ROBUST RE-IDENTIFICATION WITH LARGE VARIATIONS

A Dissertation

by

YE YUAN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Zhangyang Wang
Co-Chair of Committee,	Sing-Hoi Sze
Committee Members,	Theodora Chaspari
	Yang Shen
Head of Department,	Scott Schaefer

August 2020

Major Subject: Computer Science

ABSTRACT

Re-identification (ReID) has been one of the most intensively studied problems in computer vision and finds extensive applications in multi-camera systems such as for public safety, indoor/outdoor monitoring, and smart city/community. Being presented with a subject-of-interest (query) captured in one frame, the ReID algorithms aim to identify the occurrences (matches) of the same subject in other video frames, e.g., at different times of the day, or by other cameras. Typically, a standard ReID system contains three main components: object detection including bounding box proposal and recognition, representation learning, and evaluation for retrieval. Most of the existing ReID approaches aim to learn identity-related features or equivalently, design similarity metrics, and measure identity similarities between image pairs. The main goal for ReID problem is to correctly match two images of the same object under intensive appearance changes caused by either intrinsic factors *i.e.* various pose and viewpoint, or extrinsic factors *e.g.* occlusion, illumination change, and various environmental background.

With the rapidly increasing demand for ReID in multi-camera video surveillance systems, the core technical challenge of the ReID problem is not only just the performance in an enclosed or fixed environment, but also the model's robustness and transferability to diverse and large-scale unseen cases. In seeking a highly robust ReID algorithm for large-scale real-world scenarios, we strive to tackle this challenging problem from four interlinked perspectives: image understanding in poor visibility environments, robust representation learning with noisy labels, domain-invariant learning for better generalizability, and potential mesh recovery for video-based ReID.

To address the robustness of re-identification with large variations, we first conduct a thorough examination of how environmental variances can affect image quality and the visual task, *e.g.*, recognition and detection, and propose a low-level enhancement pipeline as image preprocessing module to help eliminate degradations in complex environmental variations. The proposed image enhancement pipeline wins the second prize in CVPR 2018 UG^2 competition for automatic object recognition in poor visibility environment.

In addition, we comprehensively discuss the main challenge in ReID, *i.e.*, how to correctly match two images of the same subject under intensive appearance changes caused by intrinsic and environmental factors. To be more specific, we introduce an effective yet efficient loss function a fast-approximated triplet (**FAT**) loss for representation to extract informative features from noisy data. The FAT loss provably converts the point-wise triplet loss into its upper bound form, consisting of a point-to-set loss term plus cluster compactness regularization. It preserves the effectiveness of triplet loss, while leading to linear complexity to the training set size. A label distillation strategy is further designed to learn refined soft-labels in place of the potentially noisy labels, from only an identified subset of confident examples, through teacher-student networks. We conduct extensive experiments on three most popular ReID benchmarks, and demonstrate that FAT loss with distilled labels lead to ReID features with remarkable accuracy, efficiency, robustness, and direct transferability to unseen datasets.

Meanwhile, we present an adversarial domain-invariant feature learning framework (**ADIN**) to eliminate extrinsic misleading information. The ADIN framework explicitly learns to separate identity-related features from challenging variations, where for the first time "free" annotations in ReID data such as video timestamp and camera index are utilized. Experiments on existing large-scale person/vehicle ReID datasets demonstrate that ADIN learns more robust and generalizable representations, as evidenced by its outstanding direct transfer performance across datasets, which is a criterion that can better measure the generalizability of large scale Re-ID methods.

Furthermore, we explore the possibility of modeling 3D-mesh and capturing video motion as an alternative representation for ReID to completely get rid of any environmental distraction in appearance.

DEDICATION

To my husband and parents, for their love and support.

ACKNOWLEDGMENTS

Foremost, I would like to express sincere gratitude to my adviser, Dr. Zhangyang Wang for introducing me to the Computer Vision area, sharpening my skills, broadening my perspective, and training me to be a Deep Learning researcher. His invaluable expertise and insight inspired my entire Ph.D. study, without which neither this dissertation nor my career would have been possible.

I am also truly grateful to my co-advisor Dr. Sing-Hoi Sze on guiding me to be a computer scientist. I would like to extend the thanks to my doctoral committee members Dr. Theodora Chaspari and Dr. Yang Shen, for offering valued suggestions in the preparation of this dissertation.

Over the past years, it has been my great honor to collaborate with many outstanding researchers, including my previous internship supervisors: Dr. Yang Yang from Walmart Tech; Dr. Chen Fang, Dr. Xiaohui Shen, and Dr. Jianchao Yang from ByteDance AI Lab; and Dr. Zhaowen Wang, Dr. Hailin Jin from Adobe Research; as well as my project mentors: Dr. Jiaying Liu from Peking University, Dr. Zhou Ren and Dr. Gang Hua from Wormpex AI Research, and many others. The experiences with them were both productive and enjoyable and made up indispensable parts of this dissertation.

My thanks also go to many other friends, classmates, and alumni from University of Science and Technology of China, Rice University and Texas A&M University, in particular to all the former and current members of the Visual Informatics Group at Texas A&M (VITA), whose continuing support and encouragement have made my Ph.D. career an unforgettable journey.

Lastly and most importantly, I want to thank my parents and my husband Wuyang Chen for their love, understanding, and support in all the ways in my life.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Zhangyang Wang, Dr. Sing-Hoi Sze, and Dr. Theodora Chaspari from the Department of Computer Science & Engineering and Dr. Yang Shen from the Department of Electrical and Computer Engineering.

All the work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported in part by a DAPRA grant, and a gift grant from Adobe.

NOMENCLATURE

ReID	Re-Identification
CDRM	Cascaded Degradation Removal Modules
FAT	Fast-Approximated Triplet
P2P	Point-to-Point
P2S	Point-to-Set
ADIN	Adversarial Domain-Invariant Learning Framework
SMPL-X	SMPL eXpressive
VMR	Video-Based Mesh Recovery

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Re-identification (ReID) has attracted tremendous attention owing to its many applications in video surveillance, public safety, and so on. To re-identify, by definition, is to recognize the same subject encountered on other occasions. (Fig. 1.1). Most hand-crafted ReID models [1, 2, 3, 4, 5, 6, 7] or deep learning ReID approaches [8, 9, 10, 11, 12, 13, 14, 15, 16] attempt to learn identity-related features or equivalently, to design similarity metrics [17, 18, 19, 20, 21, 22, 23, 24, 25], in order to measure identity similarities between image pairs. That makes ReID essentially an open-set problem, namely, the learned feature extractor or metric should be able to generalize to unseen queries and ideally match identities captured in any locations, at any time, and by any cameras.

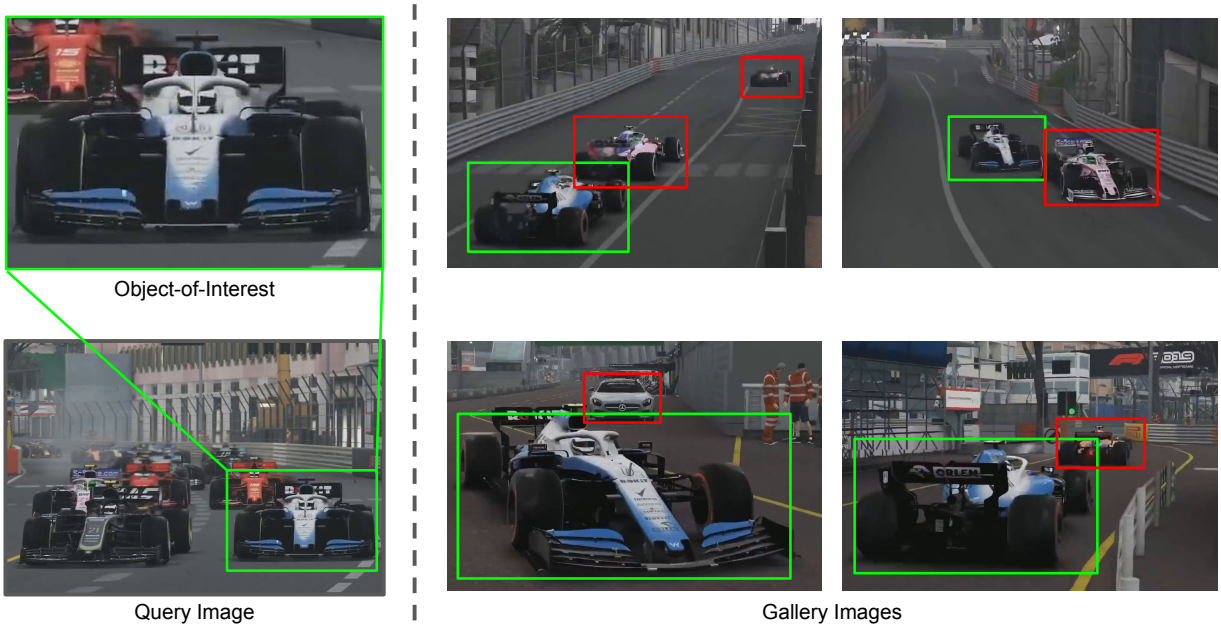

Figure 1.1: Overview of the Re-Identification Problem. Re-Identification is to identify the same subject encountered on other occasions. Left is the query image with a vehicle-of-interest; Right are gallery images of vehicles taken by other cameras. The green bounding boxes indicate that the retrieved vehicle is a correct match while the red bounding boxes indicate false positive retrievals.

With the rapidly increasing demand for ReID in multi-camera systems, the core technical challenge of the ReID problem is no longer just the performance in an enclosed or fixed environment: it has to stay effective to new subjects, scale up to new locations, and be reliable over time. We believe this issue remains yet overlooked in the ReID community, especially when the data volume explodes incredibly and the open scenarios go beyond the controlled conditions in training datasets. Despite a few notable research progress, there remain to be major gaps between the research efforts and the practical needs in large-scale deployment of ReID.

Gap #1: Oversimplified Scenarios. A recent study [26] showed that in 2014, there were 125 video surveillance cameras per thousand people in the U.S.; whereas most ReID datasets were collected only from 20 or fewer cameras (see Table 1.1). Besides, most of the existing datasets cover only single scene in a small region, making them oversimplified in reference to the complexity and diversity of the large-scale scenarios. Those dynamic environments *e.g.*, moving platforms, bad weathers, and various illumination can cause severe visual degradations such as reduced contrasts, detail occlusions, abnormal illumination, fainted surfaces, and color shift. While most ReID models are designed to perform in ideal enclosed environments, *i.e.*, where subjects are well observable without significant attenuation or alteration, practical ReID systems need to reckon with a complex unconstrained environment.

	Benchmark	#cameras	#ID	#boxes per ID	#scenarios
person	Market-1501 [27]	6	1501	21.5	campus supermarket
	DukeMTMC-ReID[28, 29]	8	1812	20.1	University campus
	CUHK03[30]	2	1467	9.0	Campus
	MSMT17[31]	15	4101	30.8	outdoor & indoor
	PKU-VehicleID[32]	2	26267	8.4	a small city
	VeRi-776[33]	20	776	63.6	1.0 km^2 area
	Vehicle-1M[34]	multiple	55527	16.9	several cities

Table 1.1: Publicly Available Benchmarks for ReID. Statistics show the number of cameras, identities, average bounding boxes per identity and scenarios covered in the dataset.

Gap #2: Low Variation Coverage. Existing datasets are constructed from short-time surveillance videos without significant lighting changes and the hand-picked video frames captured in similar outdoor environments and/or under relatively normal lighting conditions. However, practical ReID algorithms need to cope with drastically diverse locations, complex backgrounds, indoor-outdoor matching, intensive day-long illumination variations, and more. Overfitting the unitary environmental nuisances in the limited training data can prohibit ReID algorithms from extracting robust and generalized representations for unseen scenarios in large-scale ReID. Although data augmentation and domain adaptation methods have been explored to improve ReID feature generalization, they either require explicit re-training for every new dataset or redundant training data (with still limited diversity). The huge gap between enclosed datasets and large-scale diverse real cases (domains) will potentially make domain adaptation-related methods suboptimal.

Gap #3: Insufficient Number of Subjects and Spatio-temporal imbalance. Existing ReID datasets are limited in data volume: the number of identities and cameras are not large enough, especially when compared with the real surveillance video data. The limited data restricted in not only subject numbers but also variations coverage. Generally, a ReID dataset consists of images from different subjects, where every subject has an *indefinite but finite* number of images captured by several cameras (Fig.1.2 left) within some certain time periods (Fig.1.2 right). Similarly, different nuisances may also appear in a dataset with certain frequencies, and some subjects may display strong yet superficial correlations with some nuisances, making it highly challenging to decouple them. In real-world ReID, however, a subject may have infinite images that are (very likely) only captured by a small portion of cameras within a small portion of time periods. Intuitively, a person or a vehicle usually appears most in certain regions within certain hours, rather than being a wanderlust anywhere anytime "uniformly" in a city. Hence, the conditional distribution of nuisances given a subject is extremely *non-i.i.d.* The imbalances spatiotemporal coverage have placed jeopardy for practical large-scale ReID. Existing ReID algorithms trained on a single dataset are prone to overfitting nuisances of the training set, and therefore suffering from poor generalizability, as indicated by poor direct transfer performance to unseen datasets.

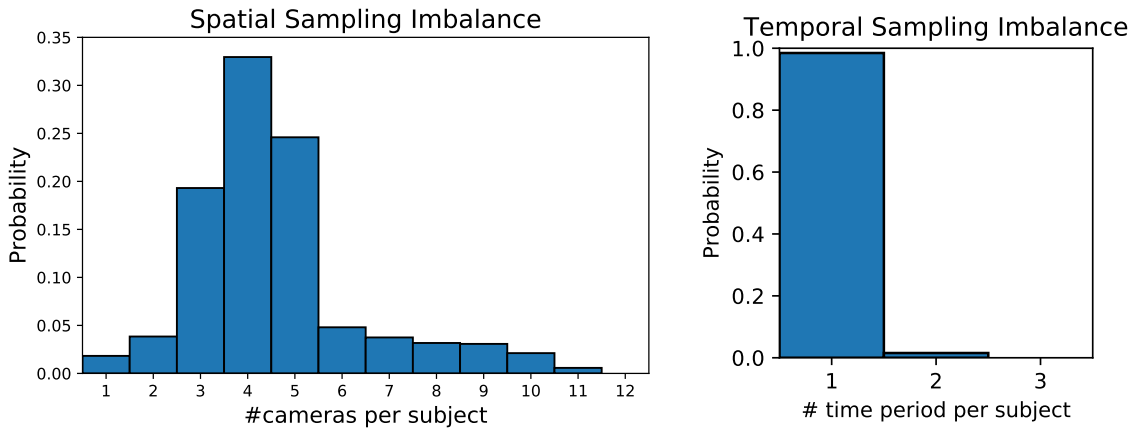

Figure 1.2: Low Spatiotemporal Coverage in ReID Datasets. The low spatiotemporal coverage issue in ReID datasets (MSMT17 [31] as an example). The imbalance manifests in: (left) histogram of camera numbers in which each subject was captured; (right) histogram of time period numbers (1/2/3: morning/noon/afternoon) during which each subject was captured. The two histograms show to be clearly skewed towards the lower end, implying the "localized" patterns of person activities in MSMT17. Similar observations can be found in other peer datasets.

Gap #4: Unpredictable Label Noise and Outliers The growing scale of training datasets embrace the potential of a more powerful model, but introduces sample outliers and label noise during data collection and annotation. [35] observed that a face recognition model trained with only a subset 30% manually cleaned-label samples can achieve comparable performance with models trained on the full dataset. Those noises and outliers provide misleading information and can significantly damage the representation learning. Unfortunately, ReID datasets are notorious to have many noisy labels and outliers, such as label flipping, mislabel, and multi-person coexistence, due to the tedious and error-prone manual annotation process. Meanwhile, sample outliers and label noise in the ReID dataset are common, yet unpredictable and random, that challenges any off-the-shelf data cleaning tools (manual, or automatic). Moreover, the popular margin-based losses for ReID are fragile to label noises. Therefore, learning for large-scale ReID becomes more daunting, due to the not only massive but also noisy datasets.

To bridge the gaps between the research efforts and the practical needs in large-scale deployment of ReID, this dissertation aims to evoke a comprehensive exploration on ReID algorithms

from four interlinked perspectives: image understanding in poor visibility environments, robust representation learning with noisy labels, domain-invariant learning for better generalizability, and potential motion capture for video-based ReID.

Section 2 first presents a literature survey of the up-to-date ReID-related research progress, in order to put this dissertation work in its context. We summarize current ReID benchmarks in Section 2.1 as well as real-world benchmark with large variations for other visual task, *i.e.* detection and recognition in Section 2.2. Then we introduce the state-of-the-art algorithms to learn robust representation with noisy labels in Section 2.3 and generalizable features in large-scale ReID in Section 2.4. Previous mesh recovery methods and options of the simplified parametric deformable human model to represent a human subject that could be assembles into ReID representations are introduced in 2.5.

In Section 3, we present an in-depth analysis on how image restoration could benifit visual recognition task on the UG^2 dataset [36, 37], a large-scale benchmark composed of video imagery captured under challenging conditions in real-world complex scenarios. To further extend this topic to the most common poor visibility scenarios for outdoor ReID, *i.e.* hazy, low-light and rainy conditions, we launch the UG^{2+} challenge [38, 39, 40]. To our best knowledge, it is the first and currently largest effort of benchmark aiming to evoke a comprehensive discussion and exploration about whether and how low-level vision techniques can benefit the high-level automatic visual recognition in various scenarios. Section 3.3 provides the detailed introduction of the haze track in UG^{2+} challenge, summarizes the interesting observations, and discusses the future directions.

In Sections 4 and 5, we address that ReID is essentially an "open-ended" retrieval problem rather than a closed-set classification, e.g., the training and testing sets usually have no overlapped identity classes. To overcome the limitation of training data volume, we introduce the comparative losses *e.g.* triplet loss, which compares the distances between the sample pairs, as a naturally better choices ReID task. Furthermore, we propose the fast approximately triplet (FAT) loss [41] to retain its effectiveness while significantly reduce the computational expensiveness in Section 4. To further improve the robustness of FAT loss as well as triplet loss, we consider a distillation

network to explicitly handle label noise and further boost ReID performance in Section 5.

In Section 6, we demonstrate how to incorporate the freely available video timestamp and camera index, provided in video surveillance as metadata, as auxiliary supervision to eliminate the scene-related nuisances [42]. We exploit the adversarial framework to extract ReID features that can: (1) be utilized to faithfully classify subjects into correct classes; (2) be resilient and invariant to those identified nuisances – in other words: no reliable classifier can be trained on those features to predict those nuisances. To our best knowledge, we are the first to utilize those "free" annotations for image-based ReID, to effectively suppress the overfitting of nuisances.

A fundamental challenge for image-based ReID is that it learns color and generic features of appearance that are shown to be fragile to image degradation and artifacts. An interesting question arises that whether we can disentangle the pose and body shape and learn an approximate "nude shape" of the body for re-identification. We consider the target object to be rigid, while the camera, pose and shape parameters from mesh recovery are treated as parameters that can be disentangled from the body nude shape. Section 7 then presents a preliminary investigation on the video mesh recovery and motion capture via an optimization-based approach, bringing new potential for the future video-based ReID task.

Finally, Section 8 concludes this dissertation with pointers to future directions.

2. LITERATURE REVIEW

2.1 Re-Identification Benchmarks

2.1.1 ReID Datasets

The disconnection between research-level datasets and community/city-level video warehouse remains to hinder the real-life applications of ReID. Table 1.1 summarizes mainstream person ReID and vehicle ReID datasets. For person ReID, considering that even in a grocery store there are usually dozens even more than 100 cameras and over 550 visitors per day, current datasets are more or less overly simplistic. More specifically, the Market-1501 [27], DukeMTMC-ReID [28, 29] and CUHK03 [30] are all collected in small outdoor regions, and in short time periods (usually well-lighted daytime). The latest MSMT17 dataset [31] led positive progress towards real large-scale usage, by including geo-spatially diverse cameras (both indoor and outdoor) and varying time periods (morning, noon and afternoon) and illuminations.

Vehicle ReID witness similar situations, where exiting benchmarks' scale and diversity are still far from being comparable to reality. Previous datasets such as VehicleID [32] have small camera or vehicle numbers, as well as limited viewpoints. A recent VeRi-776 dataset [33] presents a relatively realistic benchmark with cameras spanning a large spatial coverage and other variations, which is one step close to being representative for large-scale vehicle ReID.

In this project, we conduct extensive experiments on popular person re-identification benchmarks Market-1501 [27], DukeMTMC-reID [28, 29], and MSMT17 [31], as well as vehicle re-identification benchmarks PKU-VehicleID[32] and VeRi-776[33].

2.1.2 ReID Evaluation Metrics

The standard ReID pipeline picks a dataset, learning the model from its training set and evaluating the model's retrieval accuracy or mean average precision (mAP) on the held-out testing set (with non-overlapping subjects). However, this single-dataset evaluation is often insufficient in reflecting true generalizability (Fig.2.1) since they overlook a fact, *i.e.*, due to the low coverage

of most datasets, the training and testing sets of the same ReID dataset tend to be highly similar in terms of spatiotemporal nuisances (even overlapping or sharing camera IDs). Therefore, high accuracy on the same testing set may be misleading, as that could be a result of nuisance overfitting.

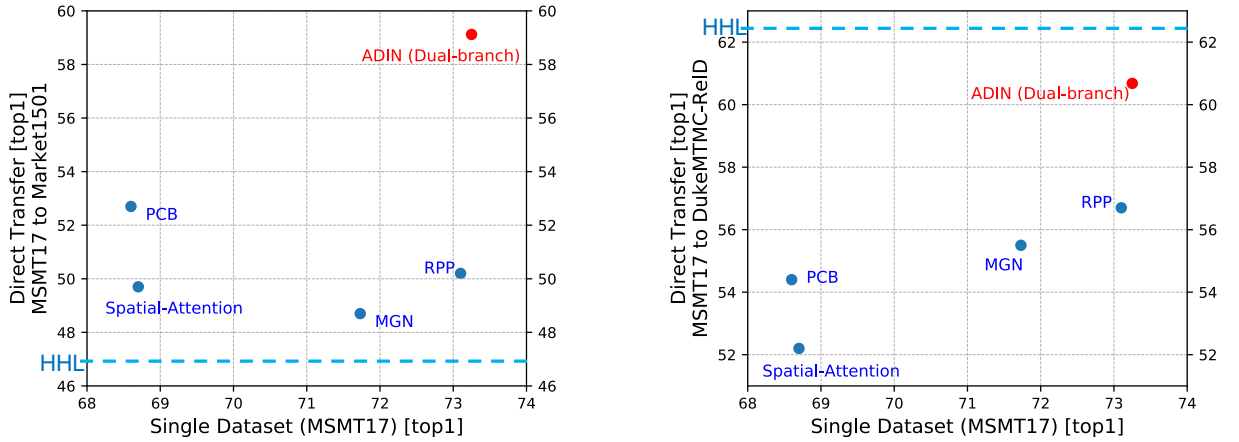

Figure 2.1: ADIN ReID Performance. Top-1 accuracy on a single-dataset (MSMT17 [31]) and direct transfer from MSMT17 to Market1501 [27] (left) and to DukeMTMC-ReID [28, 29] (right). In contrast to our ADIN (red dot in top-right) which achieved competitive performance on both a single dataset and direct transfer, we find other (single-dataset) top-performers suffer from very poor generalizability to unseen domains, indicating the misaligned goal between overfitting small-scale single dataset and generalizing to large-scale unseen scenarios in real life. See Section 6 for details. Methods studied: Spatial-Attention [43], PCB[44], RPP [44], MGN [45], HHL*[46].

Increasing attention has been paid to domain adaption in ReID recently, *i.e.*, training on one source dataset, tuning the trained model on some different target domain data, and finally evaluating model accuracy/mAP on the target dataset. Domain adaptation methods [46] emphasize the generalizability of ReID to new data. Unfortunately, they require target domain data (sometimes even auxiliary attribute annotations in target domain [47]) for re-training purposes. Considering the city growth as well as the explosive increase of cameras, it is unrealistic to collect new data and re-train ReID models for every new domain (*e.g.*, a new camera or a group of cameras in a local

*HHL uses images from both source and target domain for domain adaptation, and thus has no single-dataset performance. We use a horizontal line to represent its domain adaptation performance.

region), making it non-trivial for domain adaptation to scale up. In contrast, we advocate a far more challenging but practically evaluation criterion: direct transfer performance across datasets, to measure ReID model generalizability.

2.1.3 Triplet Loss and Hard Sample Mining

The triplet loss is first introduced in FaceNet [48] by Google to train face embeddings for the recognition task, where softmax cross-entropy loss failed to handle a variable number of classes. The goal of triplet loss is to maximize the inter-class variation while minimizing the intra-class variation. The triple loss is formulated as (2.1) below, where the triplet is defined as an anchor sample a, a positive sample p from the same class and a negative sample n from a different class (y_a, y_p, y_n denote class labels for a, p, n, respectively):

$$L_{\text{tri}} = \sum_{\substack{a,p,n \\ y_p=y_a \\ y_n \neq y_a}} \max\{d(a,p) + m - d(a,n), 0\} \tag{2.1}$$

FaceNet picks a random negative for every pair of anchor and positive, which is computationally much more expensive (cubic or quadratic w.r.t. training set size) than the simple classification loss, which prohibits its wide usage in the ReID application. Later on, [49] improves the efficiency of triplet loss for the ReID task, by proposing two triplet selection strategies: batch all and batch hard. The batch all strategy selects all valid triplets and averaged the loss. The batch hard strategy selects the hardest positive and negative samples within the batch when forming the triplets shown in Fig. 4.1. The author suggests that batch hard strategy with soft margin to yield better performance. A naive triplet loss that compares every possible pair of training samples will incur cubic complexity w.r.t. the training set size [49].

Also, triplet loss relatively quickly learns to correctly map most trivial triplets, rendering a large fraction of all triplets uninformative. Applying triplet loss with randomly selected triplets can accelerate training but quickly stagnates, or becomes difficult to converge. [50] reveals that selecting the hardest triplets often led to bad local minima. They argue that the bias in the triplet selection degraded the performance of learning with triplet loss, and propose a new variant of

triplet loss that adaptively corrects the distribution shift on the selected triplets.

