6.2.3. Robustness Analysis

Robustness is defined as the property that producing similar errors for the test and training samples if they are close to each other [Xu and Mannor, 2012]. In this section, we perform robustness analysis where we measure the sensitivity of the networks to input variations. To this end, following [Zhu et al., 2019], after obtaining the outputs for 250 test images, we inject input perturbation Δx on each test image by a Gaussian noise with standard deviation 0.01, run a forward pass and measure the expected $l2$ norm of the change on the output distribution.

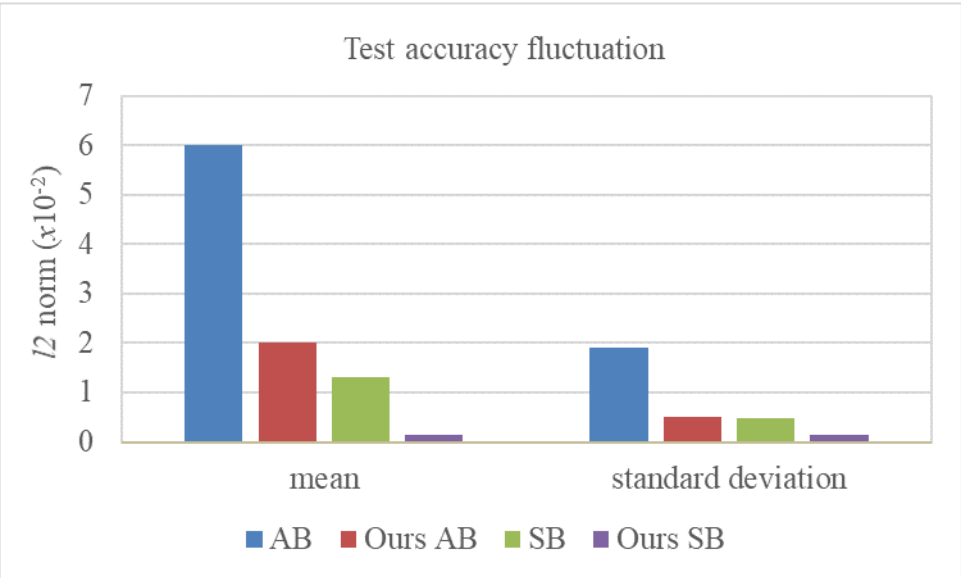

Figure 6.3 Robustness comparison of our method with BNN baseline.

We compare our model with BNN baseline for both AB and SB settings in Figure 6.3. As expected, BNN baseline in All-Binary (AB) setting is significantly more sensitive to the input perturbations compared to SB setting. On the other hand,

we observe that our ensemble model addresses the sensitivity of the AB model effectively and accomplish to reduce the magnitude of the output variation considerably. Moreover, it improves the robustness of the BNNs in SB setting.

The results of the experiments that are reported in this section indicate that the proposed end-to-end ensemble model is promising in improving the performance of BNNs in terms of accuracy, stability, and robustness. These results are consistent with very recent work of [Zhang et al., 2018], which shows that multi-branch architectures are less non-convex. However, we observe that simple adaptation of the proposed model into the SB model has no noticeable effect on the model accuracy and training stability, which needs further investigation.

6.3. Experiments on Person ReID

In this section, we adapt a deeper BNN model to evaluate the end-to-end BNN ensemble's performance on person ReID. We compare our method with the BNN baseline as well as a conventional ensemble of this baseline. As the baseline model, we use Bi-RealNet [Liu et al., 2018], a binarized version of ResNet model which improves the previous BNN models on the large-scale image classification task by . We use ResNet18 model proposed in [Liu et al., 2018], where all layers are binarized except the first and last layers. We append a spatially aware embedding layer before the classification layer for fair comparison with our model.

To convert the baseline model to our end-to-end ensemble model, we follow the same method introduced in Section 5.6.6. Specifically, we append multiple spatially-aware base learners on top of separate channel splits in the last ResNet block, which consists of 4 binary convolutional layers. The sub-networks for constructing the base learners are made up of a 1D binary convolutional layer, which serves as the feature extraction layer and a real-valued classification layer. During the inference, only the input layer requires real-valued operation because classification layer is not used.

To compare our model with conventional ensemble model, we train 8 parallel baseline models and combine their feature extraction layer outputs. We use an embedding vector size of 1024, which ends up with a 8192 length feature vector. On the other hand, our best end-to-end ensemble model consists of $8\times4=32$ base learners each of which produces 128-bit feature vectors, resulting in a 4096-length feature vector.

Table 6.2: Comparison with classical ensemble of baseline BNN.

Model	R-1	mAP
Baseline Avg.	51.22	24.57
Baseline Ensemble	60.40	34.30
Our Base Learner Avg.	53.65	28.57
Our Ensemble	71.00	47.30

We perform our experiments on Market1501 dataset. In Table 6.2, we compare our model with baseline BNN, and a conventional ensemble of this baseline model. Our average base learner performs better than the baseline model, which demonstrates that our model is an effective regularization technique. Moreover, our ensemble model surpasses the classical ensemble model by 11% in Rank-1 and 13% in mAP scores, which indicates that there is divergency between base learners. We also report the cumulative performance of our model with conventional ensemble model in Figure **6.4**. As shown in the figure, our model improves smoothly as opposed to the conventional model, which presents fluctuation due to bad performance of the baseline models.

We perform an experiment to observe whether the performance improves if we embed more base learners into our model. Specifically, after each layer in the last ResNet block we partition the output feature maps into overlapping splits and append sub-networks after each split. We keep the embedding feature size and the input feature map size the same with the previous experiments The results are given in Figure **6.5**, which shows that the performance can be increased in this way by using up to 24 base learners.

The experiments in this section have shown that our model, besides being a effective and efficient ensemble model, is a promising way of regularizing BNN models.