Besides, there are many other successful practices in applying triplet loss to ReID task. [51] proposes a multi-channel convolutional neural network to learn global-local parts features and improves the triplet loss requiring the intra-class feature distances to be less than a predefined threshold. [52] extends the triplet loss to a quadruplet form and required the intra-class variations to be smaller than any inter-class variations. [53] generalizes the point-to-point (P2P) triplet loss to the point-to-set (P2S) form by assuming a positive set (to which the anchor belongs) and a negative set (including all other clusters) for each anchor. It then penalizes the difference between the distance from the anchor to the positive set centroid and the anchor-to-negative-centroid distance. The model is also trained in a soft hard-mining scheme with greater weights to harder samples.

Being related to previous works [49, 53], FAT loss differs substantially in the following ways:

- FAT loss has linear time complexity w.r.t training dataset size: $\mathcal{O}(PK)$ or $\mathcal{O}(PK^2)$ (depending on the choice of negative set), where K denotes the average image number per identity and P the number of identities. Previous triplet losses have either cubic (vanilla) and quadratic (with hard sample mining) time complexity w.r.t training dataset size.

- FAT loss is analytically derived from the upper bound of standard triplet loss. It consists of a P2S loss term and intra-class compactness regularization. Up to our best knowledge, all previous approximations or accelerations for triplet loss, e.g., [51, 53], are only empirical.

- We studied different choices of the negative cluster/centroid, and compared their impacts. Note that FAT loss chooses the negative on "cluster" level, and does not refer to any individual sample mining.

2.1.4 Posed-/Mask-guided ReID

The 2D pose landmarks indicate the body keypoints position and are conductive to ReID problem to track the subject-of-interest [54], align body parts [55, 56, 57], introduce pose variations in training data [58] or eliminate the posture variations in learned representations [59]. In specific,

[54] combines the holistic representations and body pose layout to match and track the subject-of-interest. In order to handle the pose-variations in ReID feature learning, [59] proposes the Feature Distilling Generative Adversarial Network (FD-GAN) is proposed for learning identity-related and pose-unrelated representations. while [58] developes a pose transferred sample augmentations to enrich the pose variations in training data. Pose are widely used for parts alignment in ReID, *e.g.*, [55] explicitly leverages the human pose for local body parts matching and global/local feature fusions. [56] learns a part-aligned representation for person re-identification by aggregating the local similarities of the corresponding pose-aligned body parts. [57] exploits pose landmarks to generate attention maps for the specific body part and the occluded regions, so as to disentangle the useful information from the occlusion noise.

In the meanwhile, the binary body masks are also used in ReID problems in two respects. Firstly, it can be used to detect occlusions and help to remove the background clutters in pixel-level [60, 61]. Besides, the mask contains body shape information which can be regarded as the important gait features [61]. It has been proved that the segmentation mask can greatly improve the robustness of ReID models under various background conditions. Meanwhile, using shape information in the body mask as ReID features is robust to illumination and appearance changes.

2.2 Benchmark for Visual Recognition in Poor Circumstances

2.2.1 UG^2 Benchmark for Visual Quality and Recognition Evaluation

In the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) UG^2 prize challenge [36], the UG^2 dataset was introduced as a large-scale benchmark for to evaluate image restoration and enhancement algorithms for visual recognition. The UG^2 dataset is composed of video imagery captured under three challenging conditions: UAV, Glider, and Ground and consists of over 200,000 annotated frames representing 228 ImageNet [62] classes and super-class (most of the super-classes in the dataset are composed of more than one ImageNet synset).

The UAV collection contains videos with problematic weather/scene conditions, *e.g.* night/low light video, fog, cloudy conditions and occlusion due to snowfall, and includes eight video artifacts,

i.e. glare/lens flare, poor image quality, occlusion, over/under exposure, camera shaking and noise, motion blur, and fish-eye lens distortion. Videos in Glider collection are taken in conditions such as fog, clouds, and occlusion due to rain, and six different video artifacts were observed including glare/lens flare, over/under exposure, camera shaking and noise, occlusion, motion blur, and fish-eye lens distortion. The Ground collection contains videos taken in several weather conditions (sun, clouds, rain, snow). Several common artifacts, *i.e.* motion blur and fish-eye lens distortion, was intentionally induced for a fair comparison.

Image restoration algorithms are treated as an image pre-processing step and evaluated by performance gain of object recognition, *i.e.* the improvement of accuracy on the enhanced cropped bounding box over the un-altered tight bounding box. To measure accuracy, pretrained classification models are used, and each of them predicts a list of the ImageNet synsets. If the predicted synset belongs to the ground truth super-class, the prediction is considered to be correct. The M1 and M2 scores are then calculated as the accuracy evaluation metrics. The M1 measures the rate of achieving at least one correctly synset class in the top-5 predictions and the M2 measures the rate of placing all the correct synset classes in the super-class labels.

2.2.2 Image Restoration and Enhancement

There are numerous algorithms aiming to enhance visibility of the degraded imagery, such as image and video denoising/inpainting [63, 64, 65, 66, 67], deblurring [68, 69, 70, 71, 72], super-resolution [73, 74, 75, 76] and interpolation [77].

Here we introduce the six most powerful restoration algorithms, as evaluated on UG^2 datasets. Histogram Equalization balances the distribution of pixel intensities and increases the global contrast of images. To do this, Contrast Limited Adaptive Histogram Equalization (**CLAHE**) is adopted [78]. The image is partitioned into regions and the histogram of the intensities in each is mapped to a more balanced distribution. As the method is applied at the region level, it is more robust to locally strong over-/under-exposures and can preserve edges better. Given that removing blur effects is widely found to be helpful in fast-moving aerial cameras, and/or in low light filming conditions, **Deblur-GAN** [70] is employed as an enhancement module in which, with adversarial

training, the generator in the network is able to transform a blurred image to a visually sharper one. **Recurrent Residual Net for Super-Resolution** is proposed in [79]. Due to the large distance between objects and aerial cameras, low-resolution is a bottleneck for recognizing most objects from UAV photos. This model is a recurrent residual convolutional neural network consisting of six layers and skip-connections. **Deblocking-Net** [80] is an autoencoder-based neural network with dilation convolutions to remove blocking effects in videos, which was fine-tuned using the VGG-19 perceptual loss function, after training using JPEG-compressed images. Since lossy video coding for on-board sensors introduced blocking effects in many frames, the adoption of the deblocking net was found to suppress visual artifacts. **RED-Net** [80] is trained to restore multiple mixed degradations, including noise and low-resolution together. Images with various noise levels and scale levels are used for training. The network can improve the overall quality of images. **HDR-Net** [81] can further enhance the contrast of images to improve the quality for machine and human analysis. This network learns to produce a set of affine transformations in bilateral space to enhance the image while preserving sharp edges.

2.2.3 Haze Benchmarks and Dehazing Algorithms

Most datasets used for image restoration aims to evaluate with the quantitative (PSNR, SSIM, *etc.*) or qualitative (visual subjective quality) of enhanced images. Those datasets, *e.g.* HazeRD [82], OHAZE [83] and IHAZE [84] for dehazing, usually come with more diverse scene content, provide no integration with subsequent high-level tasks. The popularity of deep learning methods has increased demand for training and testing data. A few works [85, 86, 87] and aerial vehicles bechmarks [88, 89, 90] make preliminary attempts for visual understanding, video summarization, or face recognition in unconstrained and potentially degraded environments. However, few works specifically consider the impacts of image enhancement and high-level visual task jointly. Prior to this UG^{2+} effort, a large-scale hazy image dataset and a comprehensive study: REalistic Single Image DEhazing (RESIDE) [91], is proposed to thoroughly examine visual reconstruction and vision recognition in hazy images. The RESIDE haze benchmark brought new light on the comparisons and limitations of state-of-the-art algorithms, and suggest promising fu-

ture directions.

Besides, numerous dehazing methods have been proposed to study visual behavior on hazy scenarios. Early-stage dehanze algorithm rely on the exploitation of natural image priors and depth statistics, *e.g.* locally constant constraints and decorrelation of the transmission [92], dark channel prior [93], color attenuation prior [94], nonlocal prior [95]. In [96, 97], Retinex theory is utilized to approximate the spectral properties of object surfaces by the ratio of the reflected light. Recently, deep learning and convolutional neural models bring in the new prosperity for dehazing. Several methods [98, 99] rely on various CNNs to learn the transmission fully from data. Beyond estimating the haze related variables separately, successive works make their efforts to estimate them in a unified way. In [100, 101], the authors use a factorial Markov random field that integrates the estimation of transmission and atmosphere light. some researchers focus on the more challenging night-time dehazing problem [102, 103]. In addition to image dehazing, AOD-Net [104, 105] considers the joint interplay effect of dehazing and object detection in an unified framework. The idea is further applied to video dehazing by extending the model into a light-weight video hazing framework [106]. In another recent work [107], the semantic prior is also injected to facilitate video dehazing. In this dissertation, we follow the footsteps of predecessors to advance the fields by proposing new benchmarks.

2.3 Learning from Noisy Labels and Network Distillation

2.3.1 Label Noise in ReID Benchmarks

Deep learning models with fully-supervision assume the correctness of annotations in training datasets, which is not always the truth. Label noise is inevitable in either manually labeled or auto-annotated dataset, especially in the large-scale benchmarks. Fig. 2.2 shows some fail cases due to label noises in MSMT17, the largest person ReID dataset, which provide incorrect/imprecise information of training data and therefore significantly damage the representation learning process. We could observe three typical cases of label noise in ReID dataset:

- **Label flipping**: the query label should be 1748 but 1749 is provided, therefore the correct

retrievals are determined as false positives due to the wrong query label;

- **Mislabel**: the query image contains incomplete person (a blue jacket with orange hood) and that partial human body lacks sufficient information for a successful retrieval;

- **Multi-person coexistence**: the image contains multiple people while the label indicates only one of them, which makes the model fail to learn an accurate representation.

2.3.2 Label Denoising in ReID

Few works in ReID has been done to handle label noise. [108] learns a snippet embedding for video-based person ReID to avoid noise and outliers in pair-wise learning. [109] proposes to annotate unlabeled data with top-k counts label for unsupervised video-based person re-identification (re-ID). [110] exploites a reinforcement Learning model to free up annotation labors and auto-select attention from bounding boxes. [111] and [112] utilizes body segments to learn pose-guided features so as to overcome person body misalignment caused by detectors or pose variations. [61] introduces binary segmentation mask-guided contrastive attention model to reduce the background clutters and learn features from the foreground only.

Figure 2.2: Examples of Label Noises in Re-Identification Dataset. Label flip (first row), mislabel (second row) and coexistence (third row). The leftmost image in each row is the query image and the right ten images are the top-10 retrieval results. Red border indicates a false positive retrieval. Examples show no correct top-10 retrievals due to the label noises of the query images.

2.3.3 Label Denoising in Deep Learning

To overcome the negative effect of noisy labels, [113] proposes a bootstrap technique to modify the labels on-the-fly by augmenting the prediction objective with a notion of consistency. [114] extends [115] and proposes a re-weighting method that can be combined with any surrogate loss function for classification, to handle class-conditional random label flipping. [116] introduces an extra noise layer to absorb the label noise by adapting the network outputs to the noisy label distribution. [117] further augments the correction architecture by adding a softmax layer on top to explicitly connect the correct labels to noisy ones. [118] provids a forward-and-backward loss correction method given a class-condition label flipping probability. [119] proposes a generic conditional random field (CRF) model as a robust loss to be plugged into any existing network for label space smoothness and therefore noise resistance. [47] designs a Siamese network to distinguish clean labels from noisy labels and to simultaneously give clean labels more emphasis.

Previous approaches either employ temporal information in video-based ReID or apply a pose-/mask-guided attention to providing for feature alignment, which needs extra efforts or processing of training data and is not scalable to large-scale datasets. In the dissertation, we integrate the robustness of classification loss and effectiveness of comparative loss and propose a label distillation network to assign soft labels for samples in place of potentially noisy hard labels for ReID representation learning.

2.3.4 Network Distillation

Develops in [120], network distillation aims to transfer the knowledge in an ensemble of models to a single model, using a soft target distribution produced by the former models. [121] uses distillation to train a more efficient and accurate predictor. [122] unifies distillation and privileged information into one generalized distillation framework to learn better representations. [123] further extends data distillation to omni-supervised learning by an ensemble of predictions from multiple transformations of unlabeled data to generate new training annotations using a single network. [124, 125] applies data distillation to multi-modal training, while the testing sets might have

noisy or missing modalities. As a relevant work, [126] argues that noisy labels contain useful "side information" and shall not be discarded. The authors proposes a distillation approach to learn from noisy data guided by a knowledge graph.

Our proposed distillation algorithm to learn from noisy labels differs from previous ones in the following respects:

- We are free from the assumption of the existence of a manually-cleaned set. Instead, we train the teacher network with the entire noisy dataset, but only use most confident samples within a batch to update the parameters. We observed that the model updated based on a subset of confident samples can achieve similar or better performance, compared to the model trained with all noisy-labeled samples.

- We investigate different loss functions for distillation; the teacher network is trained with cross entropy loss with the purpose of providing pseudo soft label associated with a confidence; the student network is trained with FAT loss using the soft pseudo labels generated by the teacher network. Hence instead of mimicking a similar softmax classifier as the teacher network, the student network has the capability to “innovate” on a different task with the help of FAT loss, and eventually outperforms the teacher network.

2.4 Improving ReID Generalizability

To resolve the specific challenge of transferability among different ReID datasets, several data augmentation and unsupervised domain adaptation methods have been proposed to expand the diversity of the limited source domain (training set) and mimic the target domain (an unseen testing set for the trained feature extractor) variations and distributions.

2.4.1 Data Augmentation for Large-Scale ReID

[127, 128, 129] follows a data augmentation approach to incorporate nuisances into training data. Specifically, [127] proposes a random-background data augmentation to generate images of the same identity with a different background. [128] learns a camera-invariant descriptor subspace and transferred the camera styles to each sample as a data augmentation approach. [129] proposes

a novel two-stage pipeline to first learn a set of disentangled foreground, background, and pose factors, followed by re-composing them into novel samples.

However, data augmentation not only adds to the training burden but also introduces a considerable level of noise that leads to training oscillations. Besides, those data augmentation methods would still fail when transferred to an unseen dataset, since no single dataset can cover all possible variations existing in all real-world data. The ReID performance improvement brought by data augmentation also appears to be in general limited.

2.4.2 Domain Adaptation for Transferable ReID

[130] proposes an Unsupervised Multi-task Dictionary Learning (UMDL) model, that learns a joint dictionary to capture view-invariant identity attributes, as well as task-specific modules to capture dataset-unique appearance attributes. [47] introduces a Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) to learn attribute and identity representations under a multi-task framework. However, this method requires extra annotated attribute labels in the source domain. [131] first extracts target domain spatial-temporal patterns using a classifier trained on source domain, and further optimizes a target domain identity classifier with a Bayesian fusion model and a learning-to-rank based mutual promotion procedure. [132] proposes a method to generate a new dataset consisting of images whose identities are from the labeled source domain, while the camera styles are translated from the unlabeled target domain. [46] introduces a Hetero-Homogeneous Learning (HHL) method to learn person embedding with camera variances and domain connectedness, through inter-domain and intra-domain pairwise contrastive learning.

Unsupervised domain adaptation methods promote the transfer performance of ReID models to new datasets. One critical difference between them and our proposed method lies in that domain adaptation needs explicit joint training or fine-tuning when seeing a new target domain, making it ineffective to scale up to many different and unseen domains. In contrast, ADIN framework proposed in the dissertation is designed to be directly transferable to unseen domains without any extra hassle.

2.5 Mesh Recovery and Motion Capture

2.5.1 Deformable Human Mesh Models

To recover a 3D mesh from a 2D image can be fundamentally ambiguous due to lacking precise depth information. The most effective methods to alleviate ambiguity is to integrate a strong prior.

Extensive work have been done to modeling human bodies [133, 134, 135, 136, 137], hands [138, 139, 140, 141, 142, 143, 144, 145, 146] and faces [147, 148, 149, 150, 151, 152, 153, 154, 155]. However, none of these methods, model correlations in face shape and body shape. In [156], the author trains a new, unified, 3D model of the human body, SMPL-X from thousands of 3D scans, that extends SMPL [157] with fully articulated hands and an expressive face.

The SMPL-X model uses standard vertexbased linear blend skinning with learned corrective blend shapes, $N = 10,475$ vertices and $K = 54$ joints. It is defined by a function $M(\theta, \beta, \psi) : \mathbb{R}^{|\theta \times |\beta| \times |\psi||} \rightarrow \mathbb{R}^{3N}$, where the θ encodes the human pose including hands, β encodes body shape (the first three coefficient indicates the global rotation) and ψ encodes facial expressions. The body template is fitted to four datasets of 3D human scans to get 3D alignments. The shape coefficients are trained on 3800 alignments in an A-pose capturing variations across identities, while the body pose coefficients are trained on 1786 alignments in diverse poses. The hands and faces are then learned from 1500 hand and 3800 head high resolution scans respectively.

The total number of model parameters in SMPL-X is 119: 3 parameters for the global body rotation and 72 parameters for joints rotations, 24 parameters for the lower-dimensional hand pose, 10 parameters for subject shape, and 10 parameters for the facial expressions. SMPL-X is realistic, expressive, differentiable, and easy to fit. The model parameters in SMPL-X can be potentially used for ReID purposes.

2.5.2 Human Mesh Recovery Approaches

Since SMPL-X model is recently released, few works has been done to fit it to RGB images or videos. Here, we summarize approaches that methods estimate the SMPL model (focused on body mesh) from a single image [158, 159, 160, 161, 156, 162, 163, 164].

To address this, SMPLify [157] fits the SMPL model to the 2D joint locations by penalizing the error between the projected 3D model joints and detected 2D joints in an optimization framework. HMR [165] trains an end-to-end model to reconstruct 3D mesh from 2D keypoints with the reprojection loss along with an adversarial loss to validate the human body parameter. HMMR [163] learns a representation of 3D dynamics of humans from 2D video pose annotations. They propose to utilize the unlabeled video with pseudo-ground truth 2D pose obtained from an off-the-shelf 2D pose detector and observe a monotonical improvement 3D prediction performance. HMD [162] further utilizes the constraints from body joints, silhouettes, and per-pixel shading information to ensure that the mesh reprojection can better fit to the input image. In addition to the functional tricks mentioned above, our proposed optimization-based mesh recovery algorithm further utilized pseudo-3D pose as supervisions and temporal consistency as regularization to reduce the ambiguity in 3D mesh recovery from the 2D pose.

3. IMAGE ENHANCEMENT FOR DETECTION AND RECOGNITION *

3.1 Motivation

Current ReID video data collected by a ground surveillance camera or UAV systems are commonly suffering from image degradations such as low resolution, motion blur, poor illumination, and noise problems. Those degradations prohibit the detection module of ReID systems from extracting precise bounding boxes for the subject-of-interest. In this section, we studied how low-level image enhancement algorithms can benefit high-level visual tasks and adopted an image restoration module to improve the detection problem (the first step) in ReID.

Fig. 3.1 shows an example of images taken at the hazy condition and the performance of pre-trained pedestrian or vehicle detection/recognition algorithms can be largely jeopardized by various challenging conditions in the unconstrained environments (comparing the ground truth bounding box (left column) and pre-trained Mash R-CNN [166] detection results (right column) on the raw hazy images). Failure to detect pedestrians and vehicles in traffic surveillance can cause a severe problem in automatic pilot and public safety.

One possible solution is to take advantage of image restoration and enhancement algorithms to improve the quality of video frames, and therefore boost the detection performance. As proposed to be functional, the enhancement algorithms can largely improve the visual quality of image captured in poor visibility environments (*i.e.* AOD-Net[105], DCPDN[167] and MSCNN[168] for dehazing, shown in Fig. 3.6. Besides, the image enhancement module can further improve

* Part of the material reported in this section is reprinted with permission from "Bridging the gap between computational photography and visual recognition" by W. Scheirer, R. VidalMata, S. Banerjee, B. RichardWebster, M. Albright, P. Davalos, S. Mc- Closkey, B. Miller, A. Tambo, S. Ghosh, S. Nagesh, Y. Yuan, Y. Hu, J. Wu, W. Yang, X. Zhang, J. Liu, Z. Wang, H. Chen, T. Huang, W. Chin, Y. Li, M. Lababidi, and C. Otto, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI: 10.1109/TPAMI.2020.2996538, 2020. Copyright 2020 by IEEE, and from "Advancing image understanding in poor visibility environments: A collective benchmark study" by W. Yang, Y. Yuan, W. Ren, J. Liu, W. J. Scheirer, Z. Wang, T. Zhang, Q. Zhong, D. Xie, S. Pu, Y. Zheng, Y. Qu, Y. Xie, L. Chen, Z. Li, C. Hong, H. Jiang, S. Yang, Y. Liu, X. Qu, P. Wan, S. Zheng, M. Zhong, T. Su, L. He, Y. Guo, Y. Zhao, Z. Zhu, J. Liang, J. Wang, T. Chen, Y. Quan, Y. Xu, B. Liu, X. Liu, Q. Sun, T. Lin, X. Li, F. Lu, L. Gu, S. Zhou, C. Cao, S. Zhang, C. Chi, C. Zhuang, Z. Lei, S. Z. Li, S. Wang, R. Liu, D. Yi, Z. Zuo, J. Chi, H. Wang, K. Wang, Y. Liu, X. Gao, Z. Chen, C. Guo, Y. Li, H. Zhong, J. Huang, H. Guo, J. Yang, W. Liao, J. Yang, L. Zhou, M. Feng, and L. Qin, *IEEE Transactions on Image Processing*, vol. 29, pp. 5737–5752, 2020. Copyright 2020 by IEEE.

MASK R-CNN detection performance and quantitatively (Fig. 3.7) and qualitatively by correcting coexistence (Fig. 3.8), mislabel issue (Fig. 3.9), and imprecise bounding box position (Fig. 3.10). Therefore, it is highly desirable to study how challenging visual conditions can be coped with for the goal of achieving robust pedestrian detection in the wild.

In the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) UG^2 prize challenge [36], we proposed a novel image enhancement pipeline, Cascaded Degradation Removal Modules (**CDRM**), to improve visual recognition performance on an intentionally difficult, real-world videos dataset collected by unmanned aerial vehicles, manned gliders, and ground cameras. The pipeline assembles light adjustment, super-resolution, deblurring, denoising, high-dynamic ranging, and deblocking modules into one joint deep learning-based cascade. The algorithm achieved preliminary success and won the challenge. Please see Section 3.2 for more details of our observations and proposed approaches for the UG^2 Challenge.

Considering that human visual recognition and machine vision often have considerable misalignment, existing image restoration/enhancement algorithms developed for human perception are not directly applicable to improving ReID detection performance. To better understanding how image enhancement could benefit visual task *e.g.* detection and recognition, we organized the CVPR 2019&2020 UG^{2+} prize challenge [38, 39] and further extend this topic to poor visibility enhancement under hazy, low-light and rainy conditions, the common scenarios for outdoor ReID. A detailed introduction of our datasets, challenges, evaluation protocols, and baseline results as well as the interesting observations and the reflected insights are briefly discussed Section 3.3.

To our best knowledge, this is the first and currently largest effort of benchmark aiming to evoke a comprehensive discussion and exploration about whether and how low-level vision techniques can benefit the high-level automatic visual recognition in various scenarios.

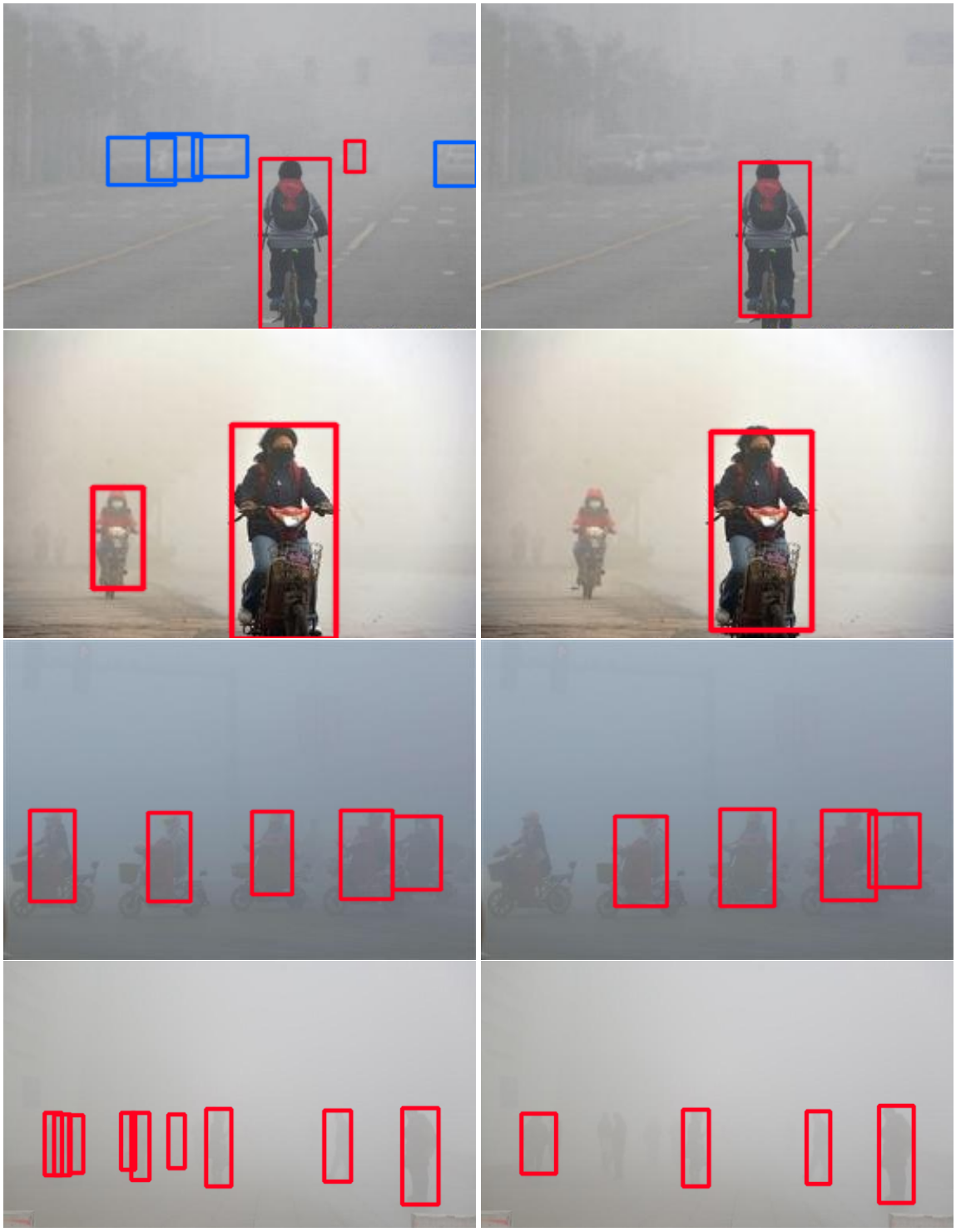

Figure 3.1: Failure Case of Object Detection in Poor Visibility Environments. The left column displays the raw hazy image annotated with ground truth bounding boxes; the right column displays raw image with bounding boxes detected by pre-trained Mask R-CNN.