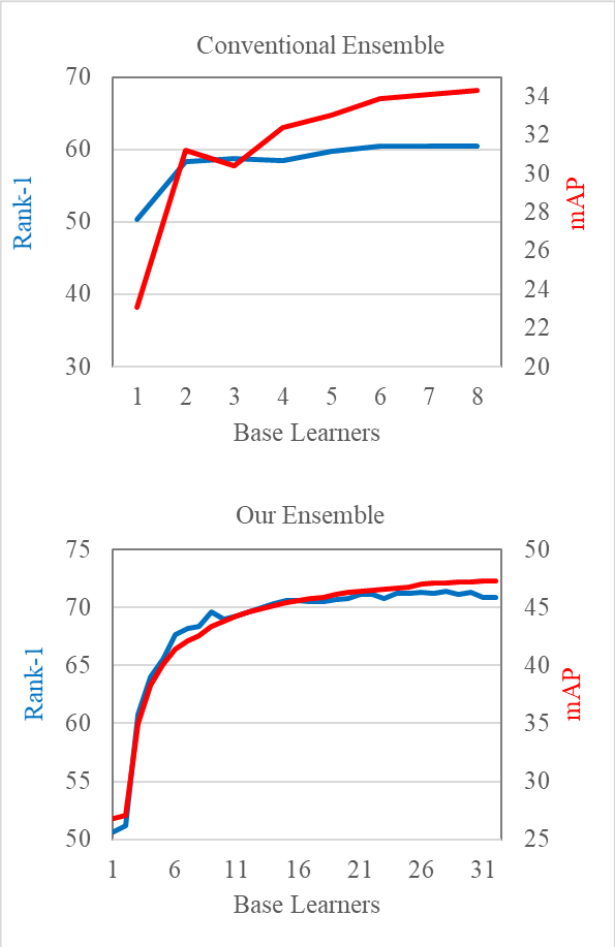

Figure 6.4: Cumulative performance of classical ensemble and our ensemble.

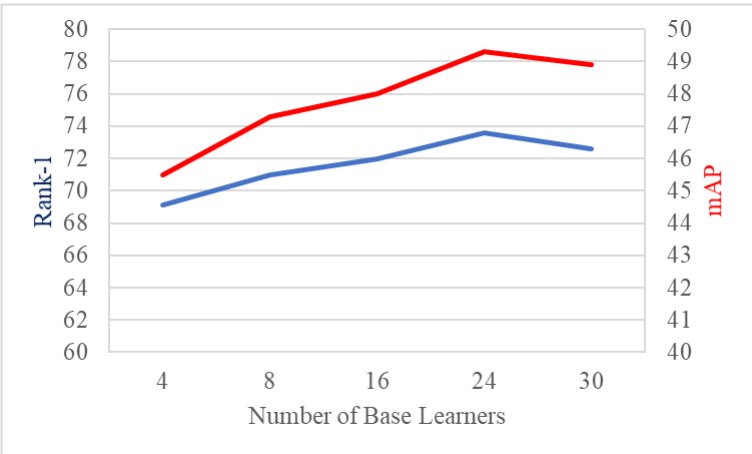

Figure 6.5: Convergence in performance as the ensemble size increases.

7. CONCLUSION AND FUTURE WORK

7.1. Conclusion

This thesis has aimed at finding novel regularization methods for person ReID problem. To this end, we designed an ensemble learning model of deep networks that consists of many diverse and accurate base learners. The computational overhead of training multiple deep networks is avoided by embedding multiple base learners in a single network. Specifically, a multi-branch architecture is designed by integrating individual sub-networks on top of different network stages. Each sub-network, together with the shared backbone architecture, constitutes a base learner. The base learners are trained jointly in an end-to-end manner, which ended up an efficient ensemble model.

The detailed experiments demonstrate that the base learners are diverse and accurate enough so that they form a convenient ensemble model. We observe significant improvement in the ReID accuracy when the feature representations of multiple base learners are combined. As a result, the proposed model outperforms most state-of-the-art approaches on four benchmark dataset. Moreover, the impressive performance gain on the relatively small scale CUHK03 dataset indicates that the proposed model effectively addresses the overfitting problem. The analysis of the computational cost of the proposed model shows that our model adds a negligible number of FLOPs over the baseline model. Moreover, it is very efficient compared to the conventional ensemble model, which consists of multiple deep networks trained independently.

We have also evaluated the performance in Hamming space to allow fast similarity calculation between the query and gallery images. The experiments demonstrate that our approach is much more effective in this space. Although the base learners are weaker compared to the full-precision embedding, their combination shows comparable performance, which indicates that there is sufficient diversity of between base learners.

One of the most critical features of the proposed approach is its easy applicability to other problems. Since it has no ReID-specific sub-modules, and relies only on the backbone architecture, one can easily adapt the model for other tasks such as image classification or image retrieval. On the basis of its generability, and the comparable

performance in Hamming space, we have adapted our model into BNN's for image classification task. Our initial experiments have shown that the proposed approach improves the performance of BNN's in terms of accuracy and training stability when all weights and activation are binarized. We have also performed a sensitivity analysis to the input variations, where the experiments demonstrate that our model significantly improves the model robustness.

Ensemble learning is a widely used method for regularizing the learning models and has a rich literature with many applications. The computational overhead of deep networks hinders designing ensemble of deep networks in a conventional way. Combining deep networks with ensemble learning has only recently started to emerge. Considering the challenges, we can expect more research effort to be made in the future for combining other ensembling techniques with deep learning and improving their performance in terms of accuracy and efficiency.

7.2. FUTURE WORK

In this section, we first identify some future directions which may utilize from the proposed model to overcome some task-specific problems in person ReID, such as end-to-end or unsupervised ReID. These problems have attracted more attention in recent years, and the proposed model should be adapted accordingly to keep it up to date. Secondly, we propose some further research subjects from machine learning or computational perspectives regarding the proposed ensemble model.

7.2.1. Task-Specific Problems

A natural direction of future research is to improve the proposed model to overcome some specific problems in person ReID. Below, we review these significant issues, which are very attractive research topics in the ReID community.

7.2.1.1.Open-Set ReID

The person ReID models, which have shown a significant improvement in recent years, are mostly based on the closed-set scenario. In the closed-set scenario, it is assumed that a known set of people all appear in the view of each camera [Cancela, et

al., 2014]. This assumption implies that every probe person certainly exists in the gallery set, and the ultimate goal is, given a probe image, retrieving the most similar person from the gallery set [Liao et al., 2014]. The closed-set scenario considerably simplifies the ReID problem, in which the previous work has shown tremendous improvement.

However, the above-mentioned assumption does not hold in the practical applications, such as searching a suspect in a video. In a real-world ReID system, the identities of people that appear in different cameras may or may not overlap. Thus, the probe person can or cannot match with a person in the gallery set. In other words, there are more possible outcomes in this unconstrained setting [Cancela et al., 2014], which makes the problem even more challenging. Therefore, a real-world person ReID problem can be decomposed into two sub-tasks: detection and identification. To deal with the large gallery sets, a practical model should first determine the videos in which the probe person appears (detection), and then retrieve the corresponding frames.

The difficulty of the open-set scenario has led most researchers to deal with the simplified closed-set version. As a result of the significant improvement in the closed-set scenario, the attention to the more challenging open-set scenario has increased in the ReID community [Leng et al., 2019].

The open-set ReID problem requires making a hard decision (match/no match) during the detection process as opposed to the closed-set problem, where the final decision is based on the soft computation of similarity scores. Therefore, a strong decision-making process is required, which should not be solely relied on a single classifier. Ensemble models are very appropriate tools which can be utilized for improving the decision-making process in such problems. In order to deal with large video gallery sets, compact ensemble models with improved performance may provide significant performance gain. The proposed end-to-end ensemble model, which shares a substantial number of convolutional operations among the base learners can be adapted and improved to be used for decision making in the open-set person ReID problem.

7.2.1.2.End-to-End ReID

A practical person ReID system takes the raw videos as input and performs person detection and tracking along with re-identification [Zheng et al., 2016]. It has been shown that integrating person detection and re-identification can improve model performance [Xu et al. , 2014]. Therefore, the proposed model can be adapted for joint learning of these two tasks. The base learners in an ensemble model may compensate for the mistakes of each other's in bounding box detection, and this may further improve the ReID accuracy.

7.2.1.3.Domain-Invariant or Unsupervised Person ReID

A trained person ReID model must be transferable to new domains where there is no labeled training data because annotating data for each domain is impracticable. However, domain discrepancy causes a significant performance drop. Therefore, the generalization ability of the trained model is of crucial importance in ReID systems. Therefore, a considerable amount of research has been invested in unsupervised domain adaptation, where a trained model is adapted to a target domain via unsupervised training or a domain-invariant model is proposed which benefits from both labeled source domain and unlabeled target domains [Zhong et al., 2019], [Song et al., 2019]. One future research direction is to adapt our model to unsupervised domain adaptation.

7.2.2. Further Directions

In this section, we suggest some future research subjects to improve the model efficiency and effectiveness using tools from machine learning and deep learning research.

7.2.2.1.More Ensemble Methods

A potential research direction may aim to incorporate different ensembling approaches to improve the end-to-end training strategy in the proposed model. For example, to increase the base learner diversity, bagging is a widely used model, where

each base learner is trained with a variant of the training data, which is via sampling by replacement. In this way, each base learner slightly differs from the others without compromising from the accuracy. Another ensembling scheme that can be employed is boosting, where each base learner aims to compensate for the previous learner's mistakes on training data. This is a more effective strategy for obtaining diverse base learners. These methods can be adapted in end-to-end CNN ensembles by calculating the loss of each base learner based on weighted training instances.

7.2.2.2.Binary Neural Networks

Binary Neural Networks (BNNs) is one of the most promising topics for training and deployment of deep models on low-end portable devices. A BNN uses only bitwise operations during the forward pass by quantizing the weights and activations to binary values and operates in full-precision during the calculation of gradients. This makes them very favorable compared to the full-precision networks, which are very expensive to train and test. BNNs theoretically enable 52x faster convolutional computation.

Compression rate and accuracy present a trade-off in BNN's. Nevertheless, many advances have been accomplished [Rastegari et al., 2016] . Recent research has shown that binary neural networks can be hardware accelerated [Conti et al., 2018], so the area is open to further improvements. [Zhu et al., 2019] have shown that ensembling is an effective strategy for obtaining accurate binary neural networks. This finding is consistent with our experimental results, which demonstrates that it is more enhancing to ensemble the binarized features than the high precision ones. Therefore, more research effort should be made on adapting the end-to-end ensemble approach for binary neural networks, which is a promising approach.

7.2.2.3.New Problems

The proposed end-to-end ensemble model increases the mAP and Rank-1 scores for person ReID, where the ultimate task is ranking. The results indicate that the proposed model can be adapted to related tasks such as image retrieval where the final goal is similarity ranking.

In this thesis, the proposed end-to-end ensemble model is adapted for image classification with BNN's. The initial results are promising and show that our model improves the classification accuracy, robustness, and stability of the BNN's. Future work should investigate the effects on large-scale image classification task, which suffers from severe overfitting

We have further investigated whether the end-to-end ensemble model improves the image classification task and perform experiments on CIFAR-10. We observed some improvements in the training stability and convergence such as the ensemble model is more robust to hyper-parameter changes and converges faster, which is consistent with the recent research [Zhang et al., 2018]. However, the classification accuracy remained the same. This result implies that the base learners are not diverse enough to compensate for the mistakes of each other in the classification layer. Therefore, more research effort should be made to increase the base learner diversity while adapting the end-to-end ensemble model for other tasks.

REFERENCES

Ahmed E., Jones M., Marks T. K., (2015), “An improved deep learning architecture for person re-identification”, IEEE Conference on Computer Vision and Pattern Recognition, 3908–3916.

Ainam J. P., Qin K., Liu G., Luo G., (2019), “Sparse Label Smoothing Regularization for Person Re-Identification”, IEEE Access, 7, 27899–27910.

Almazan J., Gajic B., Murray N., Larlus D., (2018), “Re-ID done right: towards good practices for person re-identification”, ArXiv Preprint ArXiv:1801.05339.

Alpaydin E., (1993), “Multiple networks for function learning”, IEEE International Conference on Neural Networks, 9–14.

Alpaydin E., Kaynak C., (1998), “Cascading classifiers”, Kybernetika, 34(4), 369–374.

Andriluka M., Pishchulin L., Gehler P., Schiele B., (2014), “2d human pose estimation: New benchmark and state of the art analysis”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3686–3693.

Balntas V., Riba E., Ponsa D., Mikolajczyk K., (2016), “Learning local feature descriptors with triplets and shallow convolutional neural networks”, Proceedings of the British Machine Vision Conference, 1(2), 3.

Baltieri D., Vezzani R., Cucchiara R., (2011), “” 3DPeS: 3D people dataset for surveillance and forensics. MM’11 - Proceedings of the 2011 ACM Multimedia Conference and Co-Located Workshops 59–64.

Barron A. R., Barron R. L., (1988), “Statistical learning networks: A unifying view”, Symposium on the Interface: Statistics and Computing Science.

Belagiannis V., Farshad A., Galasso F., (2019), “Adversarial network compression”, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 431–449.

Bengio Y., Lamblin P., Popovici D., Larochelle H., (2007), “Greedy layer-wise training of deep networks”. Advances in Neural Information Processing Systems, 153–160.

Bishop C. M., (1995), “Training with noise is equivalent to Tikhonov regularization”, Neural Computation, 7(1), 108–116.

Boulgouris N. V., Hatzinakos D., Plataniotis K. N., (2005), “Gait recognition: A challening signal processing technology for biometric identification”, IEEE Signal Processing Magazine, Vol. 22, 78–90.

Breiman L., (1996a), “Bagging predictors”, Machine Learning, 24(2), 123–140.