3.2 Visual Recognition in Challenging Circumstances

To study how image restoration and enhancement could help visual understanding in less than ideal circumstances, we participated in UG^2 Image Recognition Challenge [36] composed of video imagery captured under challenging conditions. Assume the bounding box is provided, we take advantage of the image restoration modules as reprocess step and validate the visual recognition improvement on enhanced images compared to unaltered images. We conducted a thorough examination on the UG^2 benchmark (Fig.3.2), and observed that independently removing any single type of degradation could, in fact, undermine performance in the recognition task since other degradations were not simultaneously considered and those artifacts might be amplified during this process. To measure accuracy, pretrained classification models are used, and each of them predicts a list of the ImageNet synsets. If the predicted synset belongs to the ground truth super-class, the prediction is considered to be correct. The M1 and M2 scores are then calculated as the accuracy evaluation metrics. The M1 measures the rate of achieving at least one correctly synset class in the top-5 predictions and the M2 measures the rate of placing all the correct synset classes in the super-class labels.

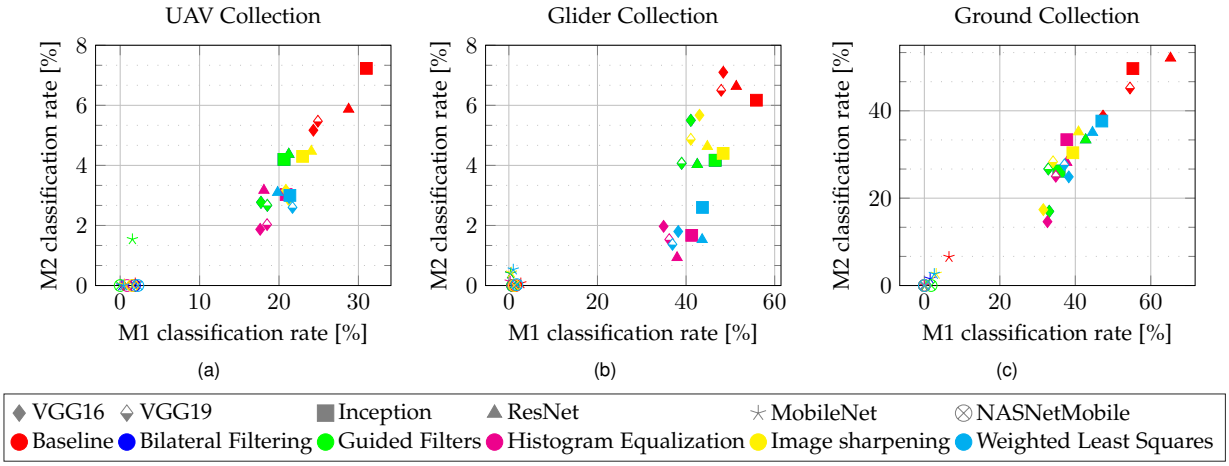

Figure 3.2: Evaluation of Single Restoration Module on UG^2 Dataset. We could observe that (1) None of the image restoration algorithms could outperform the raw degradated images without any enhancement; (2) Image enhancement can have a different impact on different collections.

Besides, images captured from different real-world scenarios (*e.g.* controlled videos taken on the ground, uncontrolled videos were taken by UAVs and manned gliders) may have different degradation characteristics (motion blur are more common in UAV collected dataset while illumination change exists in ground surveillance systems that captures long-time videos). Different degradation type usually needs to be handled by a specific enhancement module, *e.g.* guided filters and image sharpening works best with UAV and glider collections while the Weighted Least Squares benefits Ground collections (Fig.3.2).

Consequently, we proposed an image preprocessing pipeline, the Cascaded Degradation Removal Modules (**CDRM**), that consists of sequentially cascaded degradation removal modules to improve both visual quality and recognition performance. The preprocessing pipeline functions by first identifying the incoming images as belonging to one of the three degradation collections as a form of quality estimation, and then deploying a specific processing model for each collection (Fig. 3.3).

We adopted six state-of-the-art image enhancement modules in CDRM pipeline, including light adjustment (Histogram Equalization [78]), deblurring (Deblur-GAN [70]), super-resolution (Recurrent Residual Net for Super-Resolution [79]), deblocking (Deblocking-Net [80]), denoising (RED-Net [80]), high-dynamic ranging (HDR-Net [81]), to alleviate the detection/recognition performance drop caused by the most common image degradations (*e.g.* glare/lens flare, compression artifacts, occlusion, over/under exposure, camera shaking, sensor noise, motion blur, and fish-eye lens distortion).

We submitted our CDRM to UG^2 Challenge to evaluate the recognition on the hold-out testing set, along with the other three algorithms submitted by participants (CCRE, MA-CNN, and TM-DIP). As can be observed in Fig. 3.4, most of the submitted algorithms were able to improve the recognition performance on the Ground collection but failed in improving that for the aerial collections. Most noteworthy is our CDRM achieves the highest classification improvement with an improvement of 5.30% and 5.21% over the baselines for the Inception M1 and M2 metrics.

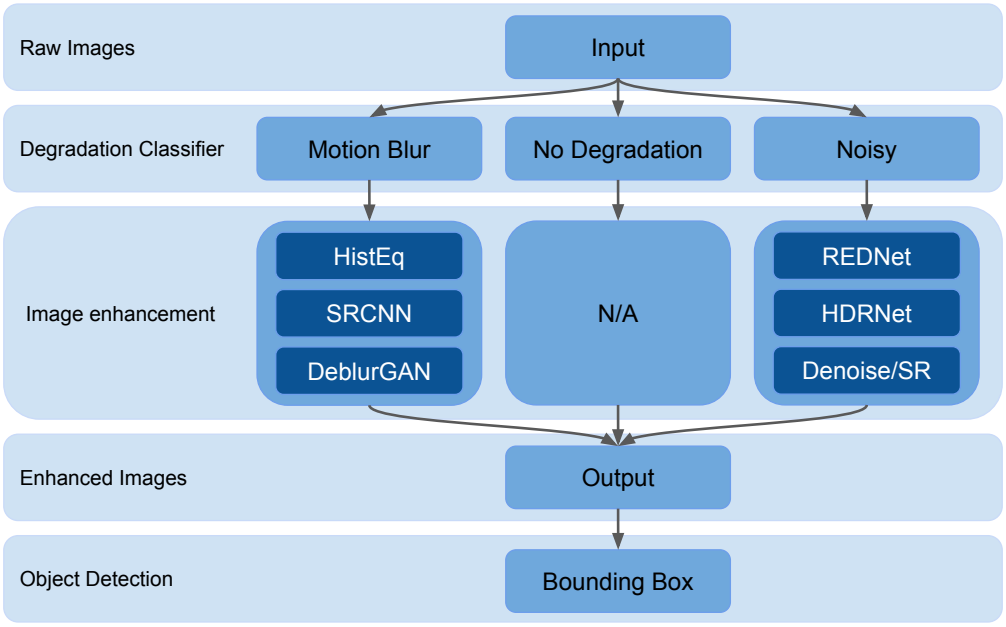

Figure 3.3: Overview of the CDRM Image Enhancement Pipleline. If no degradation is detected, no action is taken (task performance is deemed to be good enough by default). This pipeline process an image in 14 seconds.

3.3 Object Detection in Poor Visibility Environments

Existing enhancement methods are empirically expected to help the high-level-end computer vision task *i.e.* pedestrian detection: however, that is observed to not always be the case in practice. To provide a more thorough examination and fair comparison for detection in poor visibility enhancements caused by haze, we introduced the UG^{2+} [38, 39] haze benchmark.

3.3.1 Collection and Annotation

Collected in real-world hazy conditions, the UG^{2+} haze benchmark consists of 4,322 annotated real-world hazy images from the RESIDE RTTS set [] as the training and/or validation sets, 4,807 unannotated real-world hazy images collected from the same traffic camera sources, for the possible usage of semi-supervised training, and 2,987 real-world hazy images collected from the same sources as the test set. Five categories of objects (car, bus, bicycle, motorcycle, pedestrian) are la-

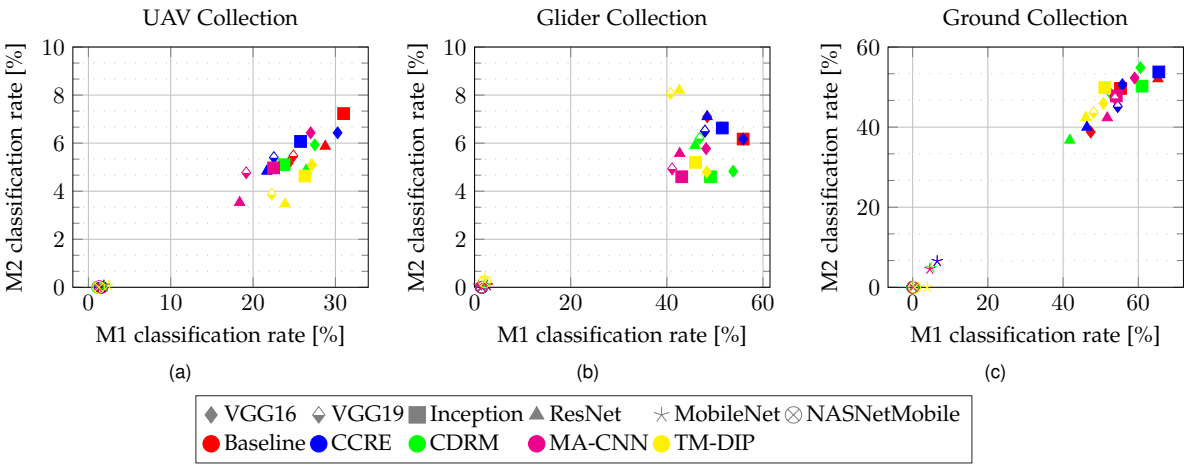

Figure 3.4: Evaluation of CDRM on UG^2 Dataset. Comparison of classification rates at rank 5 for each collection after applying the submitted algorithms from UG2 Challenge participants.

beled with tight bounding boxes. Table 3.1, 3.2 and Fig. 3.5 presents a summary of the benchmark statistics.

	#Images	#Bounding Boxes
Training/Validation	4,310	41,113
Test (held-out)	2,987	24,201

Table 3.1: Image and Object Statistics of UG^{2+} Haze Benchmarks.

3.3.2 Evaluation Metrics

The evaluation criteria are set to be the Mean average precision (mAP) on each held-out test set, with a default Interception-of-Union (IoU) threshold as 0.5. If the ratio of the intersection of a detected region with an annotated object is greater than 0.5, a score of 1 is assigned to the detected region, and 0 otherwise. When mAPs with IoU as 0.5 are equal, the mAPs with higher IoUs (0.6, 0.7, 0.8) will be compared sequentially.

Categories	Car	Person	Bus	Bicycle	Motorcycle
Training/Validation	25,317	11,366	2,590	698	1,232
Test (held-out)	18,074	1,562	536	225	3,804

Table 3.2: Label Statistics of UG^{2+} Haze Benchmarks.

3.3.3 Baseline Composition

We report baseline results using cascading off-the-shelf enhancement methods and popular pre-trained detectors without joint training performed Table 3.3. We test four state-of-the-art object detectors: **Mask R-CNN** [166], **RetinaNet** [169], **YOLO-V3** [170], and Feature Pyramid Network **(FPN)** [171]. We also try three state-of-the-art dehazing approaches: **AOD-Net** [104], Multi-Scale Convolutional Neural Network (**MSCNN**) [99], and Densely Connected Pyramid Dehazing Network (**DCPDN**) [167]. All dehazing models adopt officially released versions.

3.3.4 Results and Analysis

Fig. 3.1 shows the object detection performance on the original hazy images of RESIDE RTTS set using Mask RCNN. The detectors are pretrained on Microsoft COCO, a large-scale object detection, segmentation, and captioning dataset.

The overall detection performance (Table 3.3) has an mAP of only 41.83% using Mask RCNN and 42.54% using YOLO-V3. More detailed detection performance on the five objects can be found in Table 3.4. Results show that without preprocessing or dehazing, the object detectors pretrained on clean images fail to predict a large number of objects in the hazy image. Among all the five object categories, the person has the highest detection AP, while the bus has the lowest AP.

https://github.com/matterport/Mask_RCNN, pretrained on Microsoft COCO dataset.
https://github.com/fizyr/keras-retinanet, , pretrained on Microsoft COCO dataset.
https://github.com/ayooshkathuria/pytorch-yolo-v3, , pretrained on Microsoft COCO dataset.
https://github.com/DetectionTeamUCAS/FPN_Tensorflow, FPN using ResNet-101 backbone is pretrained on the PASCAL Visual Object Classes (VOC) dataset.
https://github.com/Boyiliee/AOD-Net
https://github.com/rwenqi/Multi-scale-CNN-Dehazing
https://github.com/hezhangsprinter/DCPDN

Besides, the choice of pre-trained detectors also matters here: Mask R-CNN outperforms the other two detectors on both validation and test sets, before and after dehazing.

We also compare the validation and test set performance in Table 3.3 and Table 3.4. One possible reason for the performance gap between validation and test sets is that the bounding box size of the latter is smaller compared to the former, as shown in Fig. 3.5.

	mAP	hazy	AOD-Net [104]	DCPDN [99]	MSCNN [167]
validation	RetinaNet	36.18	33.87	36.37	**37.27**
	Mask R-CNN	41.83	39.55	**42.56**	42.28
	YOLO-V3	42.54	41.64	42.06	**43.52**
	FPN	32.25	31.82	31.17	**34.02**
test	RetinaNet	12.79	12.69	12.87	**14.18**
	Mask R-CNN	16.92	17.02	17.42	**18.09**
	YOLO-V3	14.69	14.83	15.08	**15.78**
	FPN	10.69	10.77	9.88	**11.61**

Table 3.3: Overall Detection Performance (mAP) on UG^{2+} Haze Benchmarks.

3.3.5 Effect of Dehazing

We further evaluate the current state-of-the-art dehazing approaches on the hazy dataset, with pre-trained detectors subsequently applied without tuning or adaptation. Fig. 3.6, 3.8, 3.9, 3.10 shows examples that dehazing algorithms can improve not only the visual quality of the images but also the detection accuracies. More detection results are included in Table. 3.4. Detection mAPs of dehazed images using DCPDN and MSCNN approaches are 1% higher on average compared to those of hazy images.

3.3.6 Conclusions

The results of our baseline results lead to some surprises. Even though the dehaze algorithms tends to improve the visual quality for the hazy imagery, no approach can able to uniformly im-

prove the detection performance on all classes over the baseline detection on the raw hazy image. Moreover, in some cases, the dehaze process can even introduce artifacts (*e.g.* color distortion) that might be detrimental to the ReID task. Based on this observation, the image enhancement procedure is only utilized for the detection step in the ReID task, and the bounding box extracted from the original frames is used for the representation learning and retrieval process.

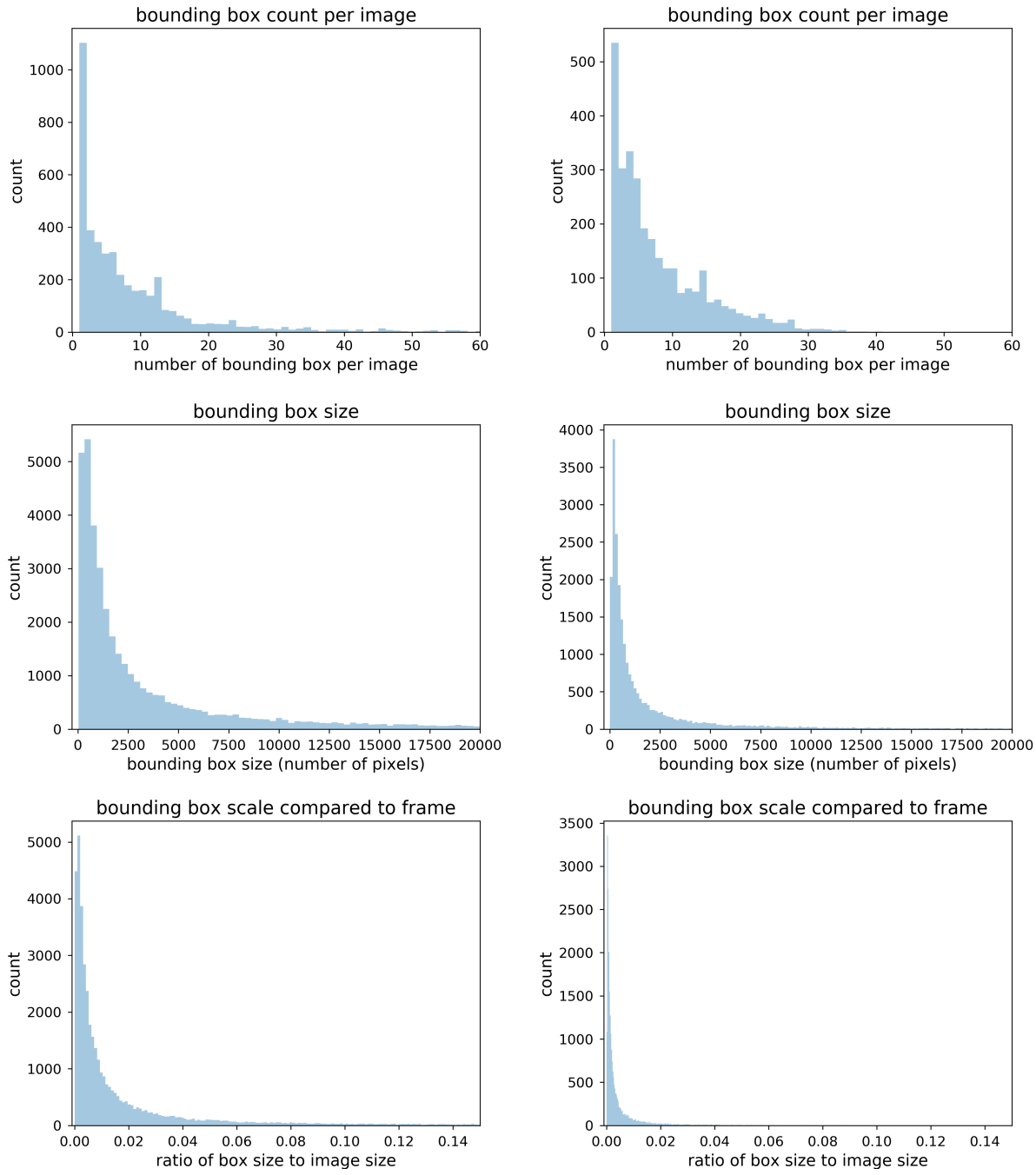

Figure 3.5: Statistical Distributions of UG^{2+} Haze Benchmarks. Training/validation set (the top row) and the held out test set (the bottom row). The first column shows the image size distribution (number of pixels per image), The second column the bounding box count distribution (number of bounding boxes per image), the third column the bounding box size distribution (number of pixels per bounding box), and the last column the ratios of bounding box size compared to frame size.

mAP			hazy	AOD-Net [104]	DCPDN [99]	MSCNN [167]
validation	RetinaNet	Person	55.85	54.93	56.70	**58.07**
		Car	41.19	37.61	42.68	**42.77**
		Bicycle	**39.61**	37.80	36.96	38.16
		Motorcycle	27.37	23.31	**29.18**	29.01
		Bus	16.88	15.70	16.34	**18.34**
	Mask R-CNN	Person	67.52	66.71	67.18	**69.23**
		Car	48.93	47.76	**52.37**	51.93
		Bicycle	**40.81**	39.66	40.40	40.42
		Motorcycle	33.78	26.71	**34.58**	31.38
		Bus	18.11	16.91	18.25	**18.42**
	YOLO-V3	Person	60.81	60.21	60.42	**61.56**
		Car	47.84	47.32	48.17	**49.75**
		Bicycle	41.03	**42.22**	40.18	42.01
		Motorcycle	39.29	37.55	38.17	**41.11**
		Bus	23.71	20.91	**23.35**	23.15
	FPN	Person	51.85	52.35	51.04	**54.50**
		Car	37.48	36.05	37.19	**38.88**
		Bicycle	35.31	35.93	32.57	**37.01**
		Motorcycle	23.65	21.07	22.97	**23.86**
		Bus	12.95	13.68	12.07	**15.83**
test	RetinaNet	Person	17.64	18.23	16.65	**19.34**
		Car	31.41	29.30	33.31	**32.97**
		Bicycle	0.42	**0.84**	0.38	0.75
		Motorcycle	1.69	1.37	1.93	**2.03**
		Bus	12.77	13.70	12.07	**15.82**
	Mask R-CNN	Person	25.60	26.63	24.59	**27.94**
		Car	39.31	39.71	**42.76**	42.57
		Bicycle	**0.64**	0.52	0.22	0.37
		Motorcycle	**3.37**	2.81	2.83	2.99
		Bus	15.66	15.41	**16.69**	16.55
	YOLO-V3	Person	20.64	21.41	21.42	**22.11**
		Car	34.68	33.90	34.52	**35.93**
		Bicycle	0.50	0.38	**0.98**	0.57
		Motorcycle	4.26	4.10	4.72	**5.27**
		Bus	13.55	14.35	13.75	**15.04**
	FPN	Person	12.65	12.57	11.13	**14.19**
		Car	30.54	31.24	27.81	**32.68**
		Bicycle	**1.91**	0.39	1.12	0.97
		Motorcycle	**2.25**	1.7	1.96	1.89
		Bus	6.08	7.93	7.39	**8.31**

Table 3.4: Detailed Detection Performance (mAP) on UG^{2+} Haze Benchmarks.

Figure 3.6: Image Enhancement Improves Visual Quality. The left column displays the raw hazy image; the right column displays images enhanced using pre-trained MSCNN.

Figure 3.7: Image Enhancement Benefits Detection Quantitatively. The top row displays the raw hazy image annotated with ground truth bounding boxes; the second row displays raw image with bounding boxes detected by pre-trained Mask R-CNN; the third and forth row displays image enhanced by pre-trained MSCNN and DCPDN with bounding boxes detected by pre-trained Mask R-CNN.

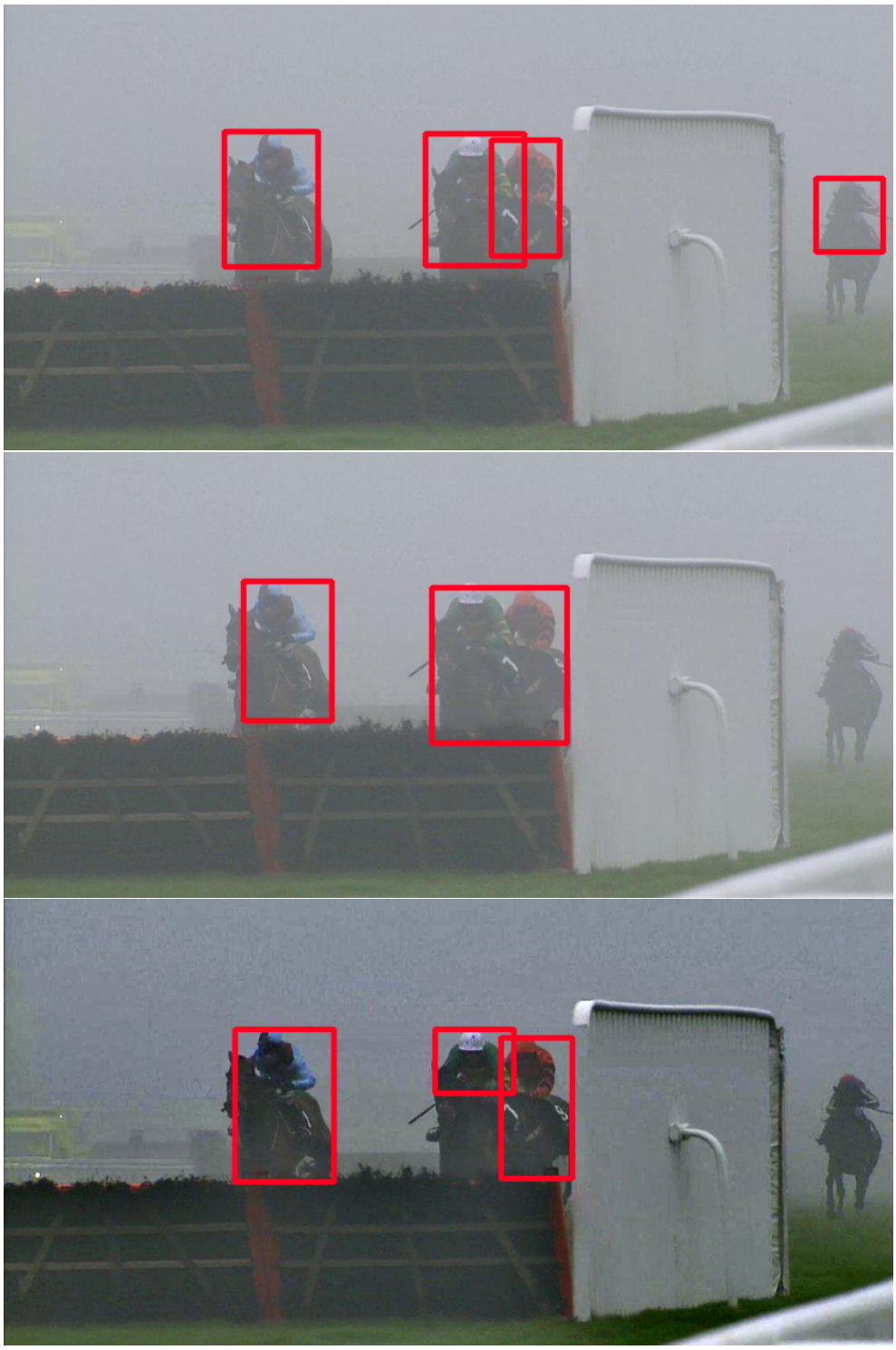

Figure 3.8: Image Enhancement Corrects Coexistence in Detection. The top row displays the raw hazy image annotated with ground truth bounding boxes; the middle row displays raw image with bounding boxes detected by pre-trained Mask R-CNN; the bottom row displays image enhanced by pre-trained MSCNN with bounding boxes detected by pre-trained Mask R-CNN.