Breiman L., (1996b), “Bias, variance, and arcing classifiers”, Tech. Rep. 460, Statistics Department, University of California, Berkeley, CA, USA..

Breiman L., (1996c), “Heuristics of instability and stabilization in model selection”, Annals of Statistics, 24(6), 2350–2383.

Breiman L., (2001), “Random forests”, Machine Learning, 45(1), 5–32.

Cai Z., Vasconcelos N,., (2018), “Cascade R-CNN: Delving into High Quality Object Detection”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 6154–6162.

Cancela B., Hospedales T. M., Gong S., (2014), “Open-world person re-identification by multi-label assignment inference”, Proceedings of the British Machine Vision Conference 2014.

Chang X., Hospedales T. M., Xiang T., (2018), “Multi-level factorisation net for person re-identification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2109–2118.

Chen T., Guestrin C., (2016), “XGBoost: A scalable tree boosting system”, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Chen W., Chen X., Zhang J., Huang, K., (2017a) , “A multi-task deep network for person re-identification”, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 3988–3994.

Chen W., Chen X., Zhang J., Huang K., (2017b) , “Beyond triplet loss: A deep quadruplet network for person re-identification”, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, 1320–1329.

Chen Y.C., Zhu X., Zheng W.S., Lai J.H., (2018) , “Person re-identification by camera correlation aware feature augmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(2), 392–408.

Chen Y., Zhu X., Gong S., (2017) , “Person re-identification by deep learning multi-scale representations”, Proceedings of the IEEE International Conference on Computer Vision, 2590–2600.

Cheng D., Gong Y., Zhou S., Wang J., Zheng N., (2016) , “Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function”, IEEE Conference on Computer Vision and Pattern Recognition, 1335–1344.

Chung D., Delp E. J., (2019) , “Camera-aware image-to-image translation using similarity preserving stargan for person re-identification”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 1517–1525.

Conti F., Schiavone P. D., Benini L., (2018) , “XNOR Neural engine: A hardware accelerator IP for 21.6-fJ/op binary neural network inference”, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 37(11), 2940–2951.

Courbariaux M., Bengio Y., David J. P., (2015), “Binaryconnect: Training deep neural networks with binary weights during propagations”, Advances in Neural Information Processing Systems, 3123–3131.

Dai Z., Chen M., Gu X., Zhu S., Tan P., (2019), “Batch dropblock network for person re-identification and beyond”, Proceedings of the IEEE International Conference on Computer Vision, 3690–3700.

Deng J., Dong W., Socher R., Li L.J., Li K., Fei-Fei L., (2009) , “Imagenet: A large-scale hierarchical image database”, Proceedings of the IEEE International Conference on Computer Vision, 248–255.

Deng W., Zheng L., Ye Q., Kang G., Yang Y., Jiao J., (2018) , “Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 994–1003.

Eigen D., Puhrsch C., Fergus R., (2014), “Depth map prediction from a single image using a multi-scale deep network”, Advances in Neural Information Processing Systems, 2366–2374.

Fan X., Jiang W., Luo H., Fei M., (2019), “SphereReID: Deep hypersphere manifold embedding for person re-identification”, Journal of Visual Communication and Image Representation, 60, 51–58.

Farenzena M., Bazzani L., Perina A., Murino V., Cristani M., (2010), “Person re-identification by symmetry-driven accumulation of local features”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2360–2367.

Freund Y., (1995) , “Boosting a weak learning algorithm by majority. Information and Computation”, 121(2), 256–285.

Freund Y., Schapire R. E., (2005), “A desicion-theoretic generalization of on-line learning and an application to boosting BT”, Computational learning theory. Computational Learning Theory, 904, 23–37.

Friedman J. H., (2001), “Greedy function approximation: A gradient boosting machine”, Annals of Statistics, 29(5), 1189–1232.

Friedman J. H., (2002), “Stochastic gradient boosting. Computational Statistics and Data Analysis”, 38(4), 367–378.

Fu H., Gong M., Wang C., Tao D., (2018), “MoE-SPNet: A mixture-of-experts scene parsing network”, Pattern Recognition, 84, 226–236.

Fu Y., Wei Y., Zhou Y., Shi H., Huang, G., Wang X., Huang T., (2019), “Horizontal

pyramid matching for person re-identification", Proceedings of the AAAI Conference on Artificial Intelligence, 33, 8295–8302.

Garipov T., Izmailov P., Podoprikhin D., Vetrov D. P., Wilson A. G., (2018), "Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs", ArXiv Preprint ArXiv:1802.10026.

Ge Y., Li Z., Zhao H., Yin G., Yi S., Wang X., Li H., (2018), "FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-identification", In Advances in neural information processing systems, 1222-1233.

Geman S., Bienenstock E., Doursat R., (1992), "Neural networks and the bias/variance dilemma", Neural Computation, 4(1), 1–58.

Geng M., Wang Y., Xiang T., Tian Y., (2016), "Deep Transfer Learning for Person Re-identification", ArXiv Preprint ArXiv:1611.05244..

Gheissari N., Sebastian T. B., Hartley R., (2006), "Person Reidentification Using Spatiotemporal Appearance", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2, 1528–1535.

Girosi F., Jones M., Poggio T., (1995), "Regularization theory and neural networks architectures", Neural Computation, 7(2), 219–269.

Goodfellow Ian and Bengio Y., Courville A., (2016), "Deep Learning", MIT Press.

Granitto P. M., Verdes P. F., Ceccatto H. A., (2005), "Neural network ensembles: Evaluation of aggregation algorithms", Artificial Intelligence, 163(2), 139-162.

Gray D., Brennan S., Tao H., (2007), "Evaluating appearance models for recognition, reacquisition, and tracking", 10th International Workshop on Performance Evaluation for Tracking and Surveillance, 3, 41–47.

Gray D., Tao H., (2008), "Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features", European conference on computer vision, 262-275

Güler R. A., Neverova N., Kokkinos I., (2018), "DensePose: Dense Human Pose Estimation In The Wild", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7297-7306.

Guo J., Yuan Y., Huang L., Zhang C., Yao J. G., Han K., (2019), "Beyond human parts: Dual part-aligned representations for person re-identification", Proceedings of the IEEE International Conference on Computer Vision, 3641–3650.

Guo Y., Yao A., Zhao H., Chen Y., (2017), "Network sketching: Exploiting binary structure in deep CNNs", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4040–4048.

Guo Y., Zhao X., Ding G., Han J., (2018), "On trivial solution and high correlation problems in deep supervised hashing", Thirty-Second AAAI Conference on Artificial

Intelligence.

Guyon I., Vapnik V., Boser B., Bottou L., Solla S. A., (1992), “Structural risk minimization for character recognition”, Advances in Neural Information Processing Systems, 471–479.

Hamdoun O., Moutarde F., Stanciulescu B., Steux B., (2008), “Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences”, Second ACM/IEEE International Conference on Distributed Smart Cameras, 1–6.

He K., Zhang X., Ren S., Sun J., (2016), “Deep residual learning for image recognition”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

Healey S. P., Cohen W. B., Yang Z., Kenneth Brewer C., Brooks E. B., Gorelick N., Zhu Z., (2018), “Mapping forest change using stacked generalization: An ensemble approach”, Remote Sensing of Environment, 204, 717–728.

Hermans A., Beyer L., Leibe B., (2017), “In Defense of the Triplet Loss for Person Re-Identification”, ArXiv Preprint ArXiv:1703.07737.

Hinton G. E., Osindero S., Teh Y. W., (2006), “A fast learning algorithm for deep belief nets.”, Neural Computation, 18(7), 1527–1554.

Hinton G., Vinyals O., Dean J., (2015), “Distilling the knowledge in a neural network”, ArXiv Preprint ArXiv:1503.02531.

Hirzer M., Beleznai C., Roth P. M., Bischof H., (2011), “Person re-identification by descriptive and discriminative classification”, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 91–102.

Hu J., Zhang J., Zhang C., Wang J., (2016), “A new deep neural network based on a stack of single-hidden-layer feedforward neural networks with randomly fixed hidden neurons”, Neurocomputing, 171, 63–72.

Huang G., Li Y., Pleiss G., Liu Z., Hopcroft J. E., Weinberger K. Q., (2017), “Snapshot ensembles: Train 1, get M for free”, ArXiv Preprint ArXiv:1704.00109.

Huang G., Liu Z., Van Der Maaten L., Weinberger K. Q., (2017), “Densely connected convolutional networks”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4700–4708.

Huang H., Li D., Zhang Z., Chen X., Huang K., (2018), “Adversarially Occluded Samples for Person Re-identification”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 5098–5107.

Huang Y., Xu J., Wu Q., Zheng Z., Zhang Z., Zhang J., (2019), “Multi-pseudo regularized label for generated data in person re-identification”, In IEEE Transactions

on Image Processing, 28(3), 1391-1403

Hubara I., Courbariaux M., Soudry D., El-Yaniv R., Bengio Y., (2016), “Binarized Neural Networks”, Advances in Neural Information Processing Systems, 4107–4115.

Ioffe S., Szegedy C., (2015), “Batch normalization: Accelerating deep network training by reducing internal covariate shift”, ArXiv Preprint ArXiv:1502.03167.

Izmailov P., Podoprikhin D., Garipov T., Vetrov D., Wilson A. G., (2018), “Averaging Weights Leads to Wider Optima and Better Generalization”, ArXiv Preprint ArXiv:1803.05407.

Jacobs R. A., Jordan M. I., (1993), “Learning Piecewise Control Strategies in a Modular Neural Network Architecture”, IEEE Transactions on Systems, Man and Cybernetics, 23(2), 337–345.

Jacobs R. A., Jordan M. I., Nowlan S. J., Hinton G. E., (1991), “Adaptive Mixtures of Local Experts”, Neural Computation, 3(1), 79–87.

Jaderberg M., Simonyan K., Zisserman A., others., (2015), “Spatial transformer networks”, Advances in Neural Information Processing Systems, 2017–2025.

Kim W., Goyal B., Chawla K., Lee J., Kwon K., (2018), “Attention-based Ensemble for Deep Metric Learning”, ArXiv Preprint ArXiv:1804.00382.

Kingma D. P., Ba J. L., (2015), “Adam: A method for stochastic optimization”, arXiv preprint arXiv:1412.6980,

Krizhevsky A., Sutskever I., Hinton G. E., (2012), “Imagenet classification with deep convolutional neural networks”, Advances in Neural Information Processing Systems, 1097–1105.

Kukačka J., Golkov V., Cremers D., (2017), “Regularization for Deep Learning: A Taxonomy”, arXiv preprint arXiv:1710.10686.

Lakshminarayanan B., Pritzel A., Blundell C., (2017), “Simple and scalable predictive uncertainty estimation using deep ensembles”, Advances in Neural Information Processing Systems, 6403–6414.

Lee D. H., (2013), “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural network”, Workshop on Challenges in Representation Learning, ICML, 3, 2.

Lee S., Purushwalkam S., Cogswell M., Crandall D., Batra D., (2015), “Why M Heads are Better than One: Training a Diverse Ensemble of Deep Networks”, arXiv preprint arXiv:1511.06314.

Leng Q., Ye M., Tian Q., (2019), “A Survey of Open-World Person Re-identification”, IEEE Transactions on Circuits and Systems for Video Technology, 1–1.

Li D., Chen X., Zhang Z., Huang K., (2017), “Learning deep context-aware features over body and latent parts for person re-identification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 384–393.

Li H., Lin Z., Shen X., Brandt J., Hua G., (2015), “A convolutional neural network cascade for face detection”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 5325–5334.

Li W., Zhao R., Xiao T., Wang X., (2014), “DeepReID: Deep Filter Pairing Neural Network for Person Re-identification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 152–159.

Li W., Zhu X., Gong S., (2017), “Person re-identification by deep joint learning of multi-loss classification”, ArXiv Preprint ArXiv:1705.04724.

Li W., Zhu X., Gong S., (2018), “Harmonious attention network for person re-identification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1, 2.

Liang X., Gong K., Shen X., Lin L., (2018), “Look into person: Joint body parsing pose estimation network and a new benchmark”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(4), 871–885.

Liao S., Mo Z., Zhu J., Hu Y., Li S. Z., (2014), “Open-set Person Re-identification”, arXiv preprint arXiv:1408.0872.

Lin Y., Wu Y., Yan C., Xu M., Yang Y., (2020), “Unsupervised person re-identification via cross-camera similarity exploration”, IEEE Transactions on Image Processing, 29, 5481–5490.

Lin Y., Xie L., Wu Y., Yan C., Tian Q., (2020), “Unsupervised Person Re-identification via Softened Similarity Learning”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3390-3399.

Liu H., Feng J., Qi M., Jiang J., Yan S., (2017), “End-to-end comparative attention networks for person re-identification”, IEEE Transactions on Image Processing, 26(7), 3492–3506.

Liu J., Zha Z. J., Chen D., Hong R., Wang M., (2019), “Adaptive transfer network for cross-domain person re-identification”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 7195–7204.