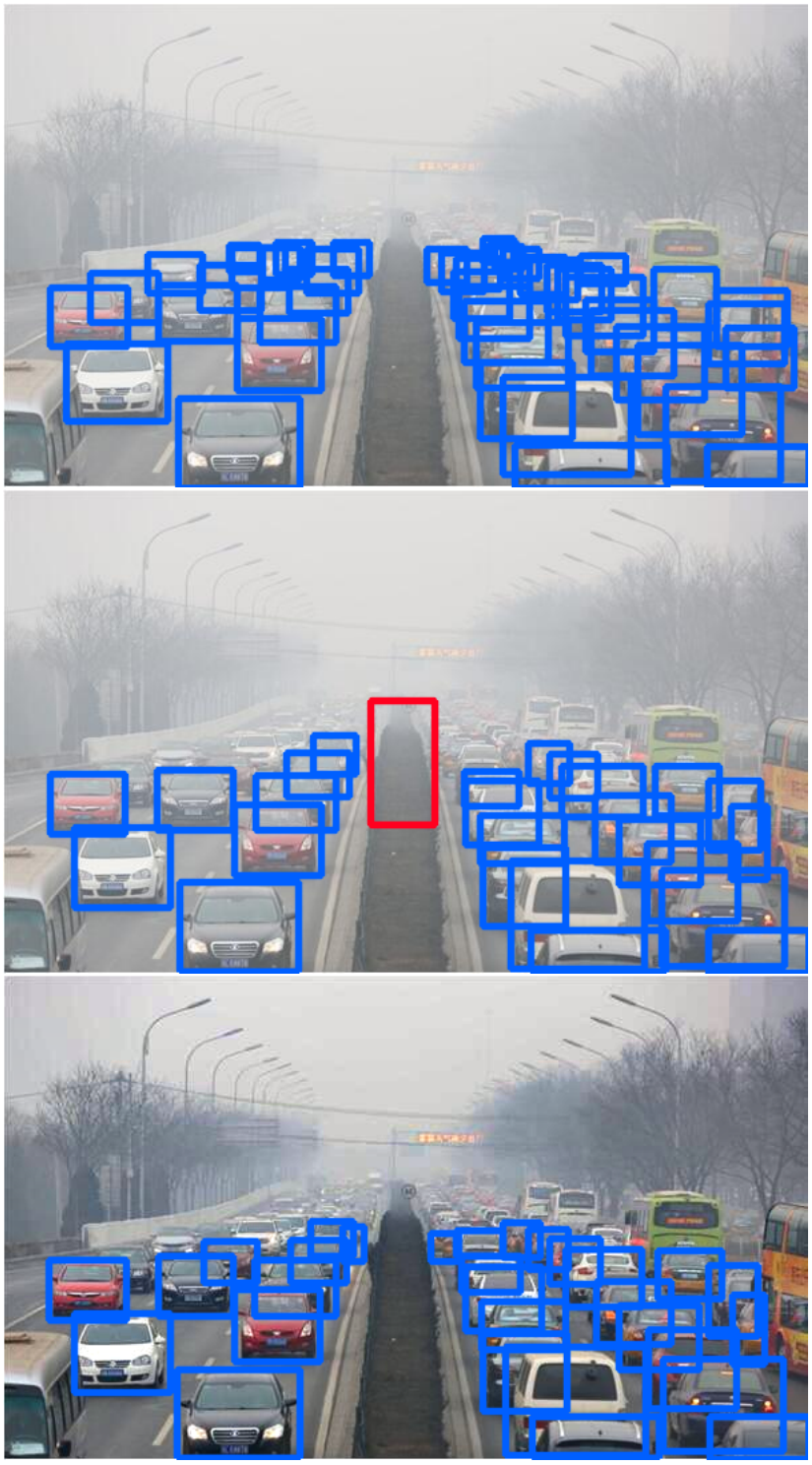

Figure 3.9: Image Enhancement Corrects Mislabel in Detection. The top row displays the raw hazy image annotated with ground truth bounding boxes; the middle row displays raw image with bounding boxes detected by pre-trained Mask R-CNN; the bottom row displays image enhanced by pre-trained MSCNN with bounding boxes detected by pre-trained Mask R-CNN.

Figure 3.10: Image Enhancement Corrects Bounding Box Position in Detection. The top row displays the raw hazy image annotated with ground truth bounding boxes; the middle row displays raw image with bounding boxes detected by pre-trained Mask R-CNN; the bottom row displays image enhanced by pre-trained MSCNN with bounding boxes detected by pre-trained Mask R-CNN.

4. A FAST-APPROXIMATED TRIPLET LOSS FOR REPRESENTATION LEARNING *

4.1 Motivation

Existing deep learning ReID algorithms usually use a classification loss to train their feature learning backbones [172, 173, 174, 175, 16, 42]. However, ReID is essentially an "open-ended" retrieval problem rather than closed-set classification, *e.g.*, the training and testing sets usually have no overlapped identity classes. The learned feature extractor should be able to generalize to matching unseen identities. The testing performance is evaluated by the precision and recall of the matching instances, rather than classification accuracy. Therefore, classification-driven learning could be misaligned with the end goal. Instead, the comparative losses [48, 176, 177, 178], which compares the distances between two sample pairs, are naturally better choices, as empirically validated by a handful of works [179, 175, 52, 180, 181]. Among many, the triplet loss [49] (illustrated in Fig. 4.2a), which maximizes the margin between the intra-class distance and the inter-class distance, has been mostly used in ReID, in order to explicitly embed the relative orders between right and wrong matches (*i.e.*, the correct matches should always be closer to the query than the wrong ones).

An important downside of triplet loss lies in its computational expensiveness, which prohibits its wide usage in the large-scale ReID applications. The vanilla triplet loss needs to calculate over all $PK(K-1)(PK-K)$ possible triplets, where K denotes the average number of images per identity and P identities in total [49]. Hard sample mining [50, 182] has recently become the standard practice in using triplet loss, to select only "informative" (a.k.a. hard) pairs rather than all pairs to enforce the loss. Given the query image, the correct retrievals are positive samples that come from the same class as the query image, while the false positive is negative samples from other classes. Typically, the positive hard sample is an image from the same class but is least

* Part of the material reported in this section is reprinted with permission from "In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation" Y. Yuan, W. Chen, Y. Yang, and Z. Wang, *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, June 2020. Copyright 2020 by IEEE.

similar to the query image (Fig. 4.1 top left). The negative hard sample is an image that is most similar to the query image but belongs to another class (Fig. 4.1 bottom right). Although the hard sample mining can help to reduce the time complexity to $PK(PK-1)+PK$, it runs the risk of causing sample bias [50], often appears fragile to outliers, and suffers from "sample imbalance" issue that the negative pairs (images from a different class) are quantitatively much more than positive ones (images from the same class) during pairwise training [183].

Figure 4.1: Illustration of Hard Samples Mining in Triplet Loss. The positive samples are the images from the query vehicle, but of different viewpoints. The negative hard samples are image that come from different vehicles but are in a similar viewpoints

In this paper, we propose a new fast-approximated triplet (**FAT**) loss to trim down the computational cost of triplet loss without hampering its effectiveness. Viewing all images belonging to the same identity class as a cluster, the proposed FAT loss re-defines a triplet to include an anchor, its corresponding cluster centroid, and the centroid of another cluster. The main idea of FAT loss is to replace point-to-point distances with point-to-cluster distances, through an upper bound re-

laxation of the triplet form. Such a relaxation simultaneously requires the query to be closest to its ground-truth-cluster centroid and enforces each cluster to have a compact radius.

Being related to previous triplet loss [49], FAT loss differs substantially in the following ways:

- FAT loss has **linear time complexity** w.r.t training dataset size: $\mathcal{O}(PK)$ or $\mathcal{O}(PK^2)$ (depending on the choice of the negative set), where K denotes the average image number per identity and P the number of identities. Previous triplet losses have either cubic (vanilla) and quadratic (with hard sample mining) time complexity w.r.t training dataset size.

- FAT loss is analytically derived from the **upper bound** of standard triplet loss. It consists of a point-to-set loss term and intra-class compactness regularization. Up to our best knowledge, all previous approximations or accelerations for triplet loss, e.g., [51, 53], are only empirical.

- We studied different choices of the negative cluster/centroid and compared their impacts. Note that FAT loss chooses the negative on the "cluster" level, and does not refer to any individual sample mining. Therefore the model is least affected by the sample imbalance issue.

Given an anchor image a with the identity label y_a, the triplet loss attempts to find a positive sample p with the same identity label $y_p = y_a$ and a negative sample n with a different label $y_n \neq y_a$, and then maximizes the difference of distances between the positive pair $d(a, p)$ and the negative pair $d(a, n)$ by a margin m. We typically use the euclidean distance (or cosine similarity) between learned ReID features $f_E(a), f_E(p), f_E(n)$ as distance metrics. However, computing triplet loss exhaustively over all possible pairs is too expensive to be practical.

4.2 Fast-Approximated Triplet Loss

To retain the effectiveness of triplet loss while improve its efficiency, we propose a relaxation of the triplet loss 2.1 into its upper bound form. We first have the following two triangle inequalities:

$$\begin{aligned} \max\{0, d(a, c_a) - d(c_a, p)\} \leq d(a, p) \leq d(a, c_a) + d(c_a, p) \\ \max\{0, d(a, c_n) - d(c_n, n)\} \leq d(a, n) \leq d(a, c_n) + d(c_n, n) \end{aligned} \tag{4.1}$$

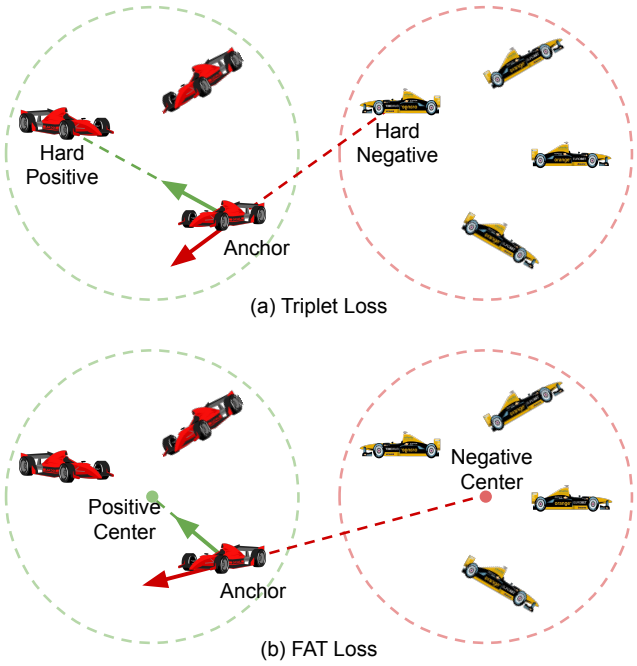

Figure 4.2: Comparison of the Standard Triplet Loss and the FAT Loss. The former compares point-to-point distances, while the latter compares point-to-set distances while regularizing all cluster sets to be compact. The solid arrows depict the "push and pull" effect of triplet loss and the point-to-set term of FAT loss. The dash arrows represents the compactness regularization of FAT loss.

where c_a, c_n are defined as the centroids (average) of the clusters that a, n belong to, respectively. Their proofs are self-evident, given that $d()$ is a well-defined distance function in some metric space. Notice that although we use Euclidean distance for $d()$ by default, our derivations are applicable to other distances too.

We next expand our derivation as in the algorithm (1). Interestingly, the upper bound consists of two terms: a point-to-set (P2S) term which depends on the anchor point; plus a penalty term on the cluster compactness, defined as the largest cluster "radius" among all clusters, whose value is decided by the entire dataset and is agnostic to the anchor. We minimize this upper bound instead, and name it as the *fast approximated triplet* (**FAT**) loss:

$$L_{\text{FAT}} = \sum_{\substack{a,n \\ n \neq y_a}} \max\{0, d(a, c_a) + m - d(a, c_n)\} + R(a) + R(n). \tag{4.2}$$

Algorithm 1 Derivation of FAT loss as an upper bound for triplet loss (2.1).

$$
\begin{aligned}
L_{\text{tri}} &= \max\{0, d(a,p) + m - d(a,n)\} \\
&\leq \max\{0, d(a,c_a) + d(c_a,p) + m - \max\{0, d(a,c_n) - d(c_n,n)\}\} \\
&\qquad \triangleright \text{ refer to both inequalities in (4.1)} \\
&= \max\{0, d(a,c_a) + d(c_a,p) + m - d(a,c_n) + \min\{d(c_n,a), d(c_n,n)\}\} \\
&\qquad\qquad \triangleright \text{ move d(a, } c_n\text{) out of inner } \max \text{ then reverse sign} \\
&= \max\{0, d(a,c_a) + m - d(a,c_n) + d(c_a,p) + \min\{d(c_n,a), d(c_n,n)\}\} \\
&= \max\{0, d(a,c_a) + m - d(a,c_n)\} + d(c_a,p) + \min\{d(c_n,a), d(c_n,n)\} \\
&\qquad\qquad \triangleright \text{ move non-negative sums out of } \max \\
&\leq \max\{0, d(a,c_a) + m - d(a,c_n)\} + d(c_p,p) + d(c_n,n) \\
&\qquad\qquad \triangleright \ c_a = c_p;\ \min\{\text{d}(c_n, \text{a}), \text{d}(c_n, \text{n})\} \leq \text{d}(c_n, \text{n}) \\
&\leq \underbrace{\max\{0, d(a,c_a) + m - d(a,c_n)\}}_{\textbf{anchor-dependent point-to-set loss}} + \underbrace{R(a) + R(n)}_{\textbf{cluster compactness}} \\
&\qquad\qquad \triangleright \text{ R() defines the radius of the cluster (pre-computed)}
\end{aligned}
$$

As the name suggests, the new loss will give rise to similarly competitive ReID performance compared to the full triplet loss, but with tremendously better efficiency. We now analyze FAT loss w.r.t. the triplet loss from two aspects.

As can be obviously seen from its form, FAT loss greatly accelerates the cubic/quadratic time complexity of computing triplet loss, to linear complexity, w.r.t. the training set size. We also examine how tight it approximates the original triplet loss. Observing (1), three relaxations take place in the second, sixth and seven lines. For the first one, the equality in (4.1) could be met when: a, c_a, p are co-linear with a, p on the same side of c_a; while a, c_n, n are also co-linear with a, n on different sides of c_n. The second relaxation becomes tight if and only if $d(a, c_n) \geq d(n, c_n)$, which implies that a is sufficiently far away from the cluster of c_n. For the last one, the exact equality can only be taken in a very special case, when every cluster has the same radius and every sample in a cluster distributes on a circle. In sum, when clusters are well-separated and balanced in size, FAT loss can provide a relatively tighter approximation for triplet loss. However, it is always reasonable to expect that minimizing this upper bound would lead to suppressing the original triplet loss value.

4.3 Normalized FAT Loss

As a margin loss, FAT loss, as well as triplet loss, is sensitive to input scales. Given the fact that ReID features are also scale-sensitive: neighboring features in the normalized space can be far away from each other in the original feature space, the learned feature are often normalized before feeding into the evaluation metrics. That could be reflected in a normalized FAT loss:

$$L_{\text{FATnorm}} = \max\{0, d(\frac{a}{||a||}, c'_a) + m - d(\frac{a}{||a||}, c'_n)\} + R'(a) + R'(n), \tag{4.3}$$

where R' is similarly defined as the radius of the normalized sample set. In practice, we empirically find that adding a cross entropy (CE) loss L_{CE} term will help stabilize training with FAT or Normalized FAT loss notably. That leads to minimizing a hybrid loss ($L_{\text{CE-FAT}}$ can be replaced to $L_{\text{FAT-N}}$; λ is a scalar):

$$L_{\text{CE-FAT}} = L_{\text{FAT}} + \lambda * L_{\text{CE}} \tag{4.4}$$

4.4 Choices of Centroids

The choice of cluster centroids is also found to be critical to the effectiveness of FAT loss. Four options of cluster centroids are available: i) mean of cluster features; ii) mean of normalized cluster features; iii) normalized mean of cluster features; and iv) normalized mean of normalized cluster features. A visual comparison of the four options are in Figure 4.3. Mathematically:

$$\begin{aligned} C_{i1} &= \frac{1}{N_i}\sum_{y_k=i} f_E(X_k), \qquad C_{i2} = \frac{1}{N_i}\sum_{y_k=i} \frac{f_E(X_k)}{\|f_E(X_k)\|} \\ C_{i3} &= \frac{\sum_{y_k=i} f_E(X_k)}{\|\sum_{y_k=i} f_E(X_k)\|}, \quad C_{i4} = \frac{\sum_{y_k=i} \frac{f_E(X_k)}{\|f_E(X_k)\|}}{\|\sum_{y_k=i} \frac{f_E(X_k)}{\|f_E(X_k)\|}\|} \end{aligned} \tag{4.5}$$

Since the original FAT loss (4.2) is calculated based on un-normalized features, only the first centroid option C_{i1} makes sense for it. The remaining three options can all be utilized for the normalized FAT loss (4.3). Our experiments indicate that the normalized mean of normalized cluster features C_{i4} works best with the normalized FAT loss.

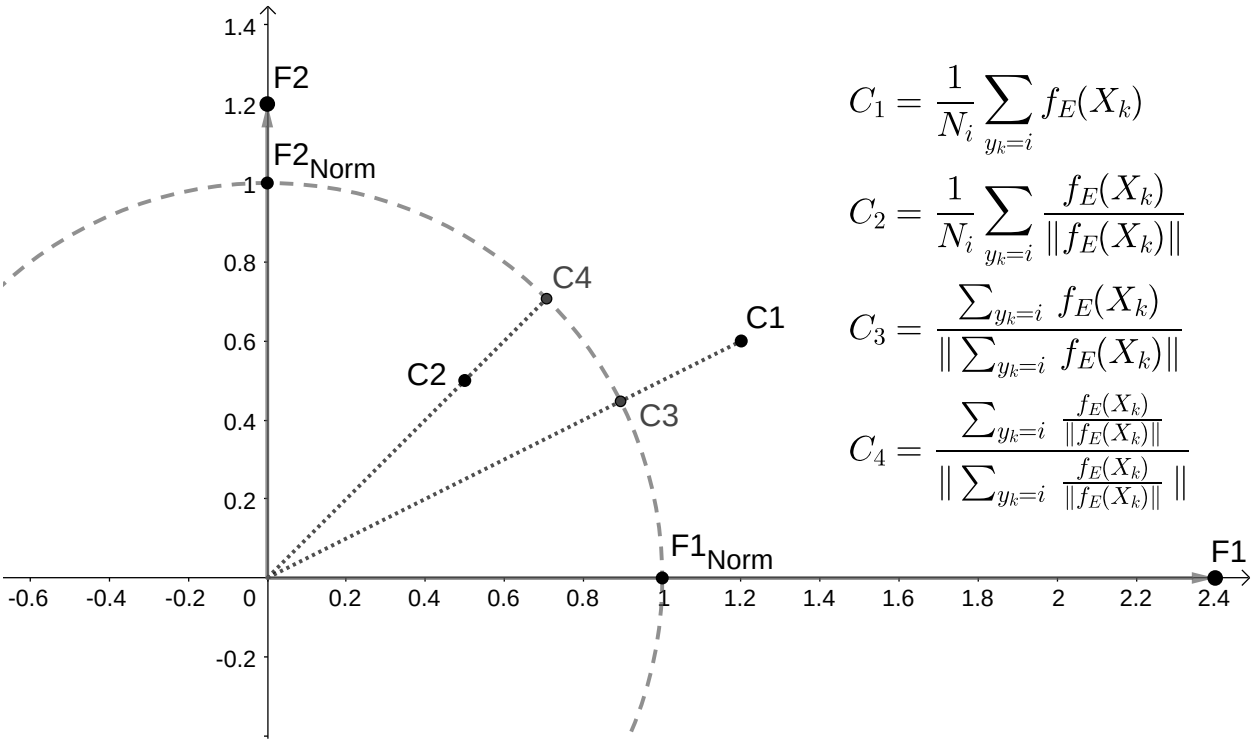

Figure 4.3: Example of Four Different Centroid Options.

4.5 Implementation of FAT Loss

We implement our FAT loss in PyTorch deep learning framework. In the training phase, all images are resized to 144×432 and then randomly cropped into 128×384 sub-images. Standard horizontal flipping is adopted for data augmentation. In the test phase, all images are re-sized to 128×384 and no data augmentations are applied. All images have the training set mean subtracted and then normalized by the training set standard deviation, before feeding into the network.

Following a standard ReID protocol, we use ResNet [184] or Densenet [185] backbone as the feature extractor f_E towards learning a pedestrian representation directly supervised by FAT loss L_{fat}. The cluster centroids are computed at the beginning of each epoch, using C_{i1} for FAT loss and C_{i4} for normalized FAT loss in Equation 4.5. Besides, we also compare four different options of choosing the negative cluster c_n for computing FAT loss each time: i) *ctrdAll:* identity classes that are different from the one a belong to; ii) *ctrdAvg:* consider all other classes, except the one that a

belongs to, as one cluster and obtain one negative centroid by computing the average of all negative centroids, which is similar to [53] but differs in the way of calculating all negative samples' mean; iii) *ctrdHM:* find a hard negative cluster (in terms of closest centroid to the one that a belongs to), from all classes of the whole dataset; iv) *batchHM:* find a hard negative sample on "batch level", *e.g.*, from all classes that are sampled by the current batch.

4.6 Results and Analysis of FAT loss

We first present a comprehensive ablation study on the effectiveness of FAT loss in Table 4.1, using the Market1501 dataset. By default, we use the CE-FAT loss defined in (4.4), with $\lambda = 1$, as it consistently improves over either FAT or CE loss alone. The margin m is chosen as 1 for FAT loss and 0.1 for normalized FAT loss, as validated to be effective in experiments. We study on the four choices of the negative cluster (only ctrdAvg was previously explored in a similar form [53]), as well as the FAT loss hyperparameter (margin m). We also compare CE-FAT with CE-P2S, the latter defined by removing the cluster compactness term in FAT loss; as well as the normalized versions for both, denoted as CE-FATnorm and CE-P2S norm, respectively.

We evaluate different methods in terms of their top-1/top-5/top-10 accuracy and mean average precision (mAP) values obtained on the Market1501 testing set. Moreover, we use the **direct transfer** performance of the Market1501-trained feature extraction to the DukeMTMC-reID dataset, as an additional performance criterion, to avoid overfitting small ReID datasets. A few popular ReID loss options proposed in previous works [186, 175, 187, 49] are also included into comparison, so is a CycleGAN [132] baseline for transfer evaluation. Note that CycleGAN is a domain adaption method that demands re-training on the target domain, while the direct transfer needs no extra re-training.

First, comparing CE-FAT with ctrdAll, ctrdAvg, ctrdHM, and batchNeg, it is clear that batchNeg outperforms the other three. Second, comparing CE-P2S with CE-FAT in fair settings, we show the necessity of cluster compactness regularization in addition to the P2S loss; for example, without the compactness term, we will see 1.8% (ctrdAll) and 2.2% (batchNeg) top-1 accuracy drops on the Market1501 test case, and 7.5% (ctrdAll) and 9.2% d(batchNeg) top-1 accuracy drops

on the transfer case to DukeMTMC-reID. The performance gaps clearly differentiate FAT loss from previous empirical P2S losses, thanks to our more rigorous upper-bound derivation. Third, no performance gain has been observed on Market1501, when using normalized features for FAT/P2S. Finally, CE-FAT outperforms all state-of-the-art losses trained with the same ResNet50, on the Market1501 testing set. Furthermore, after we replace the backbone into DenseNet161, CE-FAT achieves not only further boosted Market1501 testing results, but also impressive direct transfer performance to DukeMTMC-reID, even surpassing Cycle-GAN domain adaption [132] that is re-trained with the target domain data.

Tables 4.2 and 4.3 report similar experiments using DukeMTMC-reID ad MSMT17 datasets, respectively. With most observations aligned with the Market1501 cases, we find the training behavior on MSMT17 to slightly differ from the other two (much) smaller datasets. In particular, while batchNeg remains effective for its own testing set, ctrdAll becomes the best option when it comes to the feature transferability evaluation. That might be attributed to the heavier label noise on MSMT17, that likely benefits from averaging the triplet effects between with current one and all other clusters. Also, we observe CE-FATnorm to outperform CE-FAT, when transferring from MSMT17 to the other two datasets. That implies that normalization may become essential to overcome feature scale variances on large datasets. Finally, training ResNet50 with CE-FAT loss and batchNeg has surpassed the state-of-the-art performance [31] ever reported on MSMT17.

Besides, we use t-SNE to visualize the feature distributions learned using cross entropy loss (Figure 4.4 top) and FAT loss (Figure 4.4 bottom). Twenty identities are randomly selected from the MSMT17 dataset and their IDs are listed below the graphs. We can see that the distances between identity features become much larger when we switch from the cross entropy loss to the FAT loss, indicating that the proposed FAT loss is a better optimization target for maximizing the inter-class distance.

4.7 Conclusion

This work proposes the fast-approximated triplet (**FAT**) loss, which remarkably improves the efficiency over the standard triplet loss in ReID models. Instead of using point-to-point distances,

the FAT loss uses a point-to-set distance with cluster compactness regularization, which is derived rigorously as an upper bound of standard triplet loss, with linear complexity to the training set size. A distillation network is also designed to assign soft labels for samples in place of potentially noisy hard labels. Extensive experiments demonstrate the high effectiveness and promise of the proposed FAT loss along with label distillation.