Liu Jinxian Ni B., Yan Y., Zhou P., Cheng S., Hu J., (2018), “Pose Transferrable Person Re-identification”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Liu Y., Yao X., (1999), “Ensemble learning via negative correlation”, Neural Networks, 12(10), 1399-1404.

Liu Z., Wu B., Luo W., Yang X., Liu W., Cheng K.-T., (2018), “Bi-Real Net: Enhancing the Performance of 1-bit CNNs With Improved Representational Capability and Advanced Training Algorithm”, Proceedings of the European conference on computer vision, 722-737.

Loy C. C., Xiang T., Gong S., (2010), “Time-delayed correlation analysis for multi-camera activity understanding”, International Journal of Computer Vision, 90(1), 106–129.

Luo H., Gu Y., Liao X., Lai S., Jiang W., (2019), “”, Bag of Tricks and A Strong Baseline for Deep Person Re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.

Luo J.-H., Wu J., Lin W., (2017), “ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression”, Proceedings of the IEEE international conference on computer vision, 5058-5066.

McCue C., (2006), “Data mining and predictive analysis: Intelligence gathering and crime analysis”, Butterworth-Heinemann.

Meir R., (1995), “Bias, variance and the combination of least squares estimators”, Advances in Neural Information Processing Systems, 295–302.

Melville P., Mooney R. J., (2005), “Creating diversity in ensembles using artificial data”, Information Fusion, 6(1), 99-111.

Miao J., Wu Y., Liu P., Ding Y., Yang Y., (2019), “Pose-Guided Feature Alignment for Occluded Person Re-Identification”, Proceedings of the IEEE International Conference on Computer Vision, 542-551.

Miech A., Laptev I., Sivic J., (2018), “Learning a Text-Video Embedding from Incomplete and Heterogeneous Data”, ArXiv Preprint ArXiv:1804.02516.

Nielsen D., (2016), “Tree Boosting With XGBoost: Why does XGBoost win every machine learning competition?”, Master’s Thesis, Norweigian University of Science and Technology.

Odena A., (2016), “Semi-supervised learning with generative adversarial networks”, ArXiv Preprint ArXiv:1606.01583.

Oh Song H., Xiang Y., Jegelka S., Savarese S., (2016), “Deep metric learning via lifted structured feature embedding”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4004–4012.

Opitz D. W., Shavlik J. W., (1996), “Generating Accurate and Diverse Members of a Neural-Network Ensemble”, Advances in Neural Information Processing Systems, 535-541.

Opitz M., Waltner G., Possegger H., Bischof H., (2017), “BIER-Boosting Independent Embeddings Robustly”, Proceedings of the IEEE International

Conference on Computer Vision, 5189–5198.

Palangi H., Ward R., Deng L., (2017), “Convolutional Deep Stacking Networks for distributed compressive sensing”, Signal Processing, 131, 181–189.

Pedagadi S., Orwell J., Velastin S., Boghossian B., (2013), “Local Fisher Discriminant Analysis for Pedestrian Re-identification”, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 3318–3325.

Qian X., Fu Y., Jiang Y. G., Xiang T., Xue X., (2017), “Multi-scale Deep Learning Architectures for Person Re-identification”, Proceedings of the IEEE International Conference on Computer Vision, 5409–5418.

Qian X., Fu Y., Xiang T., Jiang Y.-G., Xue X., (2019), “Leader-based Multi-Scale Attention Deep Architecture for Person Re-identification”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 1–1.

Qian X., Fu Y., Xiang T., Wang W., Qiu J., Wu Y., … Xue X., (2018), “Pose-normalized image generation for person re-identification”, Proceedings of the European Conference on Computer Vision, 650–667.

Ranjan R., Castillo C. D., Chellappa R., (2017), “L2-constrained softmax loss for discriminative face verification”, ArXiv Preprint ArXiv:1703.09507.

Rastegari M., Ordonez V., Redmon J., Farhadi A., (2016), “XNOR-net: Imagenet classification using binary convolutional neural networks”, European conference on computer vision, 525-542.

Renda A., Barsacchi M., Bechini A., Marcelloni F., (2019), “Comparing ensemble strategies for deep learning: An application to facial expression recognition”, Expert Systems with Applications, 136, 1–11.

Sabokrou M., Fayyaz M., Fathy M., Klette R., (2017), “Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes”, IEEE Transactions on Image Processing, 26(4), 1992–2004.

Sarfraz M. S., Schumann A., Eberle A., Stiefelhagen R., (2018), “A Pose-Sensitive Embedding for Person Re-identification with Expanded Cross Neighborhood Re-ranking”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 420–429.

Schapire R. E., (1990), “The Strength of Weak Learnability”, Machine Learning, 5(2), 197–227.

Schroff F., Philbin J., (2015), “FaceNet: A Unified Embedding for Face Recognition and Clustering”, Proceedings of the IEEE conference on computer vision and pattern recognition, 815-823.

Serbetci A., Akgul Y. S., (2020), “End-to-end training of CNN ensembles for person re-identification. Pattern Recognition”, 104, 107319.

Sermanet P., Eigen D., Zhang X., Mathieu M., Fergus R., Lecun Y., (2014), “OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks”, arXiv preprint arXiv:1312.6229..

Shazeer N., Mirhoseini A., Maziarz K., Davis A., Le Q., Hinton G., Dean J., (2019), “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer” , arXiv preprint arXiv:1701.06538.

Shen Y., Li H., Xiao T., Yi S., Chen D., Wang X., (2018), “Deep Group-shuffling Random Walk for Person Re-identification”, Proceedings of the IEEE conference on computer vision and pattern recognition, 2265-2274.

Shi Z., Zhang L., Liu Y., Cao X., Ye Y., Cheng M.-M., Zheng G., (2018), “Crowd counting with deep negative correlation learning”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5382–5390.

Si J., Zhang H., Li C.-G., Kuen J., Kong X., Kot A. C., Wang G., (2018), “Dual attention matching network for context-aware feature sequence based person re-identification”, ArXiv Preprint ArXiv:1803.09937.

Simonyan K., Zisserman A., (2014), “Very deep convolutional networks for large-scale image recognition”, ArXiv Preprint ArXiv:1409.1556.

Singh S., Hoiem D., Forsyth D., (2016), “Swapout: Learning an ensemble of deep architectures”, Advances in Neural Information Processing Systems, 28–36.

Song J., Yang Y., Song Y. Z., Xiang T., Hospedales T. M., (2019), “Generalizable person re-identification by domain-invariant mapping network”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 719–728.

Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., (2014), “Dropout: a simple way to prevent neural networks from overfitting”, The Journal of Machine Learning Research, 15(1), 1929–1958.

Su C., Li J., Zhang S., Xing J., Gao W., Tian Q., (2017), “Pose-driven deep convolutional model for person re-identification”, Proceedings of the IEEE International Conference on Computer Vision, 3960–3969.

Su C., Zhang S., Xing J., Gao W., Tian Q., (2016), “Deep attributes driven multi-camera person re-identification”, European conference on computer vision, 475–491.

Suh Y., Wang J., Tang S., Mei T., Mu Lee K., (2018), “Part-aligned bilinear representations for person re-identification”, Proceedings of the European Conference on Computer Vision, 402–419.

Sun Yi Wang X., Tang X., (2014), “Deep learning face representation from predicting

10,000 classes", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1891–1898.

Sun Y., Xu Q., Li Y., Zhang C., Li Y., Wang S., Sun J., (2019), "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 393–402.

Sun Y., Zheng L., Yang Y., Tian Q., Wang S., (2018), "Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)", Proceedings of the European Conference on Computer Vision, 480-496.

Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Rabinovich A., (2015), "Going deeper with convolutions", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–9.

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z., (2016), "Rethinking the inception architecture for computer vision", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826.

Ting K. M., Witten I. H., (1999), "Issues in stacked generalization", Journal of Artificial Intelligence Research, 10, 271–289.

Tung F., Mori G., (2020), "Deep Neural Network Compression by In-Parallel Pruning-Quantization", IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(3), 568–579.

Ullrich K., Meeds E., Welling M., (2017), "Soft Weight-Sharing for Neural Network Compression", arXiv preprint arXiv:1702.04008.

Vapnik V. N., (1999), "An overview of statistical learning theory", IEEE Transactions on Neural Networks, 10(5), 988–999.

Viola P., Jones M., (2001), "Rapid object detection using a boosted cascade of simple features", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1.

Wang C., Zhang Q., Huang C., Liu W., Wang X., (2018), "Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-identification", Proceedings of the European Conference on Computer Vision, 365-381.

Wang G., Yuan Y., Chen X., Li J., Zhou X., (2018), "Learning discriminative features with multiple granularities for person re-identification", 2018 ACM Multimedia Conference on Multimedia Conference, 274–282.

Wang J., Li Y., Miao Z., (2019), "Ensemble Feature for Person Re-Identification", ArXiv Preprint ArXiv:1901.05798.

Wang J., Song Y., Leung T., Rosenberg C., Wang J., Philbin J., Wu Y., (2014), "Learning fine-grained image similarity with deep ranking", Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition, 1386–1393.

Wang L., Tan T., Ning H., Hu W., (2003), "Silhouette Analysis-Based Gait Recognition for Human Identification", IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(12), 1505–1518.

Wang Yan Wang L., You Y., Zou X., Chen V., Li S., Weinberger K. Q., (2018), "Resource Aware Person Re-identification across Multiple Resolutions", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 8042–8051.

Wang Y., Wang L., Wang, H., Li P., (2019), "End-to-end image super-resolution via deep and shallow convolutional networks", IEEE Access, 7, 31959–31970.

Wang Z., Wang Z., Zheng Y., Chuang Y. Y., Satoh S., (2019), "Learning to reduce dual-level discrepancy for infrared-visible person re-identification", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 618–626.

Wang Z., Ye M., Yang F., Bai X., Satoh S., (2018), "Cascaded SR-GAN for Scale-Adaptive Low Resolution Person Re-identification", IJCAI, 1(2), 4.

Web-1, (2020), http://www.liangzheng.com.cn/Project/project_prw.html,
(Access Date: 15/03/2020)

Web-2, (2020), https://github.com/NEU-Gou/awesome-reid-dataset,
(Access Date: 22/07/2020)

Web-3, (2020), https://github.com/XinDongol/BENN-PyTorch,
(Access Date: 22/07/2020)

Wei L., Zhang S., Gao W., Tian Q., (2018), "Person transfer gan to bridge domain gap for person re-identification", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 79–88.

Wei L., Zhang S., Yao H., Gao W., Tian Q., (2017), "Glad: Global-local-alignment descriptor for pedestrian retrieval", Proceedings of the 25th ACM International Conference on Multimedia, 420–428.

Wojke N., Bewley A., (2018), "Deep cosine metric learning for person re-identification. Proceedings", IEEE Winter Conference on Applications of Computer Vision, 748–756.

Wolpert D. H., (1992), "Stacked generalization", Neural Networks, 5(2), 241–259.

Wu A., Zheng W.-S., Lai J.-H., (2019), "Unsupervised Person Re-Identification by Camera-Aware Similarity Consistency Learning", Proceedings of the IEEE International Conference on Computer Vision, 6922-6931.

Wu J., Rehg J. M., Mullin M. D., (2004), "Learning a rare event detection cascade by direct feature selection", Advances in Neural Information Processing Systems, 1523-1530.

Wu Y., Lin Y., Dong X., Yan Y., Bian W., Yang Y., (2019), “Progressive Learning for Person Re-Identification with One Example”, IEEE Transactions on Image Processing, 28(6), 2872–2881.

Xiao Q., Luo H., Zhang C., (2017), “Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification”, arXiv preprint arXiv:1710.00478.

Xiao T., Li H., Ouyang W., Wang X., (2016a), “Learning deep feature representations with domain guided dropout for person re-identification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1249–1258.

Xiao T., Li H., Ouyang W., Wang X., (2016b), “Learning deep feature representations with domain guided dropout for person re-identification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1249-1258.

Xu H., Mannor S., (2012), “Robustness and generalization”, Mach Learn, 86, 391–423.

Xu Y., Huang R., Ma B., Lin L., (2014), “Person search in a scene by jointly modeling people commonness and person uniqueness”, Proceedings of the ACM Conference on Multimedia, 937–940.

Xuan H., Souvenir R., Pless R., (2018), “Deep Randomized Ensembles for Metric Learning”, Proceedings of the European Conference on Computer Vision, 723–734.

Yang F., Yan K., Lu S., Jia H., Xie X., Gao W., (2019), “Attention driven person re-identification”, Pattern Recognition, 86, 143–155.

Yang Z., Salakhutdinov R., Cohen W. W., (2017), “Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks”, arXiv preprint arXiv:1703.06345..

Yao H., Zhang S., Hong R., Zhang Y., Xu C., Tian Q., (2019), “Deep representation learning with part loss for person re-identification”, IEEE Transactions on Image Processing, 2860-2871.

Ye M., Liang C., Yu Y., Wang Z., Leng Q., Xiao C., Hu R., (2016), “Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing”, IEEE Transactions on Multimedia, 18(12), 2553–2566.