Settings			Test on Market1501			
loss	negative	margin	top1	top5	top10	mAP
Histogram Loss [186]	NA	NA	59.5	80.7	86.9	
Multi-loss class [175]	NA	NA	83.9	-	-	64.4
Point to Set Similarity [187]	NA	NA	70.7	-	-	44.3
Triplet loss [49]	NA	1	84.9	94.2	-	69.1
Support Neighbor Loss [188]	NA	NA	88.3	-	-	73.4
CE-FAT	ctrdAll	1	89.1	95.0	96.7	71.6
CE-FAT	ctrdAvg	1	89.2	95.3	97.0	72.4
CE-FAT	ctrdHM	1	87.1	94.7	96.3	69.9
CE-FAT	batchNeg	1	89.4	95.6	97.1	73.1
CE-P2S	ctrdAll	1	87.4	95.0	96.7	68.9
CE-P2S	batchNeg	1	87.2	94.6	96.7	67.0
CE-P2Snorm	batchNeg	0.1	87.5	95.3	96.8	68.1
CE-FATnorm	batchNeg	0.1	88.6	95.1	96.7	69.7
CE-FAT* (DenseNet161)	batchNeg	1	**91.4**	**96.6**	**97.7**	**76.4**

Settings			Transfer to DukeMTMC-reID			
loss	negative	margin	top1	top5	top10	mAP
CycleGAN [132]	NA	NA	38.5	54.6	60.8	19.9
CE-FAT	ctrdAll	1	34.4	51.5	57.6	18.9
CE-FAT	ctrdAvg	1	35.1	51.2	57.6	19.2
CE-FAT	ctrdHM	1	34.3	50.8	56.9	18.0
CE-FAT	batchNeg	1	37.3	52.3	58.4	20.3
CE-P2S	ctrdAll	1	27.6	42.9	50.0	14.1
CE-P2S	batchNeg	1	28.1	42.6	49.2	14.3
CE-P2Snorm	batchNeg	0.1	27.8	41.7	48.7	13.6
CE-FATnorm	batchNeg	0.1	35.0	50.6	57.4	18.9
CE-FAT* (DenseNet161)	batchNeg	1	**40.8**	**57.1**	**63.2**	**23.4**

Table 4.1: Comparison Analysis of FAT Loss on Market-1501 Dataset. Evaluation results on Market1501 and transfer results from Market1501 to DukeMTMC-reId. We use Resnet50 as our default backbone and trained on Market1501, with only one exception indicated by * using DenseNet161 backbone.

Settings			Test on DukeMTMC-reID			
loss	negative	margin	top1	top5	top10	mAP
Deep-Person [189]	NA	NA	**80.9**	-	-	**64.8**
CE-P2Snorm	batchNeg	0.1	76.5	87.3	90.6	57.3
CE-FATnorm	batchNeg	0.1	77.9	87.8	91.4	58.3
CE-P2S	batchNeg	1	78.2	88.5	91.8	59.5
CE-FAT	batchNeg	1	78.8	88.7	91.5	60.8
CE-FAT* (DenseNet161)	batchNeg	1	80.8	**89.5**	**92.0**	63.1

Settings			Transfer to Market1501			
loss	negative	margin	top1	top5	top10	mAP
CycleGAN [132]	NA	NA	48.1	66.2	72.7	20.7
CE-P2Snorm	batchNeg	0.1	46.5	63.9	71.0	19.9
CE-FATnorm	batchNeg	0.1	49.8	65.8	73.2	21.2
CE-P2S	batchNeg	1	47.0	64.6	71.4	19.7
CE-FAT	batchNeg	1	49.1	67.1	73.9	21.8
CE-FAT* (DenseNet161)	batchNeg	1	**54.7**	**70.8**	**77.4**	**25.2**

Table 4.2: Comparison Analysis of FAT Loss on DukeMTMC-reID Dataset. Evaluation results on DukeMTMC-reID and transfer results from DukeMTMC-reID to Market1501. We use Resnet50 as our backbone, and trained on DukeMTMC-reID, with only one exception indicated by * using DenseNet161 backbone.

loss	negative set	Test on MSMT17			
CE-P2Snorm	batchNeg	64.8	78.3	83.0	33.8
CE-FATnorm	batchNeg	66.2	79.4	83.7	33.1
CE-P2S	batchNeg	65.2	78.5	82.9	33.7
CE-FAT	ctrdAll	68.8	81.4	85.4	39.1
CE-FAT	ctrdAvg	67.0	80.2	84.6	37.4
CE-FAT	ctrdHM	67.7	80.2	84.5	36.2
CE-FAT	batchNeg	**69.4**	**81.5**	**85.6**	**39.2**

loss	negative set	Transfer to DukeMTMC-reID				Transfer to Market1501			
HHL [46]	NA	45.0	59.4	64.4	23.0	**56.0**	**75.8**	**81.2**	26.7
CE-P2Snorm	batchNeg	49.1	64.9	70.6	29.2	51.6	68.9	75.5	23.9
CE-FATnorm	batchNeg	**51.2**	**66.1**	71.1	29.5	54.8	70.9	76.5	25.1
CE-P2S	batchNeg	49.9	67.6	**74.5**	22.9	48.7	63.5	69.3	**28.5**
CE-FAT	ctrdAll	50.9	65.0	70.2	**30.7**	51.5	69.4	75.9	24.4
CE-FAT	ctrdAvg	45.0	61.7	67.0	25.4	48.3	65.6	73.0	21.5
CE-FAT	ctrdHM	50.1	64.4	70.2	28.4	48.4	66.0	72.5	21.5
CE-FAT	batchNeg	49.2	64.8	69.6	28.7	50.6	68.0	74.9	23.6

Table 4.3: Comparison Analysis of FAT Loss on MSMT17 Dataset. Evaluation results on MSMT17, DukeMTMC-reID, and Market1501. We use ResNet50 as our backbone and trained on MSMT17 with different negative sets.

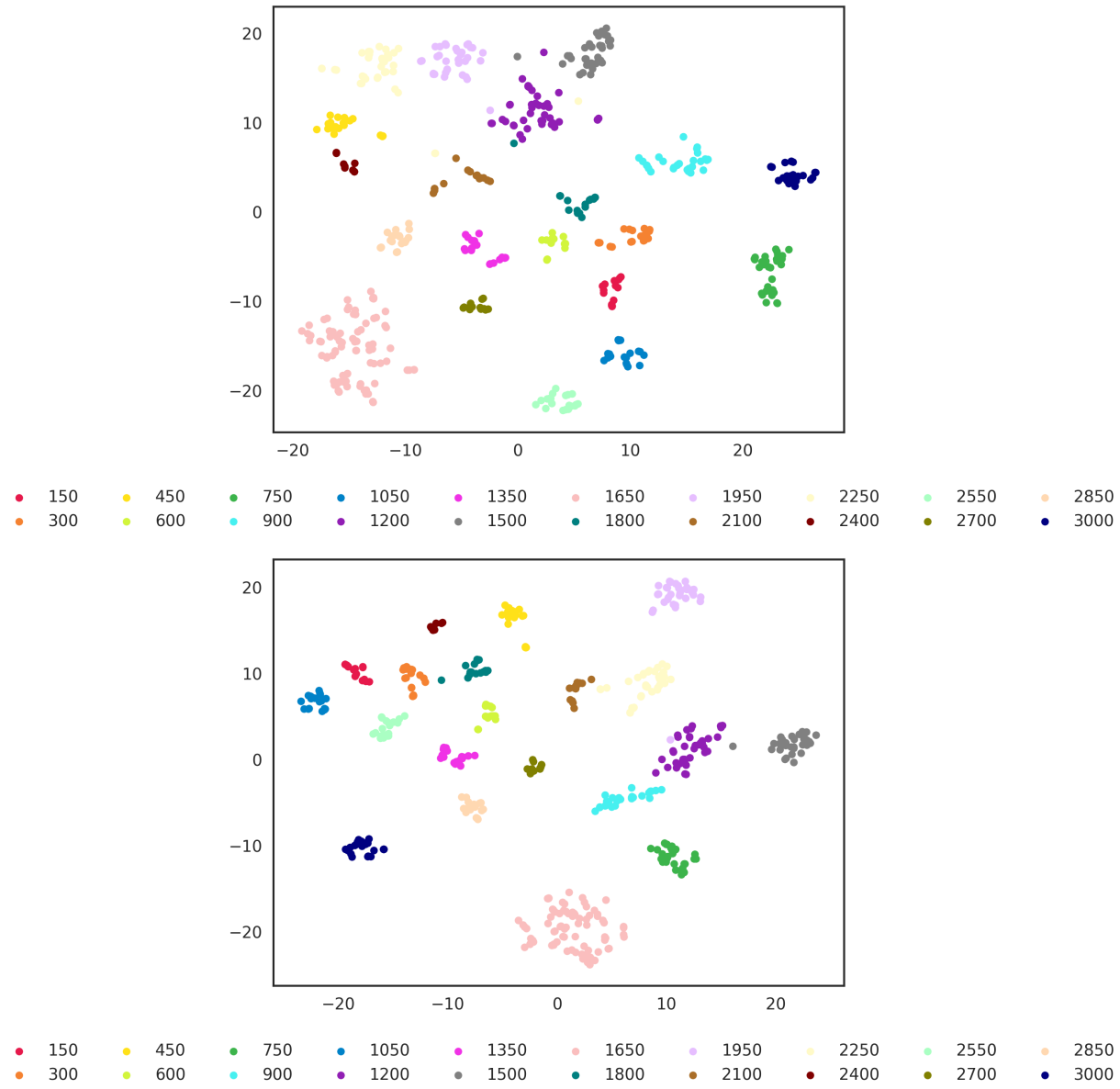

Figure 4.4: T-SNE Visualization of Feature Learned via FAT Loss. Cross entropy loss (top) and FAT loss (bottom). Twenty identities are randomly selected from the MSMT17 dataset and their IDs are listed below the graphs.

5. ROBUST LEARNING WITH NOISY LABEL VIA DISTILLATION NETWORK *

5.1 Motivation

Typically, there are three common label noises in ReID datasets 2.2: i) label flip, i.e., an image is assigned to a wrong identity class; ii) mislabeling, i.e., an image does not belong to any known identity class; iii) multiple identities co-exist in one image. Similar to other margin-based losses, triplet loss is highly sensitive to label noise. Since the proposed FAT loss has a P2S term where all samples within the same cluster are averaged, hence alleviating noisy labels to some extent. We hereby propose a label distillation approach based on a teacher-student model, to improve FAT loss robustness to label noise further, using "soft labels" predicted from another teacher model, trained with a loss that is less sensitive to label noise, *e.g.*, cross-entropy. The pipeline is plotted in Fig.5.1: the teacher network generates soft pseudo labels for each sample associated with a confidence coefficient; then the feature extractor of the teacher network is loaded to student network as pretrained extractor and those soft labels instead of the original noisy labels are feed into the student network for fine-tuning, where each individual samples' contribution to the model update is re-weighted by their label confidence.

Our proposed distillation algorithm is free from the assumption of the existence of a manually-cleaned set. Instead, we train a teacher network with the entire noisy dataset but only use the most confident samples within a batch to update the parameters. We observed that the model updated based on a subset of confident samples can avoid overfitting on the noisy labels and achieve similar performance, compared to the model trained with all noisy-labeled samples.

Besides, we investigate different loss functions for distillation; the teacher network is trained with cross-entropy loss (relatively robust to label noise) with the purpose of providing pseudo soft label associated with a confidence coefficient; the student network is trained with FAT loss (more

* Part of the material reported in this section is reprinted with permission from "In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation" Y. Yuan, W. Chen, Y. Yang, and Z. Wang, *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, June 2020. Copyright 2020 IEEE.

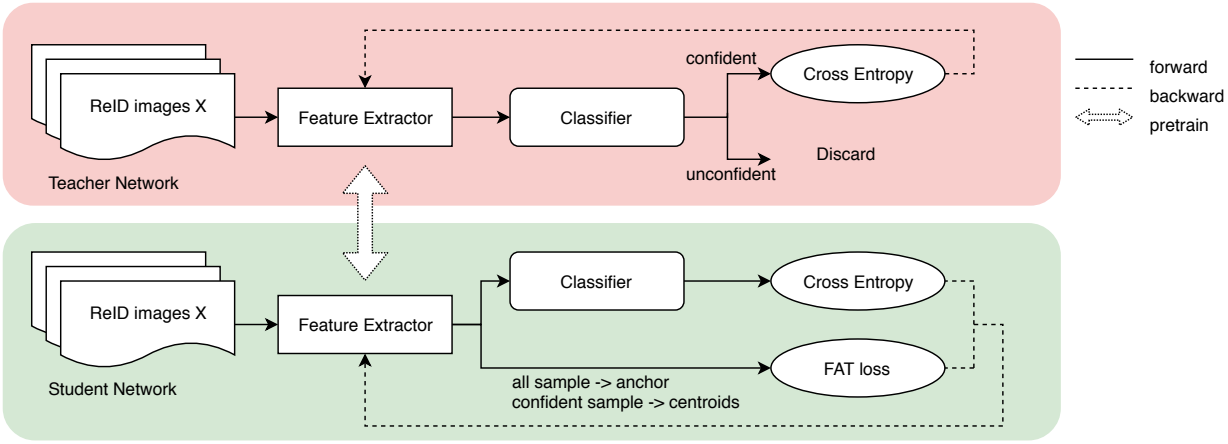

Figure 5.1: Overview of the Label Distillation Pipeline. A distillation network is proposed to assign soft labels learned by teacher network for samples in place of the original potentially-noisy hard labels in the student network.

effective for retrieval task) using the soft pseudo labels generated by the teacher network. Hence instead of mimicking a similar classification behavior as the teacher network, the student network has the capability to "innovate" on a different retrieval task, and eventually outperforms the teacher network for ReID performance.

5.2 Implementation of Label Distillation

Overall, the model is composed of two components, the feature extractor, and a classifier. The feature extractor learns a robust representation of the person images while the auxiliary classifier computes the cross-entropy loss to supervise training.

We first use a self-bootstrapping approach to learn the teacher model robustly. The teacher net is first trained with cross-entropy loss on classifying all samples (including noisy labels) for 5 epochs. It was previously observed that the network would be more inclined to learning with high confidence for "easy samples", within the early stage of training [190, 191]. Those confident, easy samples are hypothesized to have labels that are semantically consistent and correct, less confusing and ambiguous, and therefore more reliable. We identify those most confidently predicted samples based on the entropy of their currently predicted softmax vectors. We then resume training for another 5 epochs; but now in each epoch, we will keep using those identified confident samples,

while not using or only partially using the others that are more likely to contain label noise or outliers. We periodically repeat the above process, and each time we may gradually enlarge the pool of confident examples as the training continues.

After the teacher model is trained, its predictions are treated as soft labels to replace the original labels, for training the student model with FAT loss. Only the "confident" labels eventually selected by the teacher net will participate in averaging to estimate the cluster centroids. If we use the hybrid FAT loss (4.4), then soft labels are the prediction targets for the cross-entropy (softmax) loss too.

Following the basic routine described above, we further study four different modes of identifying confident samples: i) *hard threshold*: select all samples whose softmax entropy values are below a pre-set threshold t as the trusted training subset, and discard all un-selected samples; ii) *soft threshold*: select all samples whose softmax entropy values are below a pre-set threshold $t/2$, and then randomly select 50% of the remaining (unselected) samples to add into the trusted training subset; iii) *hard percentage*: always select 50% samples with lowest softmax entropy values, as the trusted training subset; iv) *hard percentage*: always select 25% samples with lowest softmax entropy values first, and then randomly select another $1/3$ from the remaining 75% (unselected) samples to add into the trusted training subset.

The important difference between "threshold" and "percentage" methods lies in whether we keep a constant or dynamic size of the trusted training subset for the teacher model. For the first two threshold-based methods, even sticking to the same t throughout one training, the portion of samples selected into the trusted set will be dynamic, as more samples might become better confident as training continues. Figure 5.2 visualizes this trend: given $t \leq 0.1$, the final training stage will always have considered all training samples as trusted; while a larger t may lead to more "conservative" selection. We choose t = 0.1 as the empirical default value found in experiments for i) and ii). Also, for the two "soft" strategies ii) and iv), our hope is to utilize a larger set of samples while letting the stochastic selection "smooth out" the impacts from noisy labels.

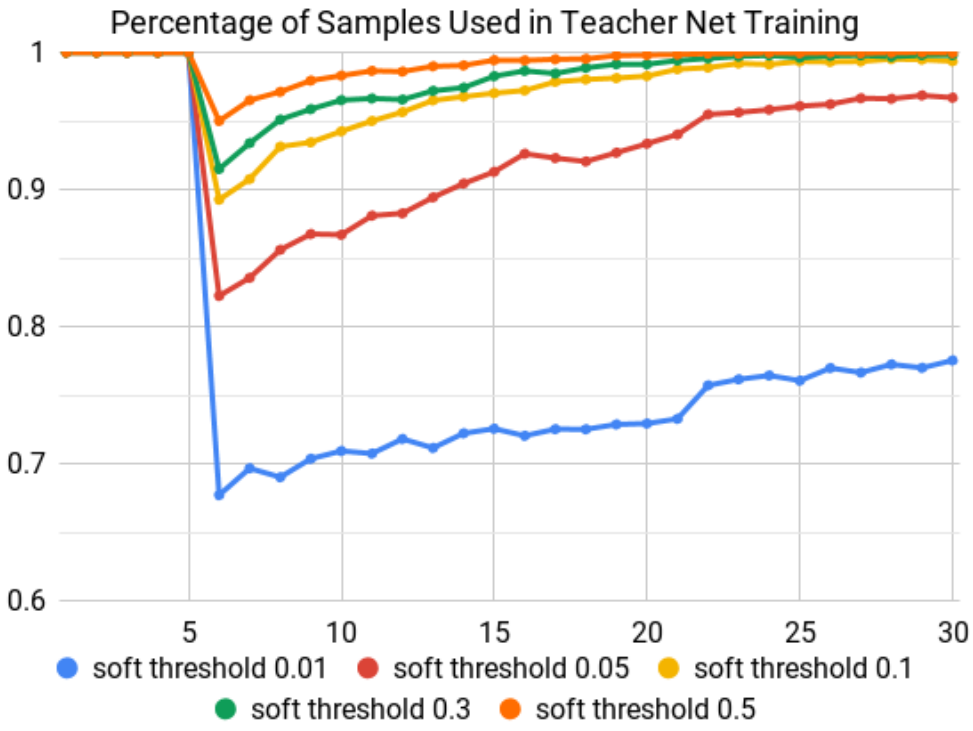

Figure 5.2: Illustration of Teacher-Student Network Training Process. The number of samples actually used as the trusted training subset, when training the ResNet-50 teacher model with different soft threshold t values, on the Market1501 dataset.

5.3 Effect of Label Distillation

To overcome the noisy label issue on MSMT17, we next investigate label distillation to further unleash the power of FAT loss. Both teacher and student nets adopt the same ResNet50 backbone for simplicity.

As shown in Table 5.1, for the training of the teacher net, the soft threshold/percentage methods appear to outperform their hard counterparts, as they can learn with a wider variety of samples (while hard methods may tend to select too many similar easy samples), meanwhile smoothing out the negative impacts of potential noisy samples due to stochastic sampling/averaging effects. In comparison, the soft threshold seems to produce superior results on the same MSMT17 testing set, whereas soft percentage leads to better feature transferability. It implies that soft percentage suffers from less overfitting, because of its curriculum-style learning (as Figure 5.2 shows) that

progressively takes into account the entire dataset information. To our surprise, our teacher net trained with only the trusted subsets by soft threshold/percentage yield competitive or even superior performance than the one trained with the whole dataset, in particular on transfer cases. That proves that the teacher net learns effectively and without being misled by noisy labels.

We then pick the teacher net trained with soft percentage, due to its best transfer performance, to provide soft pseudo labels for training the student net. The training of the student net is supervised by the CE-FAT loss with the batchNeg strategy, using the soft pseudo labels in place of original one-hot labels for both CE and FAT terms. The new model in Table 5.2, dubbed CE-FAT-distillation, does not lead to better test results on MSMT17 than our best result (CE-FAT with batchNeg). However, it produces state-of-the-art **direct transfer** performance from MSMT17 to DukeMTMC-reID. Its transfer performance to Market1501 largely surpasses that of CE-FAT without distillation, and shows competitiveness to state-of-the-art HHL domain adaption [46]. To re-iterate, the direct transfer does not re-train on target domain data as domain adaption has to.

5.4 Conclusion

This work proposes the fast-approximated triplet (**FAT**) loss, which remarkably improves the efficiency over the standard triplet loss in ReID models. Instead of using point-to-point distances, the FAT loss uses a point-to-set distance with cluster compactness regularization, which is derived rigorously as an upper bound of standard triplet loss, with linear complexity to the training set size. A distillation network is also designed to assign soft labels for samples in place of potentially noisy hard labels. Extensive experiments demonstrate the high effectiveness and promise of the proposed FAT loss along with label distillation.

Method	Test on MSMT17			
	top1	top5	top10	mAP
whole set	**65.1**	78.2	82.8	**34.5**
hard threshold	64.5	77.8	82.2	33.7
soft threshold	64.8	**78.3**	**83.0**	34.2
hard percentage	64.2	77.5	82.1	34.2
soft percentage	62.9	76.1	80.9	32.6

Method	Tranfer to DukeMTMC-reID				Tranfer to Market1501			
	top1	top5	top10	mAP	top1	top5	top10	mAP
whole set	48.2	63.8	69.9	29.0	51.1	68.3	74.2	23.5
hard threshold	46.5	62.8	69.0	27.4	49.9	66.2	73.3	23.0
soft threshold	48.2	63.5	69.0	28.9	49.6	67.3	74.1	23.1
hard percentage	49.3	64.4	69.8	29.8	52.0	69.2	**76.5**	**24.8**
soft percentage	**50.5**	**66.0**	**71.0**	**30.3**	**52.4**	**69.6**	76.0	24.6

Table 5.1: Performance of Teacher Network in Label Distillation. Evaluation results of the Teacher Net on MSMT17, DukeMTMC-reID, and Market1501. We use ResNet50 as our backbone and trained on MSMT17.

loss	negative set	Test on MSMT17			
CE-FAT	batchNeg	**69.4**	**81.5**	**85.6**	**39.2**
CE-FAT-distillation	batchNeg	66.2	79.2	83.6	36.5

loss	negative set	Transfer to DukeMTMC-reID				Transfer to Market1501			
HHL [46]	NA	45.0	59.4	64.4	23.0	**56.0**	**75.8**	**81.2**	**26.7**
CE-FAT	batchNeg	49.2	64.8	69.6	28.7	50.6	68.0	74.9	23.6
CE-FAT-distillation	batting	**50.9**	**66.6**	**72.2**	**31.3**	52.8	69.2	75.9	25.4

Table 5.2: Performance of Student Network in Label Distillation. Evaluation results of the Student Net on MSMT17, DukeMTMC-reID, and Market1501. We use ResNet50 as our backbone and trained on MSMT17.

6. DOMAIN-INVARIANT LEARNING FOR LARGE-SCALE APPLICATIONS *

6.1 Motivation

With rapidly increasing demand for ReID in multi-camera systems such as for public safety, indoor/outdoor monitoring, traffic surveillance and smart city/community, the core technical challenge of ReID problem is no longer just the performance in an enclosed or fixed environment: it has to stay effective to new subjects, scale up to new locations, and be reliable over time.

However, the scale and diversity of existing ReID datasets are still far from being comparable to real scenarios. A recent study [26] showed that in 2014, there were 125 video surveillance cameras per thousand people in the U.S.; whereas most ReID datasets were collected only from 10 or fewer cameras (see Section 2.1.1). Trained on limited data, most existing ReID algorithms may not have addressed the generalization issue well: the model's robustness and transferability can not be extend to diverse and large-scale unseen cases, such as changing background, illumination, viewpoint, and other camera parameters, which may hinder their deployment in practice.

Most scene-related nuisances are caused by camera-specific and/or time-specific factors. Fortunately, video timestamp or camera index are freely available in video surveillance as metadata and are provided by almost all the existing ReID datasets. The nuisance labels can be potentially utilized as auxiliary supervision, although few image-based ReID methods have taken advantage of them. Inspired by [192, 193], we aim to improve the generalizability of ReID models in large-scale settings, by resorting to a novel domain-invariant feature learning perspective.

We consider samples (of different subjects) with the same nuisance to be from one domain (such as images captured by the same fixed camera, or in the same time period). This is because scene-related changes (background, illumination, viewpoint, etc.) heavily dominate the appearances of images. Different types of nuisances hence becomes domain-specific features. In con-

* Part of the material reported in this section is reprinted with permission from "Calibrated domain-invariant learning for highly generalizable large scale re-identification" by Y. Yuan, W. Chen, T. Chen, Y. Yang, Z. Ren, Z. Wang, and G. Hua, *in Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2019. Copyright 2019 by IEEE.

trast, one subject can be captured at different cameras and time periods, and the subject's identity features should apparently remain domain-invariant. Therefore, our main idea is to extract ReID features that can: (1) be utilized to faithfully classify subjects into correct classes; (2) be resilient and invariant to those identified nuisances – in other words: no reliable classifier can be trained on those features to predict those nuisances.

We formulate our adversarial domain-invariant learning framework (**ADIN**), by taking advantage of "free" annotations like video timestamp and camera index, to separate identity-related features from scene-specific nuisances. To our best knowledge, we are the first to utilize those "free" annotations for image-based ReID, to effectively suppress the overfitting of nuisances. Moreover, we find the imbalance of nuisance distribution w.r.t. subjects hampers the adversarial learning. A novel calibrated adversarial loss is therefore introduced to tackle the nuisance class imbalance for ADIN. Measured by a new direct transfer performance criterion (discussed in Section 2 Section 2.1.2) on several popular large-scale ReID benchmarks, our ADIN demonstrates outstanding generalizability and outperform previously reported results and even some that rely on domain adaptation using target data.

6.2 Domain-Invariant Learning Formulation

Given a training image X with the identity labels Y_I and the (freely) available nuisances label Y_N (one or multiple, such as camera ID, video timestamp, etc.), our goal is to learn a feature representation $f_E(X)$ that is highly **relevant** to the identity label, yet being invariant or **irrelevant** to the nuisances label. Using a function R to represent the correlation between the feature and the label, our learning goal is mathematically described as:

$$R(f_E(X), Y_I) \approx R(X, Y_I), \qquad R(f_E(X), Y_N) \ll R(X, Y_N). \tag{6.1}$$

We adopted an identity prediction module f_I which projects the feature $f_E(X)$ into identity-related features, and a nuisance prediction module f_N that extracts scene-specific nuisances from $f_E(X)$. Without loss of generality, both of them are assumed to have softmax-form outputs. Note

that f_E, f_I and f_N all need to be learned together. Their interactions provide mutual supervision. In particular, f_N will serve as an "adversary" role.