Ye M., Yuen P. C., (2020), “PurifyNet: A Robust Person Re-Identification Model with Noisy Labels”, IEEE Transactions on Information Forensics and Security, 15, 2655–2666.

Yi D., Lei Z., Liao S., Li S. Z., (2014), “Deep metric learning for person re-identification”, International Conference on Pattern Recognition, 34–39.

Yu Q., Chang X., Song Y.-Z., Xiang T., Hospedales T. M., (2017), “The Devil is in the Middle: Exploiting Mid-level Representations for Cross-Domain Instance Matching”, ArXiv Preprint ArXiv:1711.08106.

Yu W., Zhuang F., He Q., Shi Z., (2015), “Learning deep representations via extreme learning machines. Neurocomputing”, 149, 308–315.

Yuan C., Guo J., Feng P., Zhao Z., Xu C., Wang T., (2019), “A jointly learned deep embedding for person re-identification”, Neurocomputing, 330, 127-137.

Yuan Y., Yang K., Zhang C., (2017), “Hard-aware deeply cascaded embedding”, Proceedings of the IEEE international conference on computer vision, 814-823.

Yuksel S. E., Wilson J. N., Gader P. D., (2012), “Twenty years of mixture of experts”, IEEE Transactions on Neural Networks and Learning Systems, 23(8), 1177–1193.

Zhai Y., Guo X., Lu Y., Li H., (2019), “In Defense of the Classification Loss for Person Re-Identification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.

Zhang C., Recht B., Bengio S., Hardt M., Vinyals O., (2019), “Understanding deep learning requires rethinking generalization”, arXiv preprint arXiv:1611.03530.

Zhang H., Shao J., Salakhutdinov R., (2018), “Deep neural networks with multi-branch architectures are less non-convex”, ArXiv Preprint ArXiv:1806.01845.

Zhang Y., Zhong Q., Ma L., Xie D., Pu S., (2019), “Learning Incremental Triplet Margin for Person Re-Identification”, Proceedings of the AAAI Conference on Artificial Intelligence, 9243-9250.

Zhang Z., Lan C., Zeng W., Chen Z., (2019), “”, Densely semantically aligned person re-identification. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 667–676.

Zhao H., Tian M., Sun S., Shao J., Yan J., Yi S., Tang X., (2017), “Spindle net: Person re-identification with human body region guided feature decomposition and fusion”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1077–1085.

Zhao L., Li X., Zhuang Y., Wang J., (2017), “Deeply-learned part-aligned representations for person re-identification”, Proceedings of the IEEE International Conference on Computer Vision, 3219–3228.

Zheng J., Cao X., Zhang B., Zhen X., Su X., (2018), “Deep ensemble machine for video classification”, IEEE Transactions on Neural Networks and Learning Systems, 30(2), 553–565.

Zheng L., Bie Z., Sun Y., Wang J., Su C., Wang S., Tian Q., (2016), “Mars: A video benchmark for large-scale person re-identification”, European Conference on Computer Vision, 868–884.

Zheng L., Huang Y., Lu H., Yang Y., (2019), “Pose-invariant embedding for deep person re-identification”, IEEE Transactions on Image Processing, 28(9), 4500–4509.

Zheng L., Shen L., Tian L., Wang S., Wang J., Tian Q., (2015), “Scalable person re-identification: A benchmark”, Proceedings of the IEEE International Conference on Computer Vision, 1116–1124.

Zheng L., Yang Y., Hauptmann A. G., (2016), “Person re-identification: Past, present and future”, ArXiv Preprint ArXiv:1610.02984.

Zheng W. S., Gong S., Xiang T., (2009), “Associating groups of people”, British Machine Vision Conference.

Zheng Z., Yang X., Yu Z., Zheng L., Yang Y., Kautz J., (2019), “Joint Discriminative and Generative Learning for Person Re-identification”, Proceedings of the IEEE conference on computer vision and pattern recognition, 2138-2147.

Zheng Z., Zheng L., Yang Y., (2017), “Unlabeled Samples Generated by GAN Improve the Person Re-Identification Baseline in Vitro”, Proceedings of the IEEE International Conference on Computer Vision, 3754-3762.

Zheng Z., Zheng L., Yang Y., (2018), “Pedestrian alignment network for large-scale person re-identification”, IEEE Transactions on Circuits and Systems for Video Technology, 29(10), 3037-3045.

Zhong Z., Zheng L., Cao D., Li S., (2017), “Re-ranking person re-identification with k-reciprocal encoding”, Proceedings of the IEEE Computer Vision and Pattern Recognition, 3652–3661.

Zhong Z., Zheng L., Kang G., Li S., Yang Y., (2017), “Random erasing data augmentation”, ArXiv Preprint ArXiv:1708.04896.

Zhong Z., Zheng L., Luo Z., Li S., Yang Y., (2019), “Invariance matters: Exemplar memory for domain adaptive person re-identification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 598–607.

Zhong Z., Zheng L., Zheng Z., Li S., Yang Y., (2018), “Camera Style Adaptation for Person Re-identification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5157–5166.

Zhou Shuchang Wu Y., Ni Z., Zhou X., Wen H., Zou Y., (2016), “DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients”, arXiv preprint arXiv:1606.06160.

Zhou Shuren Ke M., Luo P., (2019), “Multi-camera transfer GAN for person re-identification”, Journal of Visual Communication and Image Representation, 59, 393–400.

Zhu J.Y., Park T., Isola P., Efros A. A., (2017), “Unpaired image-to-image translation using cycle-consistent adversarial networks”, Proceedings of the IEEE International Conference on Computer Vision, 2223–2232.

Zhu S., Dong X., Su H., (2019), “Binary Ensemble Neural Network: More Bits per Network or More Networks per Bit?”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4923-4932.

Zhu X., Wu B., Huang D., Zheng W. S., (2018), “Fast open-world person re-identification”, IEEE Transactions on Image Processing, 27(5), 2286–2300.

Zhuo J., Chen Z., Lai J., Wang G., (2018), “Occluded Person Re-Identification”, Proceedings of the IEEE International Conference on Multimedia and Expo, 1-6.

BIOGRAPHY

Ayşe Şerbetçi was born in Erzincan, in 1985. She graduated from Hacettepe University Department of Computer Engineering in 2009. Currently, she is a PhD candidate at Gebze Technical University Department of Computer Engineering, where she worked as a research assistant between 2015-2019. Her research interests are pattern recognition and image retrieval. She is married and has a 4-years-old son.

APPENDIX

Appendix A: Publications Based on this Thesis

Serbetci A., Akgul Y. S., (2020), "End-to-end training of CNN ensembles for person re-identification", Pattern Recognition, 104, 107319.