To evaluate R practically, a straightforward choice is to use two standard classification-oriented loss functions L_I and L_N (*e.g.*, cross-entropy) for f_I and f_N respectively and minimize the classification error rate of Y_I from $f_E(X)$, while maximizing the classification error rate of Y_N from $f_E(X)$. Our task then becomes to simultaneously train f_E, f_I and f_N, so as to minimize the identity classification loss meanwhile maximizing the nuisance classification loss.

$$\min_{f_E} L_I(f_I(f_E(X)), Y_I), \qquad \max_{f_E} L_N(f_N(f_E(X)), Y_N). \tag{6.2}$$

Maximizing $L_N(f_N(f_E(X)), Y_N)$ is not straightforward to implement. Previous work [194] reversed the sign of gradient computed from minimizing $L_N(f_N(f_E(X)), Y_N)$, *i.e.*, using gradient ascent. However, we observed in experiments that the reverse gradient approach yielded unstable training process. Instead, we introduce a new L_{adv} loss, to encourage the *disparity* between $f_N(f_E(X))$ and Y_N: a smaller L_{adv} value is expected to indicate a *worse* correlation between them. A detailed discussion about the choice of L_{adv} will be presented in section 6.3.

Finally, the training goal of ADIN is represented below ($\beta > 0$ is a scalar):

$$\min_{f_E} L_I(f_I(f_E(X)), Y_I) + \beta L_{adv}(f_N(f_E(X)), Y_N). \tag{6.3}$$

Meanwhile, in order to keep adversarial domain-invariant feature learning effective so as to learn meaningful f_E, we need to also maintain f_N to be a strong competitor. That implies a *hidden constraint*, *i.e.*, avoiding $L_N(f_N(f_E(X)), Y_N)$ growing large too quickly, in which case f_N becomes to have too poor nuisance classification ability so that it cannot make a useful adversary.

6.3 Calibrated Adversarial Loss for Imbalanced Nuisances

As noted in Section 1, both subject and nuisance classes (conditioned on the subject) suffer from sample imbalances. We experimentally observed the subject imbalance to have less severe impact on ReID performance (*e.g.*, comparing using standard and reweighted softmax loss), and therefore keep using a standard softmax function for L_I. However, the nuisance class imbalance was found to cause considerable training instability and performance degradation for the adversarial learning. We thus focus a detailed discussion on how we derive a robust L_{adv} for the imbalanced nuisances.

We denote $c = [c_1, ..., c_K]$ as the softmax-form output of f_N, where K is the total nuisance class number. We next present three options that we tried for L_{adv}, among which our proposed new Option #3 is experimentally validated to be the best choice for ADIN (see section 6.6.1 for details).

Option #1: Reverse Gradient (RG). One possibility is to adopt the reversal gradient layer [195]. It computes the gradient for minimizing the cross-entropy between $f_N(f_E(X))$ and Y_N, then reversing the gradient sign. However, this objective becomes problematic in our case, as it was observed to cause large fluctuations in the training curve and failure of convergence. Moreover, when both f_N and f_E are initialized from pre-trained models (practically improving convergence and results), the gradients start with very small magnitudes and the model updates become too slow. RG is written as (Y^* is the true label):

$$L_{adv}(X, Y_N) = -L_N = \sum_{k=1}^{K} \mathbb{1}_{[k=Y^*]} \log(c_k) \tag{6.4}$$

Option #2: Negative Entropy (NE). An alternative is to minimize the negative entropy function of the softmax vector (or equivalently, its *cross-entropy* with uniform distribution), as to encourage "uncertain" predictions of nuisance attributes (*e.g.*, camera ID and video timestamps) from the extracted ReID features. The rationale is that, if the nuisance prediction is only as good as the random guess (uniform distribution over all classes), then the feature is not informed of nuisances

and therefore can generalize to unseen nuisances. NE could be written as

$$L_{adv}(X, Y_N) = \sum_{k=1}^{K} c_k \log(c_k). \tag{6.5}$$

Importantly, although Y_N does not explicitly occur in the loss form, it will still be utilized in re-training f_N to make a sufficiently strong competitor (section 6.4). We previously also tried the KL Divergence and the Jensen-Shannon Divergence between the softmax and uniform distribution, but NE appears to work best in practice.

Option #3 (Proposed): Calibrated Negative Entropy Loss (CaNE). Despite boosting uncertainty, NE overlooks the practical imbalance of nuisance class distribution w.r.t. subjects. A well-known solution is to add a modulating factor to cross-entropy loss, ensuring that the majority class/easy decisions do not overwhelm the loss [196]. We propose a reweighted form of NS, called Calibrated Negative Entropy Loss (CaNE), to make L_{adv} attentive to the skewed nuisance distribution

$$L_{adv}(X, Y_N) = \sum_{k=1}^{K} p_k c_k \log(c_k), \tag{6.6}$$

where p_k denotes the nuisance class distribution in the given training set. To our best knowledge, there has been no similar discussion addressing the class imbalance issue in (adversarial) domain adaption among existing ReID works.

6.4 Training Strategy Overview

Figure 6.1 overviews the concrete training workflow of ADIN, which consists of three modules: feature extractor f_E, subject identity classifier f_I, and nuisance classifier f_N. f_E takes the image X as input and outputs the feature $f_E(X)$, which is then passed through f_I and f_N. Both f_I and f_N aim to accurately predict their corresponding labels from the learned features. The training of

f_E strives to boost the prediction of $f_I(f_E(X))$, while suppressing the prediction of $f_N(f_E(X))$. It is important to keep f_N strong to maintain a meaningful competition for learning nontrivial f_E.

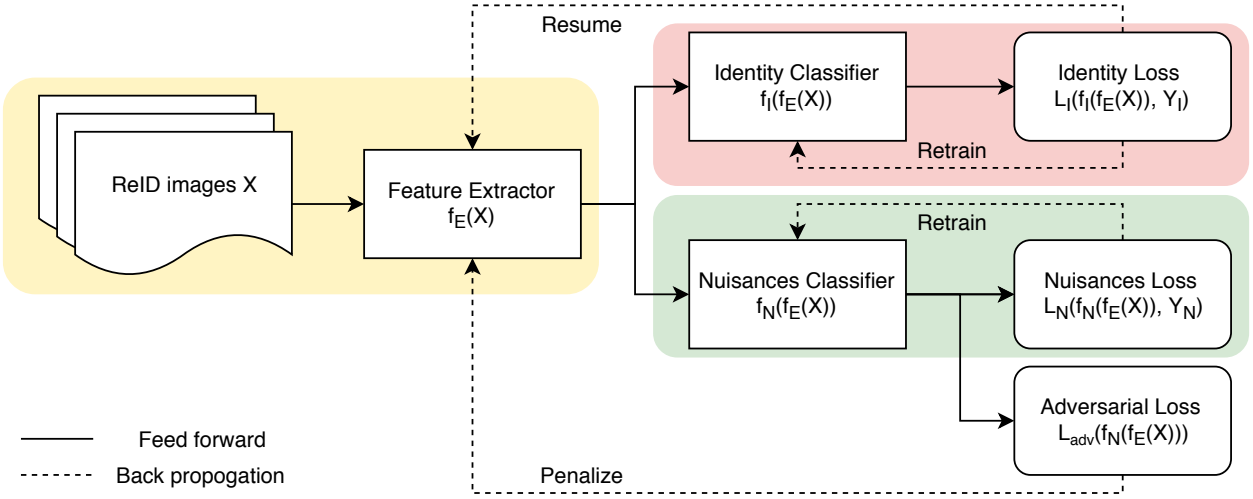

Figure 6.1: Overview of the ADIN framework. Illustration of its training strategy.

In practice, we implement the training using an iterative strategy. We initialize f_E, f_I and f_N by jointly training the feature extractor f_E and identity classifier f_I, and then fixing f_E and pre-training f_N solely on top of that. Afterwards, we alternate between optimizing two sub-problems:

$$\min_{f_E, f_N} L_{adv}(f_N(f_E(X))), \quad \min_{f_E, f_I} L_I(f_I(f_E(X)), Y_I). \tag{6.7}$$

In each alternating round, we optimize the first objective until the validation error of identity classification reducing below a pre-set $threshold_{\text{I-target}}$. We then switch to optimizing the second objective, meanwhile monitoring the resulting changes on the identity classification validation error (since f_E is altered): if it drops below another pre-set $threshold_{I-trigger}$, we will switch back to the first object and start the next round of alternations.

Algorithm 2 The Training Strategy of ADIN Framework.

Given pre-trained feature extractor f_E, identity classifier f_I and nuisances classifier f_N
$val_I, val_N \leftarrow$ identity classifier validation accuracy, nuisances classifier validation accuracy.
for number of training epoches **do**
  if $val_I < threshold_{I-trigger}$ **then** ▷ Avoid weak identity recognition performance
    while $val_I \leq threshold_{I-target}$ **do**
      for number of batches **do**
        Sample minibatch of m examples $\{X_1, ..., X_m\}$
        Jointly update the f_E and the f_I by descending its gradient with loss L_I
      end for
      $val_I \leftarrow$ identity classifier validation accuracy.
    end while
  else if $val_N > threshold_N$ **then** ▷ Suppress nuisance discriminator performance
    Feed all training examples $\{X_1, ..., X_n\}$ into the model
    Jointly update f_E and f_N by descending its gradient with the adversarial loss L_{adv}
  else ▷ Further boost identity recognition performance
    for number of batches **do**
      Sample minibatch of m examples $\{X_1, ..., X_m\}$
      Jointly update f_E and f_I by descending its stochastic gradient with loss L_I
    end for
  end if
  Re-initialize f_I, f_N
    ▷ Empirically restart the classifier to avoid it overfitting extracted features
  Train f_I, f_N by descending its gradient with classification loss L_I, f_N correspondingly
  $val_I, val_N \leftarrow$ identity classifier validation accuracy, nuisances classifier validation accuracy.
end for

6.5 Implementation of ADIN Framework

As a general framework, ADIN can take any backbone for f_E, f_I and f_N. In section 6.6.1, we first test our ADIN with f_E being a basic ResNet50 [184] to illustrate the effectiveness of our adversarial training. Afterwards, we adopt a more sophisticated dual-branch feature extractor for f_E, as inspired by [44, 45, 16], to demonstrate further boosted performance over state-of-the-arts. The configuration of the dual-branch model is depicted in Fig.6.2. For our dual-branch backbone, the first four blocks share the same design as in ResNet50. After the forth block, the network was split into a global and a local branch. In the global branch, the feature passes a global average-pooling and then is fed into the classifier. In the local branch, feature is horizontally partitioned into two equal parts, where each part adopts a separate global average-pooling layer and classifier. During inference the outputs from two branches are concatenated together as the final feature for image retrieval. On top of the f_E, we append two simple classifiers as f_I and f_N (Fig. 6.1), either taking two fully connected layers. L_I is always implemented using the hybrid loss of cross-entropy and center loss [197]. An ablation study of L_{adv} is presented in section 6.3; after that, the Calibrated Negative Entropy (CaNE) loss will be our default L_adv unless otherwise specified.

6.6 Results and Analysis

6.6.1 Ablation Study of the Adversarial Loss L_{adv}

Table 6.1 displays a step-by-step comparison for choosing L_{adv}, with the direct transfer performance from DukeMTMC-ReID (source domain) to Market1501 (target domain) as the indicator. Without the adversarial domain-invariant training, both the ResNet50 and Dual-branch backbones achieve low direct transfer accuracy, due to the domain discrepancy across two datasets. We also empirically observe the adversarial effect provided by the reverse gradient (RG) hard to converge, owing to the gradient vanishing/explosion and its sensitivity to the loss magnitude from the nuisance classifier f_N. With the negative entropy (NE) loss, our adversarial domain-invariant training forces the entropy of the nuisance classifier's prediction to be maximized, leading to nuisance-uninformative features learned by the feature extractor f_E and reliable direct transfer performance.

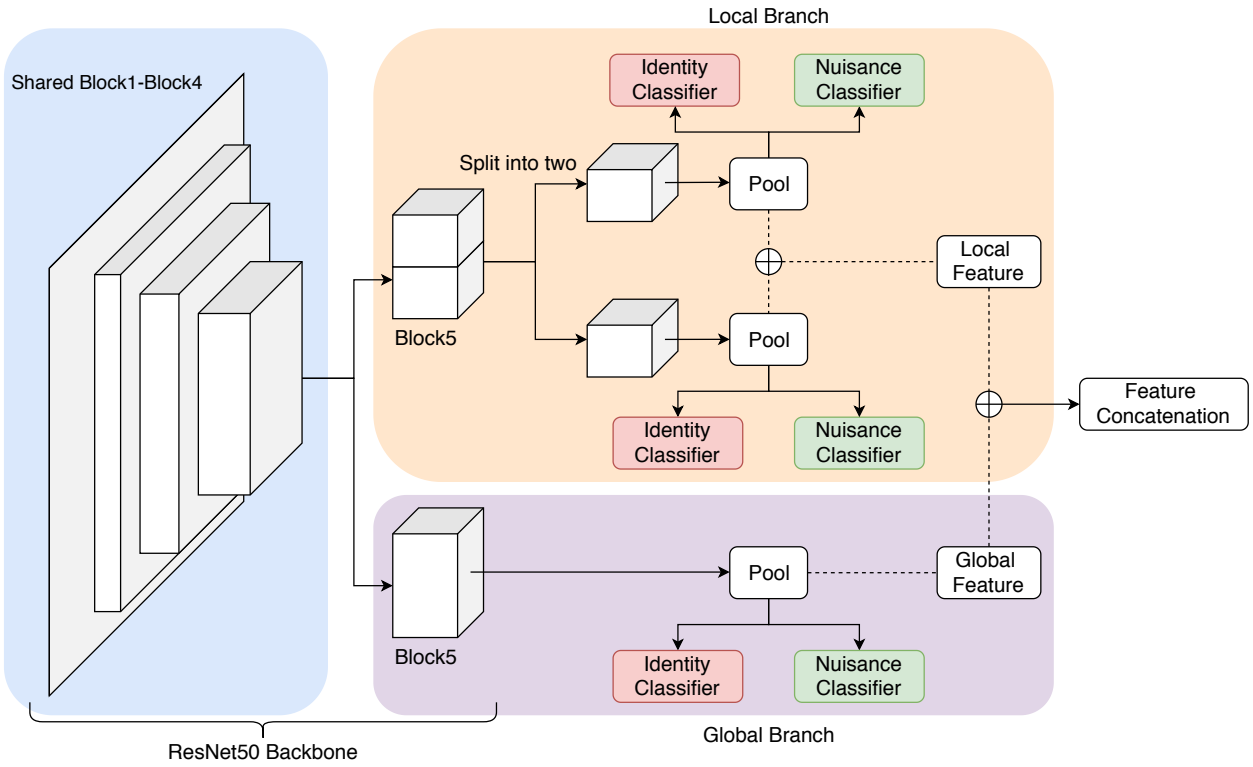

Figure 6.2: Overview of the Dual-Branch Backbone.

More importantly, as pointed out in section 2.1, the sampling of nuisances are imbalanced, which intrinsically results in imbalanced levels of adversarial effects on each nuisance. Thus our proposed calibrated negative loss (CaNE) further enables the adversarial training to be attentive w.r.t. different nuisances frequencies. Table 6.1 shows that both backbones benefit most from our proposed CaNE adversarial loss. It is worth noting that even trained within a small-scale domain like DukeMTMC-ReID, the generalizability of both of the two backbones can be boosted by ADIN.

6.6.2 Direct Transfer between Datasets without Retraining or Adaption

We evaluate three direct transfer cases, two on person ReID: MSMT17 $\rightarrow$ DukeMTMC-ReID, MSMT-17 $\rightarrow$ Market1501; and one on vehicle ReID: VeRi-776 [198] $\rightarrow$ VehicleID [32]. As comparison baselines, we train the same dual-branch backbones (without any adversarial learning) on the source datasets, and test their direct transfer performance too. We train and compare with sev-

Settings		DukeMTMC-ReID → Market1501			
		top1	top5	top10	mAP
ResNet50 (baseline)		46.8	63.5	70.3	19.0
ADIN	ResNet50 + Reverse Gradient	Unable to converge			
	ResNet50 + NE	48.8	66.2	72.7	20.4
	ResNet50 + CaNE	51.7	68.6	76.0	22.1
Dual-branch (baseline)		54.8	71.7	77.6	25.9
ADIN	Dual-branch + Reverse Gradient	Unable to converge			
	Dual-branch + NE	55.9	72.5	78.6	26.5
	Dual-branch + CaNE	**57.2**	**73.0**	**80.0**	**27.4**

Table 6.1: Ablation Study of Adversarial Loss. Performance of different L_{adv} (direct transfer from DukeMTMC-ReID to Market1501).

eral state-of-the-art ReID models on MSMT17: Spatial-Attention [43], PCB [44], RPP [44], MGN [45] (Person ReID); and RAM [199] (Vehicle ReID). We further compare with existing best performers of domain adaptation: HHL [46] (Person ReID), and DAVR [200] (Vehicle ReID), which reported the current best transfer results between DukeMTMC-ReID/Market1501, and from VeRi-776 [198] to VehicleID [32], respectively. Note that both HHL and DAVR need to use (unlabeled) target domain data and perform extra (re-)training for the source domain models, while ours need not: the comparisons are thus apparently to our competitors' advantage.

As can be seen from Tables 6.2 and 6.3, while baselines without adversarial learning fail to transfer well as expected, ADIN demonstrates highly impressive results on all three transfer cases. In particular, by training on MSMT17 and directly transferring, ADIN not only surpasses the direct transfer results from other methods but also outperforms state-of-the-art ReID domain adaption model HHL [46] and DAVR [200], while costing literally none of their hassles such as (re-)training.

In contrast to our ADIN, we find other (single-dataset) top-performers [43, 44, 45] generalize very poorly to unseen domains, indicating the misaligned goals between overfitting small-scale single dataset, and generalizing to large-scale unseen scenarios in real life. As in Fig.2.1, our ADIN framework lies in the top-right corner, while others stay in the left region with high single-

	MSMT17 → DukeMTMC-ReID			MSMT17 → Market1501		
	top1	top5	mAP	top1	top5	mAP
Spatial-Attention[43]	52.2	68.1	32.9	49.7	68.9	25.1
PCB[44]	54.4	69.6	34.6	52.7	71.3	26.7
RPP[44]	56.7	71.4	36.7	50.2	70.7	26.3
MGN[45]	55.5	70.2	35.1	48.7	66.9	25.1
HHL[46]*	**62.2**	**78.8**	31.4	46.9	61.0	27.2
ResNet50 (baseline)	49.7	65.7	28.2	47.7	64.3	21.2
ResNet50 + CaNE	52.6	67.9	30.4	50.1	66.4	22.5
Dual-branch	59.5	73.5	38.4	57.8	73.9	29.4
Dual-branch + CaNE	60.7	74.7	**39.1**	**59.1**	**75.4**	**30.3**

Table 6.2: Generalizability Evaluation of ADIN on Person ReID datasets. Direct transfer performance from MSMT17 [31] to DukeMTMC-ReID [28, 29] and to Market1501 [27].
* indicates method using images from both source and target domain.

Method	Test size = 1600			Test size = 3200		
	top1	top5	mAP	top1	top5	mAP
RAM[199]	30.5	49.5	39.5	24.5	40.3	32.4
Spatial-Attention[43]	39.5	57.2	47.9	33.7	49.6	41.6
PCB[44]	41.3	58.8	49.7	35.4	51.4	43.2
RPP[44]	40.6	58.4	49.1	35.0	51.1	42.9
MGN[45]	39.9	62.4	50.6	32.7	53.1	42.7
DAVR[200]*	45.2	64.0	49.7	38.7	55.9	42.9
ResNet50 (baseline)	42.3	58.5	46.2	36.1	52.2	39.9
ResNet50 + CaNE	43.3	59.7	47.2	37.0	53.4	40.9
Dual-branch	47.3	65.3	51.6	41.2	57.9	45.3
Dual-branch + CaNE	**48.7**	**67.3**	**53.1**	**42.1**	**59.5**	**46.3**

Table 6.3: Generalizability Evaluation of ADIN on Vehicle ReID datasets. Direct transfer performance from VeRi-776 [198] to VehicleID[32].
* indicates method using images from both source and target domain.

dataset accuracy but poor direct transfer performance. We believe the effective direct transfer is the right choice for evaluating and promoting larger-scale ReID practice, and hope our proposals and arguments could invoke more discussions in the community.

6.6.3 Visulization of Retrieval Correctness via ADIN Framework

Figure 6.3 shows five visual retrieval examples. In both queries, the spatiotemporal nuisances (*e.g.* the door of same geo-location, certain lighting condition or glare) have a strong presence in images. As can be seen in the top row of each case, the baselines overfit background, tending to retrieve images with similar nuisances (illumination, viewpoints, scene backgrounds, etc.). In contrast, ADIN successfully eliminates them, and leads to much more robust matching under drastic visual appearance changes, as seen in the bottom rows.

To understand the model behaviors with or without the ADIN framework, we present a visualization of sample feature space in Figure 6.4. We randomly select ten identity classes from MSMT17, and plot the t-SNE of $f_E(X)$ using ResNet50 backbone, before and after using adversarial learning, in the top and bottom rows, respectively. For the left column of Figure 6.4, we color those points based on their subject identity classes (using numerical IDs in the original dataset); for the right column, the points are colored based on their nuisances labels (timestamps). The examples are colored based on their identity IDs (the numerical showing in the left) and nuisances labels (showing in the right). Figure 6.4 indicates that without adversarial loss, the points are strongly clustered according to their nuisance labels. For example, the identity 1200 cluster can be viewed as composed of two subgroups caused by nuisance variations, with images taken in the afternoon more similar to identity 1500, while those taken in the noon are more similar to identity 1800. After nuisance elimination, the clustered structure based on timestamp is eliminated, showing the invariance of the learned features to nuisances (*e.g.*, samples of identity 1200 are now well-mixed among different timestamps). Meanwhile, the grouping structure on the identity classes is preserved with ADIN framework, and in fact displays larger inter-cluster variances, suggesting more favorable retrieval performance.

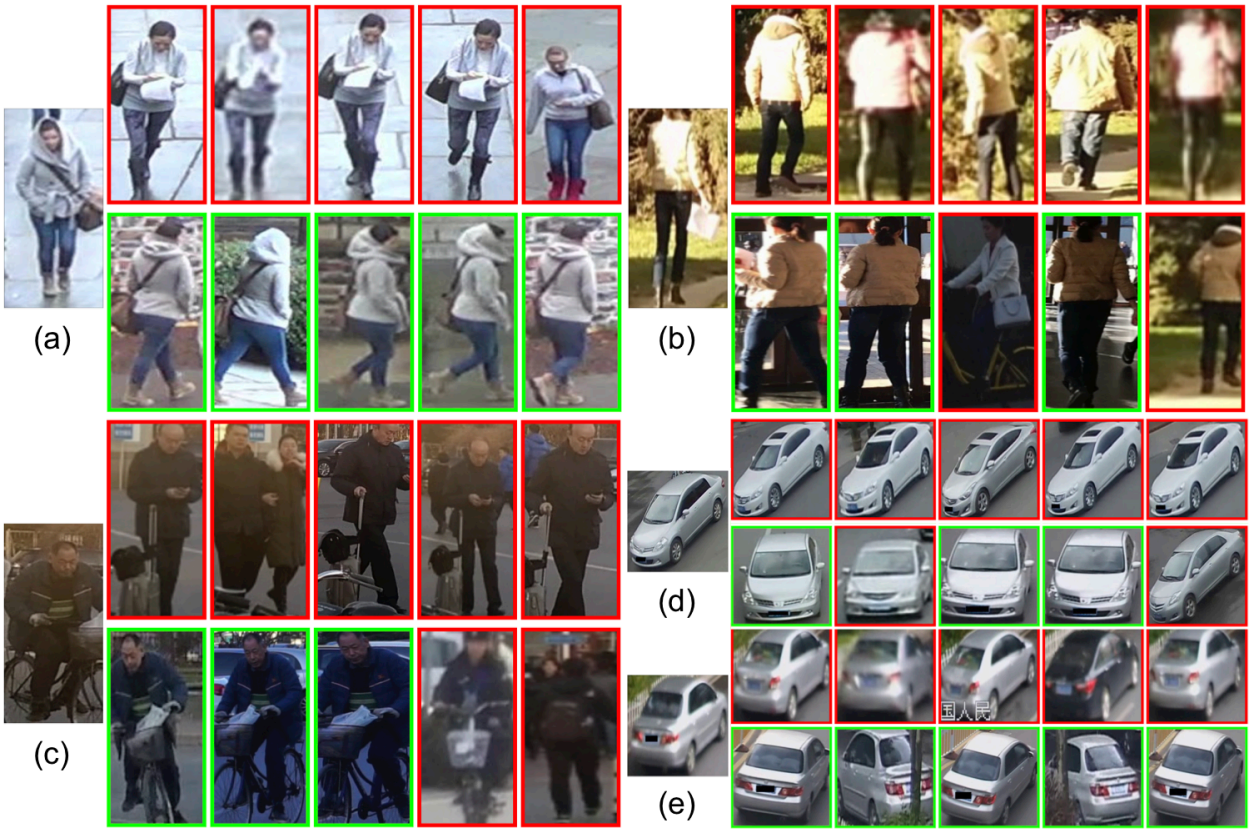

Figure 6.3: Comparison of Retrieval Results w/o ADIN. Retrieval results on the DukeMTMC-ReID (a), MSMT17 (b, c), and VeRi-776 (d, e). The leftmost image in each panel is a query image. For the five columns in each panel, the top row shows top-5 retrieval results using a vanilla ResNet50 model [184], and the bottom row shows top-5 results using ResNet50 adopted with ADIN framework. Green boxes mark the correct matches, while red boxes denote the wrong matches. The vanilla ResNet50 tend to retrieve images with similar nuisances (illumination, viewpoints, scene backgrounds, etc.), while ADIN successfully eliminates nuisances and leads to more robust matching under drastic visual appearance changes.

6.7 Conclusion

This project proposes the adversarial domain-invariant (**ADIN**) learning framework, which remarkably improves the generalizability of ReID models and resolves the nuisances-overfitting problem. Free annotations like video timestamp and camera index are for the first time utilized. In extensive experiments, measured by the new **direct transfer** performance criterion, ADIN exhibits impressive generalization to unseen datasets without any fine-tuning or re-trainig. In sum,

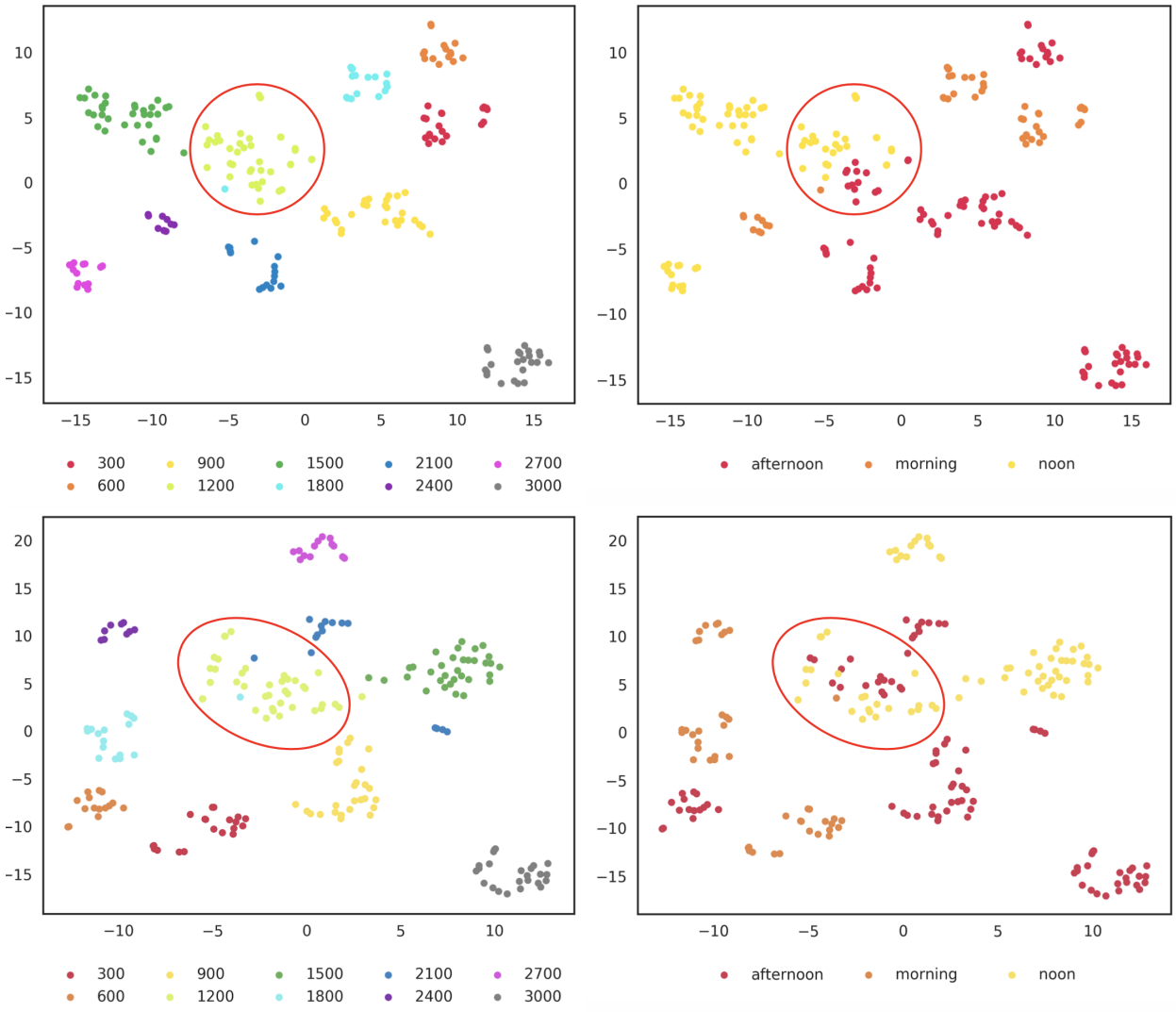

Figure 6.4: T-SNE Visualization of Representation learned via ADIN Loss. Before (top row) and after (bottom row) using adversarial disentanglement. Randomly selected ten identity classes number from MSMT17 and video time-stamps are listed below the graphs.

ADIN proves to make a substantial improvement in overcoming the generalizability challenge: the ADIN feature extractor learned on one dataset is directly generalizable to others, without seeing or adapting on any new data. To our best knowledge, ADIN is the first CNN-based ReID model that can establish strong direct transfer performance. It sets up state-of-the-art generalizability for ReID, which we believe is valuable for pursuing real-world large-scale ReID applications.

7. MESH RECOVERY FOR VIDEO-BASED REID

7.1 Motivation

As is introduced in Section 1, current ReID model suffers from numerous variations. The adversarial framework we proposed in Section 6 eliminates the environmental nuisances from the learned ReID representations. The remaining challenge for ReID lies in how to correctly match two images under appearance changes (*e.g.*, posture variations), as well as dynamic environment nuisances (*e.g.*, blur, occlusion and color distortion).

Recently, several pose-guided approaches are introduced to address the above issues: they integrate human pose/keypoints estimations to align body part regions in the pedestrian images so as to learn a structured body region features. Those algorithms [55, 56, 57, 58, 59] have been proved to be effective in extracting features from the well-organized and aligned body regions while being invariant to the entire background regions, thus reduce the negative impact of posture variations and background nuisances. However, it can not fully disentangle posture from the learned representation for the retrieval.

On the other hand, most of image-based ReID approaches learn features of the subject solely from the RGB image, which contains not only the subject-of-interest but also the background clutters. Some work proposed to utilize binary segmentation masks to accurately remove the background in the images, so as to learn a representation that only contains information of the subject body. The binary segmentation mask can also guide the algorithm to focus on the subject shape, so as to provide more information beyond appearance. However, the segmentation mask can roughly represent the silhouette of the subject which is sensitive to occlusion, deformation, and noisy boundaries. Besides, similar to the RGB image, 2D segmentation masks are lacking depth information for fully recover the 3D object and may fail in multi-viewpoint ReID systems. Therefore, the segmentation masks are not ideal to capture the body shape.

Moreover, one fundamental challenge for image-based ReID is that it learns color and generic

epitome features of appearance that are shown to be susceptible to image degradation and artifacts. Therefore, an ideal solution to capture "nude shape" of the body for ReID applications needs to (1) be able to disentangle the posture, body shape, and background; (2) be simplified so as to be spatio-temporally consistent and insensitive to appearance change; and (3) contains depth information for full recovery in any view-points. The deformable mesh recovery satisfies all the above requirements. The entire mesh assumes a rigid template for the subject and is fully determined by tens of intrinsic parameters: camera-related parameters (e.g. focal length, camera center, camera rotation, camera translation), pose and shape parameters. Those low-dimensional parameters are disentangled and subject-wise distinguishable, and can be directly used as features for ReID.

Numerous works [201, 202, 166] are proposed to accurately extract the human pose and segmentation mask from the image. Those estimated pose coordinates and binary masks do not contain any information regarding the background or the subject appearance, but capture sufficient information of body pose and shape for mesh recovery. Therefore, in this project, we assume the correctness of the offline 2D pose detector, rely on its prediction as pseudo ground truth, and take the **2D pose estimation as input** to reconstruct the 3D mesh.

The goal of 3D mesh recovery is to wrap the mesh template to best fit and represent the input image. However, it is infeasible to directly match a mesh to an RGB image. Instead, we can use the reprojection error for evaluation and optimization purposes, *i.e.*, to minimize the weighted robust distance between the reprojection of corresponding keypoints on the mesh surface and the 2D pose pseudo ground truth. More formally:

$$L_{reproj} = \sum_{i}^{P} ||v_i(x_i - \hat{x_i})||_2. \tag{7.1}$$

Here $x_i \in \mathbb{R}^2$ is the i^{th} ground truth 2D joints coordinates and $v_i \in [0, 1]$ is the confidence score provided in pose detection for each of the P joints in 2D pose.

One drawback of reconstructing mesh for ReID is that the recovery of 3D mesh from a 2D image can be ambiguous. Fortunately, temporal consistency in the video can amend the 3D esti-

mation and make up for the missing depth information in mesh recovery [203]. In this section, we propose an optimization-based video mesh recovery pipeline with 3D supervision and temporal consistency, to learn accurate 3D meshes and extract shape features for ReID representation.

7.2 Unified 3D Human Mesh

To better represent the mesh and extract features for the ReID task, we assume a rigid structure of humans and exploit the deformable unified human body model SMPL eXpressive (SMPL-X) [156] for human mesh recovery.

SMPL-X is a body model that combines SMPL body model (representing the trunk) with the FLAME head model and the MANO hand model. It is registered to 5,586 3D scans to capture the natural pose-dependent correlations between the shape of bodies, faces, and hands. SMPL-X works by factoring the human bodies into shape (individuals variations in height, weight, body proportions) and pose (articulation that deforms the 3D surface).

The shape $\beta \in \mathbb{R}^{10}$ is parameterized by the first 10 coefficients of a PCA shape space along with $\psi \in \mathbb{R}^{10}$ for facial expressions. The pose $\theta \in \mathbb{R}^{3K}$ is modeled by a relative 3D rotation of $K = 54$ joints in axis-angle representation via forwarding kinematics, including joints for the neck, jaw, eyeballs, fingers, and additional one indicating the global rotation. In all, the SMPL-X model has $N = 10,475$ vertices with a differentiable vertex-based linear blend skinning function $W(.)$ that can export a triangulated mesh with the corrective blend shapes and pose. The reconstructed mesh M can be formulated as:

$$M(\beta, \theta, \psi) = W(T_p(\beta, \theta, \psi), J(\beta), \theta, \mathcal{W}), \tag{7.2}$$

where $T_P(.)$ represents sampled vertices on the surface; $J(.)$ severs as a 3D joint sparse linear selector that regresses joint locations from mesh vertices and those joints later works as anchors to rotates the vertices; the rotation is smoothed by blend weights $\mathcal{W} \in \mathbb{R}^{N \times K}$.

Given a template $\bar{T}$, the T_P can be computed via

$$T_P(\beta, \theta, \psi) = \bar{T} + B_S(\beta; \mathcal{S}) + B_E(\psi; \mathcal{E}) + B_P(\theta; \mathcal{P}), \tag{7.3}$$

where $\mathcal{S} \in R^{3N \times |\beta|}$ in the second term is orthonormal principle components of vertex displacements capturing shape variations and the $B_S(.) : R^{|\beta|} \rightarrow R^{3N}$ is the shape blend function; Similarly, $B_E(.)$ blends the mesh with facial expression and B_P is the pose blend function, which adds corrective pose-dependent vertex displacements to the final mesh.

A weak-perspective camera model is further employed to reproject the reconstructed mesh back to the 2D image by solving a translation factor $t \in \mathbb{R}^2$ and a scale factor $s \in \mathbb{R}$. The model can also outputs $X(\theta, \beta) \in \mathbb{R}^{3P}$ 3D keypoints and the corresponding 2D pose reprojection with a linear regression from the mesh vertices, where P is the number of joints that varies from pose template (*e.g.*, $P = 17$ for COCO pose template). With an orthographic projection function Π the 2D pose reprojection $\hat{x}$ can be expressed as:

$$\hat{x} = \mathrm{s}\Pi(M(\beta, \theta, \psi)) + t. \tag{7.4}$$

7.3 Optimization-based Human Mesh Recovery

Reprojection loss along can not guarantee a reasonable mesh recovery due to the ambiguity in reconstructing 3D from 2D. The intrinsic pose/shape restriction of SMPL-X model is not sufficient to avoid local minima during the optimization which leads to some kinematically impracticable posture and abnormal shape. In order to enforce a feasible body pose and shape in the mesh model, we take inspiration from previous works [156] to employ a pose prior and shape prior to adding more constrains for mesh recovery. Here we adopted a pretrained pose prior, VPoser, as a regularization in mesh optimization. The VPoser is a variational autoencoder that learns a latent representation of human pose on a large human motion dataset and regularizes the distribution of

the latent code to be a normal distribution. The Vpose prior regularization can be formulated as:

$$L_{prior} = ||\mathcal{V}(\theta)||_2^2, \tag{7.5}$$

where $\mathcal{V}(.)$ indicates the encoder to map pose parameter θ to the latent embedding.

Learning a mesh representation from 2D pose can be deemed as estimate the 3D coordinates of all the $N = 10,475$ vertices on the mesh surface corresponding to the 2D coordinates of $P = 17$ keypoints on the pose. Directly optimizing all $\beta, \theta, \psi, \mathrm{s}, \mathrm{t}$ in one step is an extremely challenging task, especially when all these parameters contribute to the overall reprojection. The imprecise estimation of one factor can be remedied by other factors, results in accumulated errors in the optimization process. To reduce the burden on the optimizer and make progressive changes recurrently to the current estimate, we introduce a divide and conquer pipeline. Inspired by [157], we simplified the camera parameter estimation to mapping the four anchor points on T-pose (shoulders and hips) to corresponding points on 2d pose ground truth. We can then optimize the pose and shape priors separately and alternately. The overall pipeline is shown in in Fig.7.1.

7.4 Mesh Recovery with 3D Pose Supervision

One big challenge of 3D mesh recovery is lacking sufficient mesh ground truth to capture the pose and shape variations. Mesh collection usually requires special 3D laser scanners and high-speed motion capture system to capture the entire surface of a person. Even so, the mesh dataset annotations are usually "pseudo-label" obtained by fitting the mesh template to 3D scans [204]. On the other hand, 3D pose estimation has been studied extensively in recent years and a large amount of 3D pose annotations have already been collected. Given that the pose parameters on mesh models are indeed the axis-angle rotation of those articulations, we propose to estimate the pose parameters from 3D pose coordinators and ensemble 3D pose supervision to mesh recovery.

To eliminate the effect of camera parameters and body shape parameters when estimating the pose parameter from 3D pose coordinators, we build a kinetic tree to represent the correlation between pose keypoints. Considering the connection between adjacent keypoints on the kinetic

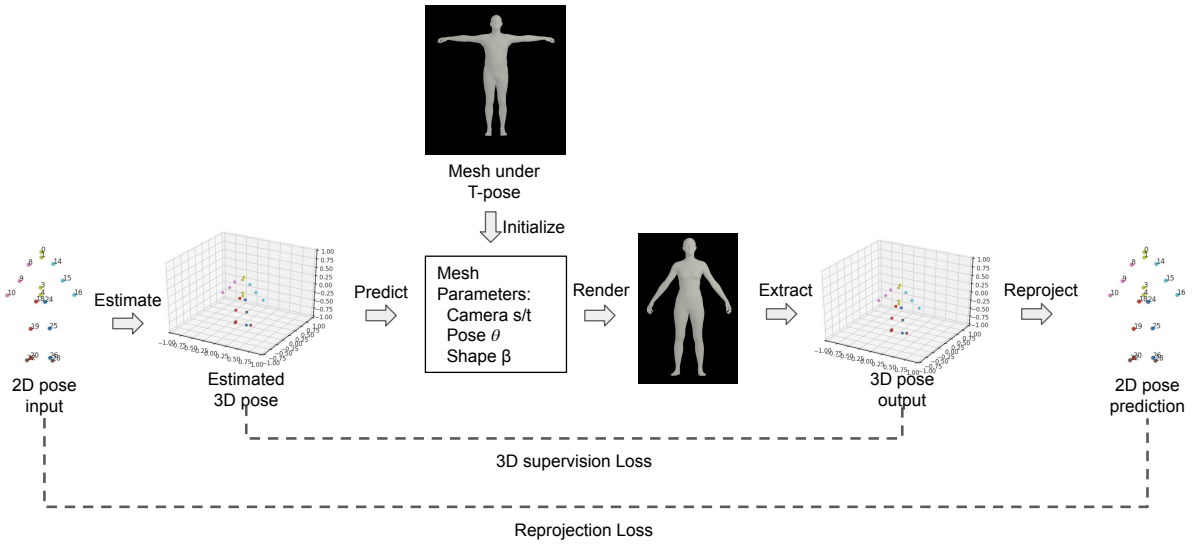

Figure 7.1: Overview of the Mesh Recovery Pipeline.

tree as a "rigid bone", we can now derive the direction from current keypoints to its kinetic child and build a vectorized 3D kinetic tree. It can be further normalized to reduce the impact of bone length, resulting in a normalized 3D kinetic tree (formulized as below), which is purely composed of human pose information and insensitive to any camera/shape changes.

$$d(i) = \frac{c_{i+1} - c_i}{||c_{i+1} - c_i||_2^2}, \tag{7.6}$$

where c_i is the 3D coordinates of the i^{th} keypoint.

In this dissertation, we propose to use the normalized 3D kinetic tree as a weak supervision of mesh recovery optimization by minimizing the difference between the rotation of the "rigid bone" on the kinetic tree and the rotation of corresponding connection from the predicted mesh.

$$L_{3d} = 1 - cos(d(i), \hat{d(i)}), \tag{7.7}$$

where the $d()$ is the kinetic direction from the i^{th} keypoint to its child keypoint on the kinetic tree.

7.5 Video-Based Mesh Recovery with Temporal Consistency

Optimization-based mesh recovery can output a rough mesh for a single frame. To further ensure the consistency and robustness the reconstructed mesh over appearance change in the long-period video, we propose a temporal consistency regularization for video-based mesh recovery (VMR). Given a video with normal framerate, the keypoints movement between adjacent frames should be in a small range, especially for keypoints lays near the root of the kinetic tree. To enforce a temporal consistency, we introduced a temporal regularization to jointly optimize the pose parameter in mesh models by minimizing its change between adjacent frames weighted by its position on the kinetic tree. Given a short clip with n frames, $\theta_{k,i}$ is the i^{th} pose parameters at the k^{th} frame and the p_i is the joint weight based on its position on the kinetic tree, the temporal consistency loss can be formulated as

$$L_{temporal} = \sum_{k=2}^{n} \sum_{i=1}^{K} p_k ||\theta_{k,i} - \theta_{k-1,i}||_2^2. \tag{7.8}$$

The impact of temporal consistency highly depends on previous frames, thus the initialization of the first frame becomes crucial. In order to start the optimization with a good initialization, we select a teaser frame in the video whose 2D pose is most similar to T-pose (the pose parameter θ of T-pose is defined to be all zeros). We start the optimization with the adjacent frames of the teaser frame and then moving the sliding window of clips for temporal consistency until all frames in the video have been covered. Temporal regularization with teaser initialization not only boost the mesh recovery performance, but also greatly accelerate optimization progress.

To summarize, the loss for reconstructing the 3D mesh from a 2D image is formulated as:

$$L_{total} = L_{reproject} + L_{prior} + L_{3d} + L_{temporal}. \tag{7.9}$$

Algorithm 3 The optimization strategy of video mesh recovery

Select a teaser frame i in the video whose 2D pose is most similar to T-pose
$\theta, \beta \leftarrow 0$, optimize camera parameters s, t with L_{total} of anchor points only
while not the last frame **do** ▷ Moving the sliding window for temporal assistency forward
  Initialize $i, i+1, ..., i+k$ with the parameters of i
  Fixed θ and β, and optimize s, t for $i, i+1, ..., i+k$ frames with L_{total} of anchor points only
  Fixed β, s, t and optimize θ for $i, i+1, ..., i+k$ frames with L_{total}
  Fixed θ, s, t and optimize β for $i, i+1, ..., i+k$ frames with L_{total}
  $i \leftarrow i+k$
end while
Set the pointer i to the teaser frame index
while not the first frame **do** ▷ Moving the sliding window for temporal assistency backward
  Temporally reverse $i, i-1, ..., i-k$ frames
  Initialize $i, i-1, ..., i-k$ with the parameters of i
  Fixed θ and β, and optimize s, t for $i, i-1, ..., i-k$ frames with L_{total} of anchor points only
  Fixed β, s, t and optimize θ for $i, i-1, ..., i-k$ frames with L_{total}
  Fixed θ, s, t and optimize β for $i, i-1, ..., i-k$ frames with L_{total}
  Reverse back $i, i-1, ..., i-k$ frames
  $i \leftarrow i-k$
end while

7.6 Implementation Details of Video Mesh Recovery

The optimization-based approach does not require a large amount of training data. The preliminary results in this dissertation are conducted on videos downloaded from the internet. We utilized the official code of OpenPose [201] for 2D keypoint detection and VideoPose3D [203] for 3D pose estimation in video from 2D keypoint trajectories.

In practice, we implement the optimization using an iterative strategy. We initialize the model with the teaser frame and set the pose parameters θ to be all zeros, and then fixed pose parameters θ and shape parameters β while optimizing camera parameters s, t by fitting the four anchor points on T-pose (shoulders and hips) to corresponding points on 2d pose ground truth. Afterward, we alternate between optimizing pose parameters θ and shape parameters β until converge. The overall training procedure is shown in Algorithm 3.

7.7 Rendering of Proposed Mesh Recovery Approach

The recovered mesh are rendered and evaluated via visual check. Without 3D pose supervision and temporal regularization, the image-based mesh recovery is well-performed on reprojection examination but fails in multi-view examination due to ambiguity. Fig. 7.2 shows the mesh recovery results on the first ten frames from the video, in which the actor is still but motion blur exists. The recovered mesh are expected to be similar among these frames, but the prediction are imprecise and noisy. We could observe that a slight change from 2D pose input can result in a large fluctuation on the recovered 3D mesh. Our proposed 3D supervision and temporal regularization can overcome the depth ambiguity issue. As is shown in Fig. 7.3, the predictions tends to consistent.

We also compare the rendering on ten continuous frames (Fig. 7.4), in which the actor is dancing and posture change could be clearly observed. The 3D supervision and temporal regularization and improve the stability of mesh recovery. Besides, we compare the rendering of proposed approach to previous state-of-the-art mesh recovery methods on ten randomly selected frames (the optimization is process on the entire video, while only ten frames are displayed here), and obviously our video-based approach outperforms all the other methods (Fig. 7.5).

The preliminary results do not include quantitative evaluations, which is supposed to be analyzed in future work. We plan to evaluate the proposed approach on Human3.6M [205, 206] and AMASS dataset [207] with (1) reconstruction error with ground truth 3D mesh annotations, (2) on the standard 3D joint estimation task, and (3) an auxiliary task of body part segmentation.

7.8 Conclusion and Future Work

In this section, we discuss the possibility to utilize human mesh recovery for ReID representation learning. To be more specific, we proposed to reconstruct 3D mesh from 2D images and ensemble the pose/shape parameters in the mesh model for ReID retrieval. Given the mesh deformable model can disentangle pose, shape, and camera factor, and is insensitive to appearance change, we believe that the body parameters could be a auxiliary feature for ReID tasks.

Figure 7.2: [Renderings of Image-Based Mesh Recovery on the First Ten Frames without 3D Supervision and Temporal Regularization. (Top) The 2D pose input; (Middle) Overlay of mesh rendering reprojection on original image; (Bottom): A different view of mesh rendering.

Figure 7.3: Renderings of Video-Based Mesh Recovery on the First Ten Frames with 3D Supervision and Temporal Regularization. (Top) The 2D pose input; (Middle) Overlay of mesh rendering reprojection on original image; (Bottom) A different view of mesh rendering.

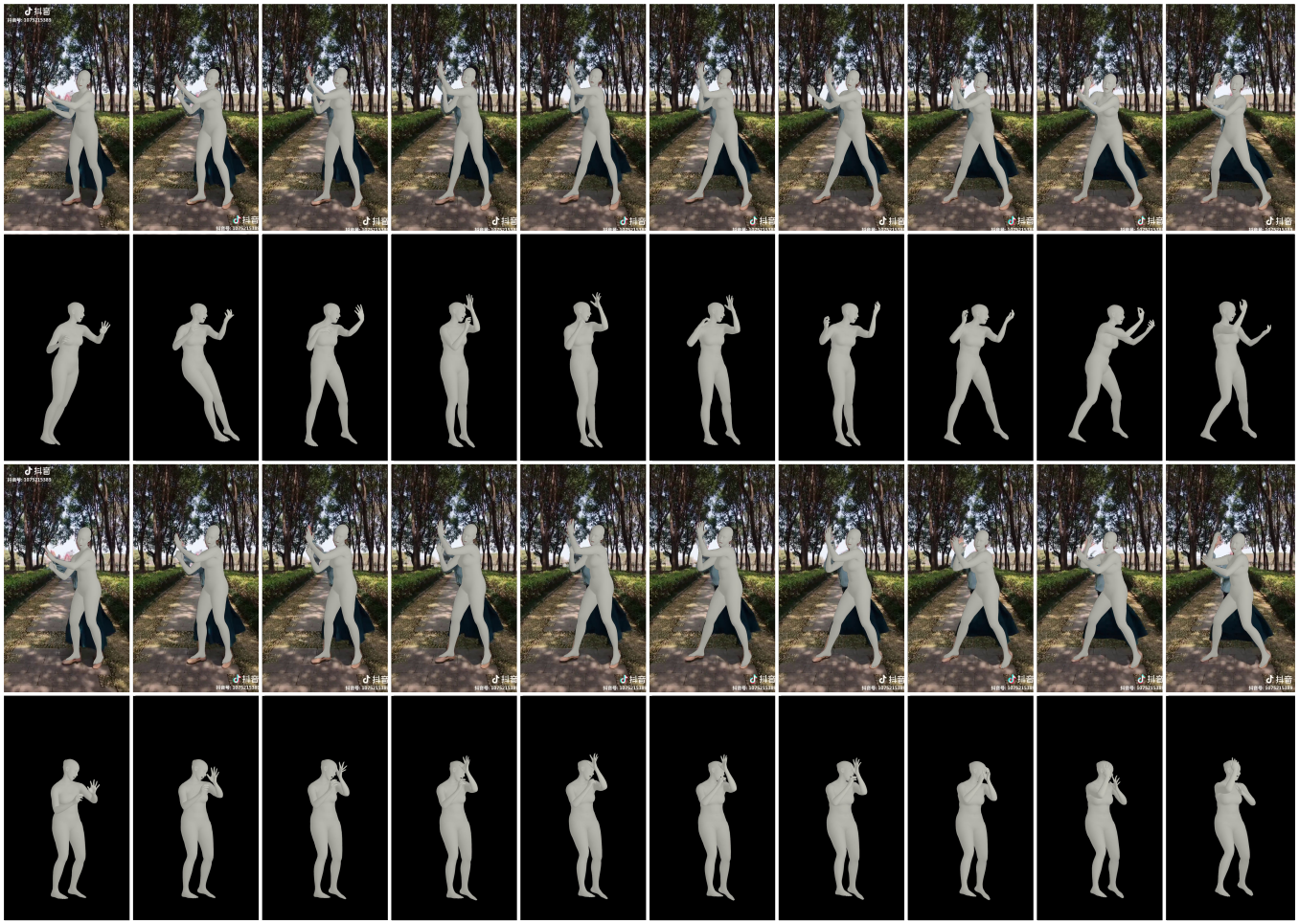

Figure 7.4: Image-Based *v.s.* Video-Based Mesh Recovery on Ten Continuous Frames. (First row) Overlay of mesh rendering re-projection from image-based mesh recovery on original image; (Second row) A different view of mesh rendering from image-based mesh recovery. (Third row) Overlay of mesh rendering reprojection from video-based mesh recovery on original image; (Forth row) A different view of mesh rendering from video-based mesh recovery.

Figure 7.5: Video-Based Mesh Recovery *v.s.* Previous Approaches on Ten Randomly Selected Frames. (First row) The 2D pose input; (Second row) Overlay of mesh rendering reprojection from HMR [165] on original image; (Third row) Overlay of mesh rendering from monocular total capture [208] reprojection on original image; (Forth row) Overlay of the rendering reprojection from video-based mesh recovery on original image;

8. SUMMARY AND FUTURE WORK

8.1 Summary

The research topics discussed throughout this dissertation share the same underlying theme: developing a generalizable and transferable ReID approach to bridge the algorithm in research and the real-world application.

My research started with an interesting question, how to bridge the low-level enhancement and high-level visual task models, to be more specific, how to remove degradations and improve the detection performance to extract precise bounding boxes and annotate them with correct labels. To address this problem, we proposed the Cascaded Degradation Removal Modules (CDRM) for image enhancement and reveal that image enhancement can indeed benefit detection qualitatively and quantitatively. The dissertation also introduced a new benchmark to better analyze the gap/alignment between image restoration and high-level visual tasks.

Degradation variation is only one of the gaps between ReID in research and life. To overcome the challenge caused by the low data volume, variation coverage, and spatio-temporal imbalance, my research was focused on effective, efficient, robust, and generalizable feature learning models, that emphasize feature selection and disentanglement in ReID application. The fast-approximate-triplet loss with label distillation and the adversarial domain-invariant network were presented along this thread, to reveal how to enforce the intra-class feature similarity and eliminate the inter-class shared nuisances.

In addition to deriving more novel ReID feature learning from appearance, I also had the ambition to introduce the body model coefficients in mesh recovery as an auxiliary feature for ReID recognition. Reconstructed mesh via a deformable parametric model can maximally disentangle between the retrieval-unnecessary feature, identity consistently shared parameters, and background factors, which brought new insight to the ReID community.

8.2 Conclusion

In this dissertation, several interesting observations could be concluded:

- The challenging visual conditions usually give rise to nonlinear and data-dependent degradation during data acquisition. Those degradation can severely impact not only the visual quality of images collected in the real-world situation but also the quality of feature extracted during the subsequent ReID process. Image restoration and enhancement algorithms that simultaneously handle multiple degradation tended to be beneficial to both of these objectives on the diverse imagery such as UG^2. On the other hand, we could also observe that separate consideration of enhancement and detection might lead to a deteriorated performance in detection as is shown on UG^{2+} dataset. How to bridge the gaps between visual quality and high-level vision tasks is still a very difficult, under-explored, yet highly meaningful class of computer vision problems in practice.

- The major gap between the research efforts and the practical needs in large-scale deployment of ReID lies in the insufficient data volume, low variation coverage, the heavily *non-i.i.d* spatio-temporal distribution, unpredictable noise and outliers, and the abundance of those nuisances in real-world scenarios. Those gaps are detrimental to the generalizability of ReID models to unseen identities. Comparative losses such as triplet loss and FAT loss can be remarkably effective in guiding the feature extractor to minimize the distances of intra-class features, while domain-invariant framework further improves the robustness by disentangling subject-irrelevant nuisances. In extensive experiments, FAT loss and ADIN framework exhibit both state-of-the-art performance on popular benchmarks and impressive generalization to unseen datasets without needing additional domain adaption.

- The standard image-based ReID method learns texture, appearance, and illumination from the subject-of-interest that are shown to be fragile to image degradation, artifacts, and appearance change. A promising direction is to learn an intrinsic representation of the subject that is insensitive to both subject behavior-related factors such as pose or viewpoint vari-

ations, and the environmental factor such as illumination changes. The deformable mesh is a perfect solution that can disentangle the posture and body shape from the background, and meanwhile it is spatio-temporally consistent and insensitive to any appearance change. Therefore, it is demanding to further explore how to ensemble mesh recovery and motion capture into the ReID representations learning to tackle the robust ReID problems.

8.3 Future Work

Surveillance systems combined with Deep Learning and AI approaches are becoming more and more prevalent, and they are playing a key role in smart cities and contribute to public safety. Yet for accurate predictions, those AI models often hinge on storing and analyzing users' private data, such as name, gender, address, or personal videos/images. The abuse or misuse of surveillance data has reinvigorated the privacy and fairness debate. It is thus also appealing to explore: on the premise of learning a robust and generalizable representation for ReID application, how to apply learning techniques to protect the privacy in the data, and ensure fairness for the users.

This question suggests a dilemma in ReID applications that we would like to develop an approach to recognize the subject-of-interest while preventing it from extracting sensitive information to preserve privacy and fairness. Classical solutions secure user privacy by adding a mosaic to the sensitive region on the image *e.g.* face or car template. However, they might erase the characteristic patterns of the natural images and consequently lead to the failure of the ReID application.

For future direction, I would be interested in studying the anonymization of user data: how to learn an appropriate auto-transform on the collected raw visual data from the local camera end, so that the cryptographic data itself will only enable the ReID task while obstructing other undesired privacy or fairness related tasks.

REFERENCES

[1] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proceedings of the European Conference on Computer Vision*, 2008.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[3] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2013.

[4] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2014.

[5] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proceedings of the European Conference on Computer Vision*, 2014.

[6] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *Proceedings of the IEEE International Conference on Computer Vision*, December 2015.

[7] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.

[8] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[9] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proceedings of the European Conference on Computer Vision*, 2016.

[10] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, pp. 13:1–13:20, Dec. 2017.

[11] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for pedestrian retrieval," in *Proceedings of the 2017 ACM on Multimedia Conference*, pp. 420–428, 2017.

[12] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[13] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[14] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[15] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision*, September 2018.

[16] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[17] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[18] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[19] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[20] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2013.

[21] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proceedings of the European Conference on Computer Vision*, September 2014.

[22] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[23] C. Jose and F. Fleuret, "Scalable metric learning via weighted approximate rank component analysis," in *Proceedings of the European Conference on Computer Vision*, September 2016.

[24] T. M Feroz Ali and S. Chaudhuri, "Maximum margin metric learning over discriminative nullspace for person re-identification," in *Proceedings of the European Conference on Computer Vision*, September 2018.

[25] P. Chen, X. Xu, and C. Deng, "Deep view-aware metric learning for person re-identification," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 620–626, July 2018.

[26] "Number of video surveillance cameras per thousand people in 2014, by country," *https://www.statista.com/statistics/484956/number-of-surveillance-cameras-per-thousand-people-by-country*, 2015.

[27] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, December 2015.

[28] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proceedings of the European Conference on Computer Vision*, pp. 17–35, Springer, 2016.

[29] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017.

[30] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[31] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[32] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2167–2175, 2016.

[33] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proceedings of the European Conference on Computer Vision*, pp. 869–884, Springer, 2016.

[34] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[35] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. Change Loy, "The devil of face recognition is in the noise," in *Proceedings of the European Conference on Computer Vision*, September 2018.

[36] *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018 UG^2 prize challenge, http://cvpr2020.ug2challenge.org/program18/challenges18.html*.

[37] W. Scheirer, R. VidalMata, S. Banerjee, B. RichardWebster, M. Albright, P. Davalos, S. McCloskey, B. Miller, A. Tambo, S. Ghosh, S. Nagesh, Y. Yuan, Y. Hu, J. Wu, W. Yang, X. Zhang, J. Liu, Z. Wang, H. Chen, T. Huang, W. Chin, Y. Li, M. Lababidi, and C. Otto, "Bridging the gap between computational photography and visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[38] *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019 UG^{2+} prize challenge, http://cvpr2020.ug2challenge.org/program19/challenges19.html*.

[39] *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020 UG^{2+} prize challenge, http://cvpr2020.ug2challenge.org*.

[40] W. Yang, Y. Yuan, W. Ren, J. Liu, W. J. Scheirer, Z. Wang, T. Zhang, Q. Zhong, D. Xie, S. Pu, Y. Zheng, Y. Qu, Y. Xie, L. Chen, Z. Li, C. Hong, H. Jiang, S. Yang, Y. Liu, X. Qu, P. Wan, S. Zheng, M. Zhong, T. Su, L. He, Y. Guo, Y. Zhao, Z. Zhu, J. Liang, J. Wang, T. Chen, Y. Quan, Y. Xu, B. Liu, X. Liu, Q. Sun, T. Lin, X. Li, F. Lu, L. Gu, S. Zhou, C. Cao, S. Zhang, C. Chi, C. Zhuang, Z. Lei, S. Z. Li, S. Wang, R. Liu, D. Yi, Z. Zuo, J. Chi, H. Wang, K. Wang, Y. Liu, X. Gao, Z. Chen, C. Guo, Y. Li, H. Zhong, J. Huang, H. Guo, J. Yang, W. Liao, J. Yang, L. Zhou, M. Feng, and L. Qin, "Advancing image understanding in poor visibility environments: A collective benchmark study," *IEEE Transactions on Image Processing*, vol. 29, pp. 5737–5752, 2020.

[41] Y. Yuan, W. Chen, Y. Yang, and Z. Wang, "In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, June 2020.

[42] Y. Yuan, W. Chen, T. Chen, Y. Yang, Z. Ren, Z. Wang, and G. Hua, "Calibrated domain-invariant learning for highly generalizable large scale re-identification," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2019.

[43] H. Wang, Y. Fan, Z. Wang, L. Jiao, and B. Schiele, "Parameter-free spatial attention network for person re-identification," *arXiv preprint arXiv:1811.12150*, 2018.

[44] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision*, September 2018.

[45] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the ACM international conference on Multimedia*, pp. 274–282, 2018.

[46] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *Proceedings of the European Conference on Computer Vision*, September 2018.

[47] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[48] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.

[49] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[50] B. Yu, T. Liu, M. Gong, C. Ding, and D. Tao, "Correcting the triplet selection bias for triplet loss," in *Proceedings of the European Conference on Computer Vision*, September 2018.

[51] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[52] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[53] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai, "Hard-aware point-to-set deep metric for person re-identification," in *Proceedings of the European Conference on Computer Vision*, September 2018.

[54] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3539–3548, 2017.

[55] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3960–3969, 2017.

[56] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision*, pp. 402–419, 2018.

[57] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 542–551, 2019.

[58] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4099–4108, 2018.

[59] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *Advances in neural information processing systems*, pp. 1222–1233, 2018.

[60] L. Qi, J. Huo, L. Wang, Y. Shi, and Y. Gao, "A mask based deep ranking neural network for person retrieval," in *IEEE International Conference on Multimedia and Expo)*, pp. 496–501, 2019.

[61] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1179–1188, 2018.

[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

[63] Z. Wang, H. Li, Q. Ling, and W. Li, "Robust temporal-spatial decomposition and its applications in video processing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 3, pp. 387–400, 2013.

[64] H. Li, Z. Lu, Z. Wang, Q. Ling, and W. Li, "Detection of blotch and scratch in video based on video decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1887–1900, 2013.

[65] J. Ren, J. Liu, and Z. Guo, "Context-aware sparse decomposition for image denoising and super-resolution," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1456–1469, 2013.

[66] Z. Wang, D. Liu, S. Chang, Q. Ling, Y. Yang, and T. S. Huang, "D3: Deep dual-domain based fast restoration of jpeg-compressed images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2764–2772, 2016.

[67] J. Liu, S. Yang, Y. Fang, and Z. Guo, "Structure-guided image inpainting using homography transformation," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3252–3265, 2018.

[68] Y. Yan, W. Ren, Y. Guo, R. Wang, and X. Cao, "Image deblurring via extreme channels prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4003–4011, 2017.

[69] W. Ren, J. Pan, X. Cao, and M.-H. Yang, "Video deblurring via semantic segmentation and pixel-wise non-linear kernel," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1077–1085, 2017.

[70] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8183–8192, 2018.

[71] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8878–8887, 2019.

[72] J. Wu, X. Yu, D. Liu, M. Chandraker, and Z. Wang, "David: Dual-attentional video deblurring," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 2376–2385, 2020.

[73] Z. Wang, Y. Yang, Z. Wang, S. Chang, J. Yang, and T. S. Huang, "Learning super-resolution jointly from external and internal examples," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4359–4371, 2015.

[74] Z. Wang, Y. Yang, Z. Wang, S. Chang, W. Han, J. Yang, and T. Huang, "Self-tuned deep super resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8, 2015.

[75] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, and T. Huang, "Robust video super-resolution with learned temporal dynamics," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2507–2515, 2017.

[76] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, X. Wang, and T. S. Huang, "Learning temporal dynamics for video super-resolution: A deep learning approach," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3432–3445, 2018.

[77] Z. Yu, H. Li, Z. Wang, Z. Hu, and C. W. Chen, "Multi-level video frame interpolation: Exploiting the interaction among different levels," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1235–1248, 2013.

[78] E. D. Pisano, S. Zong, B. M. Hemminger, M. DeLuca, R. E. Johnston, K. Muller, M. P. Braeuning, and S. M. Pizer, "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms," *Journal of Digital imaging*, vol. 11, no. 4, p. 193, 1998.

[79] W. Yang, J. Feng, J. Yang, F. Zhao, J. Liu, Z. Guo, and S. Yan, "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5895–5907, 2017.

[80] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections," *arXiv preprint arXiv:1606.08921*, 2016.

[81] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–12, 2017.

[82] Y. Zhang, L. Ding, and G. Sharma, "Hazerd: an outdoor scene dataset and benchmark for single image dehazing," in *IEEE international conference on image processing*, pp. 3205–3209, 2017.

[83] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-HAZE: a dehazing benchmark with real hazy and haze-free outdoor images," *arXiv e-prints*, p. arXiv:1804.05101, April 2018.

[84] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images," *arXiv e-prints*, p. arXiv:1804.05091, April 2018.

[85] M. Grgic, K. Delac, and S. Grgic, "Scface - surveillance cameras face database," *Multimedia Tools and Applications*, vol. 51, pp. 863–879, Feb. 2011.

[86] J. Shao, C. C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2227–2234, June 2014.

[87] X. Zhu, C. C. Loy, and S. Gong, "Video synopsis by heterogeneous multi-source correlation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 81–88, Dec 2013.

[88] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, "A large-scale benchmark dataset for event recognition in surveillance video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3153–3160, June 2011.

[89] M. Mueller, N. Smith, and B. Ghanem, " A Benchmark and Simulator for UAV Tracking," in *Proceedings of the European Conference on Computer Vision*, pp. 445–461, Springer, 2016.

[90] B. Z. Yao, X. Yang, and S.-C. Zhu, "Introduction to a large-scale general purpose ground truth database: Methodology, annotation tool and benchmarks," in *Proceedings of the International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2007.

[91] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505,

2019.

[92] R. Fattal, "Single image dehazing," *ACM transactions on graphics*, vol. 27, pp. 72:1–72:9, Aug. 2008.

[93] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2341–2353, Dec 2011.

[94] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Transactions on Image Processing*, vol. 24, pp. 3522–3533, Nov 2015.

[95] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1674–1682, June 2016.

[96] J. Zhou and F. Zhou, "Single image dehazing motivated by retinex theory," in *Proceedings of the International Symposium on Instrumentation and Measurement, Sensor Network and Automation*, pp. 243–247, Dec 2013.

[97] D. Nair, P. A. Kumar, and P. Sankaran, "An effective surround filter for image dehazing," in *Proceedings of the International Conference on Interdisciplinary Advances in Applied Computing*, ICONIAAC '14, (New York, NY, USA), pp. 20:1–20:6, ACM, 2014.

[98] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, pp. 5187–5198, Nov. 2016.

[99] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proceedings of the European Conference on Computer Vision*, 2016.

[100] L. Kratz and K. Nishino, "Factorizing scene albedo and depth from a single foggy image," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1701–1708, Sep. 2009.

[101] K. Nishino, L. Kratz, and S. Lombardi, "Bayesian defogging," *International Journal of Computer Vision*, vol. 98, pp. 263–278, July 2012.

[102] Y. Li, R. T. Tan, and M. S. Brown, "Nighttime haze removal with glow and multiple light colors," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 226–234, Dec 2015.

[103] J. Zhang, Y. Cao, S. Fang, Y. Kang, and C. W. Chen, "Fast haze removal for nighttime image using maximum reflectance prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7016–7024, July 2017.

[104] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4780–4788, Oct 2017.

[105] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "An all-in-one network for dehazing and beyond," *arXiv preprint arXiv:1707.06543*, 2017.

[106] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "End-to-end united video dehazing and detectionc," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Feb. 2018.

[107] W. Ren, J. Zhang, X. Xu, L. Ma, X. Cao, G. Meng, and W. Liu, "Deep video dehazing with semantic segmentation," *IEEE Transactions on Image Processing*, vol. 28, pp. 1895–1908, April 2019.

[108] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1169–1178, 2018.

[109] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *Proceedings of the European Conference on Computer Vision*, pp. 170–186, 2018.

[110] X. Lan, H. Wang, S. Gong, and X. Zhu, "Deep reinforcement learning attention selection for person re-identification," *arXiv preprint arXiv:1707.02785*, 2017.

[111] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1077–1085, 2017.

[112] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 420–429, 2018.

[113] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *arXiv preprint arXiv:1412.6596*, 2014.

[114] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 447–461, March 2016.

[115] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in Neural Information Processing Systems*, pp. 1196–1204, 2013.

[116] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," *arXiv preprint arXiv:1406.2080*, 2014.

[117] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," 2016.

[118] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[119] A. Vahdat, "Toward robustness against label noise in training deep discriminative neural networks," in *Advances in Neural Information Processing Systems*, pp. 5596–5605, 2017.

[120] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[121] S. R. Bulò, L. Porzi, and P. Kontschieder, "Dropout distillation," in *Proceedings of the International Conference on Machine Learning*, pp. 99–107, 2016.

[122] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," *arXiv preprint arXiv:1511.03643*, 2015.

[123] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[124] Z. Luo, J.-T. Hsieh, L. Jiang, J. C. Niebles, and L. Fei-Fei, "Graph distillation for action detection with privileged modalities," in *Proceedings of the European Conference on Computer Vision*, September 2018.

[125] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Proceedings of the European Conference on Computer Vision*, September 2018.

[126] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017.

[127] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[128] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[129] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[130] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[131] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[132] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[133] B. Allen, B. Curless, and Z. Popović, "The space of human body shapes: reconstruction and parameterization from range scans," *ACM transactions on graphics*, vol. 22, no. 3, pp. 587–594, 2003.

[134] B. Allen, B. Curless, Z. Popović, and A. Hertzmann, "Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis," in *Proceedings of the ACM SIGGRAPH Eurographics symposium on Computer animation*, pp. 147–156, 2006.

[135] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," in *ACM SIGGRAPH*, pp. 408–416, 2005.

[136] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel, "A statistical model of human pose and body shape," in *Computer graphics forum*, vol. 28, pp. 337–346, Wiley Online Library, 2009.

[137] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics*, vol. 34, no. 6, pp. 1–16, 2015.

[138] S. Khamis, J. Taylor, J. Shotton, C. Keskin, S. Izadi, and A. Fitzgibbon, "Learning an efficient model of hand shape variation from depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2540–2548, 2015.

[139] S. Melax, L. Keselman, and S. Orsten, "Dynamics based 3d skeletal hand tracking," in *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pp. 184–184, 2013.

[140] M. Oberweger, P. Wohlhart, and V. Lepetit, "Training a feedback loop for hand pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3316–3324, 2015.

[141] M. de La Gorce, D. J. Fleet, and N. Paragios, "Model-based 3d hand pose estimation from monocular video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1793–1805, 2011.

[142] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics*, vol. 36, no. 6, p. 245, 2017.

[143] T. Schmidt, R. A. Newcombe, and D. Fox, "Dart: Dense articulated real-time tracking.," in *Robotics: Science and Systems*, vol. 2, Berkeley, CA, 2014.

[144] S. Sridhar, A. Oulasvirta, and C. Theobalt, "Interactive markerless articulated hand motion tracking using rgb and depth data," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2456–2463, 2013.

[145] A. Tkach, M. Pauly, and A. Tagliasacchi, "Sphere-meshes for real-time hand modeling and tracking," *ACM Transactions on Graphics*, vol. 35, no. 6, pp. 1–11, 2016.

[146] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall, "Capturing hands in action using discriminative salient points and physics simulation," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 172–193, 2016.

[147] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3d face recognition with a morphable model," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1–6, 2008.

[148] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the annual conference on Computer graphics and interactive techniques*, pp. 187–194, 1999.

[149] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3d morphable models," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 233–254, 2018.

[150] A. Brunton, A. Salazar, T. Bolkart, and S. Wuhrer, "Review of statistical shape spaces for 3d data with comparative analysis for human faces," *Computer Vision and Image Understanding*, vol. 128, pp. 1–17, 2014.

[151] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 211–216, 2006.

[152] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans.," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 194–1, 2017.

[153] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 296–301, 2009.

[154] D. Vlasic, M. Brand, H. Pfister, and J. Popovic, "Face transfer with multilinear models," in *ACM SIGGRAPH*, pp. 24–es, 2006.

[155] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas, "Expression flow for 3d-aware face component transfer," in *ACM SIGGRAPH*, pp. 1–10, 2011.

[156] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019.

[157] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *Proceedings of the European Conference on Computer Vision*, pp. 561–578, Springer, 2016.

[158] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131, 2018.

[159] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6050–6059, 2017.

[160] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *Proceedings of the International Conference on 3D Vision*, pp. 484–494, 2018.

[161] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3d human pose and shape from a single color image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 459–468, 2018.

[162] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang, "Detailed human shape estimation from a single image by hierarchical mesh deformation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4491–4500, 2019.

[163] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5614–5623, 2019.

[164] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4501–4510, 2019.

[165] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[166] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

[167] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3194–3203, 2018.

[168] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proceedings of the European Conference on Computer Vision*, 2016.

[169] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[170] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[171] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

[172] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[173] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1249–1258, 2016.

[174] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," *arXiv preprint arXiv:1604.01850*, vol. 2, 2016.

[175] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *arXiv preprint arXiv:1705.04724*, 2017.

[176] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal loss for deep face recognition," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshop*, vol. 4, 2017.

[177] Z. Ming, J. Chazalon, M. M. Luqman, M. Visani, and J.-C. Burie, "Simple triplet loss based on intra/inter-class metric learning for face verification," in *Proceedings of the IEEE International Conference on Computer Vision Workshop*, pp. 1656–1664, 2017.

[178] X. Zheng, R. Ji, X. Sun, Y. Wu, F. Huang, and Y. Yang, "Centralized ranking loss with weakly supervised localization for fine-grained object retrieval.," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1226–1233, 2018.

[179] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[180] Q. Xiao, H. Luo, and C. Zhang, "Margin sample mining loss: A deep learning based method for person re-identification," *arXiv preprint arXiv:1710.00478*, 2017.

[181] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1335–1344, 2016.

[182] Y. Zhao, Z. Jin, G.-j. Qi, H. Lu, and X.-s. Hua, "An adversarial approach to hard triplet generation," in *Proceedings of the European Conference on Computer Vision*, pp. 501–517, 2018.

[183] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1470–1478, 2018.

[184] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[185] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks.," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, p. 3, 2017.

[186] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," in *Advances in Neural Information Processing Systems*, pp. 4170–4178, 2016.

[187] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person reidentification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 6, 2017.

[188] K. Li, Z. Ding, K. Li, Y. Zhang, and Y. Fu, "Support neighbor loss for person re-identification," in *Proceedings of the ACM international conference on Multimedia*, pp. 1492–1500, 2018.

[189] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," *arXiv preprint arXiv:1711.10658*, 2017.

[190] H. Li and M. Gong, "Self-paced convolutional neural networks," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2110–2116, 2017.

[191] A. Katharopoulos and F. Fleuret, "Not all samples are created equal: Deep learning with importance sampling," *arXiv preprint arXiv:1803.00942*, 2018.

[192] Z. Wu, Z. Wang, Z. Wang, and H. Jin, "Towards privacy-preserving visual recognition via adversarial training: A pilot study," in *Proceedings of the European Conference on Computer Vision*, pp. 606–624, 2018.

[193] H. Wang, Z. Wu, Z. Wang, Z. Wang, and H. Jin, "Privacy-preserving deep visual recognition: An adversarial learning framework and a new dataset," *arXiv preprint arXiv:1906.05675*, 2019.

[194] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[195] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[196] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

[197] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proceedings of the European Conference on Computer Vision*, pp. 499–515, Springer, 2016.

[198] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2016.

[199] X. Liu, S. Zhang, Q. Huang, and W. Gao, "Ram: a region-aware deep model for vehicle re-identification," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2018.

[200] J. Peng, H. Wang, and X. Fu, "Cross domain knowledge learning with dual-branch adversarial network for vehicle re-identification," *arXiv preprint arXiv:1905.00006*, 2019.

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

[202] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "Posetrack: A benchmark for human pose estimation and tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5167–5176, 2018.

[203] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7753–7762, 2019.

[204] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black, "Dynamic faust: Registering human bodies in motion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6233–6242, 2017.

[205] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[206] C. S. Catalin Ionescu, Fuxin Li, "Latent structured models for human pose estimation," in *International Conference on Computer Vision*, 2011.

[207] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5442–5451, 2019.

[208] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10965–10974, 2019.