From Robust to Generalizable Representation Learning for Person Re-Identification

Qilei Li

Submitted in partial fulfilment of the requirement for the degree of *Doctor of Philosophy*

School of Electronic Engineering and Computer Science

Queen Mary, University of London

2024

Abstract

Person Re-Identification (ReID) is a retrieval task across non-overlapping cameras. Given a person-of-interest as a query, the goal of ReID is to determine whether this person has appeared in another place at a distinct time captured by a different camera, or even the same camera at a different time instant. ReID is considered a **zero-shot** learning task because the identities present in the training data may not necessarily overlap with those in the test data within the label space. This fundamental characteristic adds a layer of complexity to the task, making ReID a highly challenging representation learning problem. This thesis solves the problem of learning generalizable yet discriminative representation with the following solutions:

Chapter 3 Noisy and unrepresentative frames in automatically generated object bounding boxes from video sequences cause significant challenges in learning discriminative representations in video ReID. Most existing methods tackle this problem by assessing the importance of video frames according to their local part alignments or global appearance correlations separately. However, given the diverse and unknown sources of noise that usually co-exist in captured video data, existing methods have not been effective satisfactorily. In this chapter, we explore jointly both local alignments and global correlations with further consideration of their mutual promotion/reinforcement so to better assemble complementary discriminative ReID information within all the relevant frames in video tracklets, and propose a model named Local-Global Associative Assembling (LOGA). Specifically, we concurrently optimize a Local Aligned Quality (LAQ) module that distinguishes the quality of each frame based on local alignments, and a Global Correlated Quality (GCQ) module that estimates global appearance correlations. With a local-assembled global appearance prototype, we associate LAQ and GCQ to exploit their mutual complement.

Chapter 4 While deep learning has significantly improved ReID model accuracy under the Independent and Identical Distribution (IID) assumption, it has also become clear that such models degrade notably when applied to an unseen novel domain due to unpredictable/unknown domain shift. Contemporary Domain Generalizable ReID models struggle in learning domain-invariant

representation solely through training on an instance classification objective. We consider that deep learning models are heavily influenced and therefore biased towards domain-specific characteristics, *e.g* background clutter, scale, and viewpoint variations, limiting the generalizability of the learned model, and hypothesize that the pedestrians are domain invariant owning they share the same structural characteristics. To enable the ReID model to be less domain-specific from these pure pedestrians, we introduce a Primary-Auxiliary Objectives Association (PAOA) model that guides model learning of the primary ReID instance classification objective by a concurrent auxiliary learning objective on weakly labeled pedestrian saliency detection. To solve the problem of conflicting optimization criteria in the model parameter space between the two learning objectives, PAOA calibrates the loss gradients of the auxiliary task towards the primary learning task gradients. Benefiting from the harmonious multitask learning design, our model can be extended with the recent test-time diagram to form the PAOA+, which performs on-the-fly optimization against the auxiliary objective in order to maximize the model's generative capacity in the test target domain. Experiments demonstrate the superiority of the proposed PAOA model.

Chapter 5. In this chapter, we propose a Feature-Distribution Perturbation and Calibration (PECA) method to derive generic feature representations for person ReID, which is not only discriminative across cameras but also agnostic and deployable to arbitrary unseen target domains. Specifically, we perform per-domain feature-distribution perturbation to refrain the model from overfitting to the domain-biased distribution of each source (seen) domain by enforcing feature invariance to distribution shifts caused by perturbation. In complementary, we designa global calibration mechanism to align feature distributions across all the source domains to improve the model's generalization capacity by eliminating domain bias. These local perturbation and global calibration are conducted simultaneously, which share the same principle to avoid models overfitting by regularization respectively on the perturbed and the original distributions. Extensive experiments are conducted on eight person ReID datasets and the proposed PECA model outperformed the SOTA competitors by significant margins.

Chapter 6. Existing Domain Generalizable ReID methods explore feature disentanglement to learn a compact generic feature space by eliminating domain-specific knowledge. Such methods not only sacrifice discrimination in target domains but also limit the model's robustness against per-identity appearance variations across views, which is an inherent characteristic of ReID. In this chapter, we formulate Cross-Domain Variations Mining (CDVM) model to simultaneously explore explicit domain-specific knowledge while advancing generalizable representation learning. Our key insight is that cross-domain style variations need to be explicitly modeled to represent per-identity cross-view appearance changes. CDVM retains the model's robustness against cross-view style variations that can reflect the specific characteristics of different domains whilst maximizing the learning of a globally generalizable (invariant) representation. To this end, we propose utilizing cross-domain consensus to learn a domain-agnostic generic prototype. Subsequently, this prototype is refined by incorporating cross-domain style variations, thereby achieving cross-view feature augmentation. Additionally, we further enhance the discriminative power of the augmented representation by formulating an identity attribute constraint to impose attention on the importance of individual attributes, while maintaining overall consistency across all pedestrians. Extensive experiments validate that the proposed CDVM model outperforms existing SOTA methods by significant margins.

These four solutions jointly solved the problem of domain distribution shift for OOD data by enableing the network to derive robust yet generalizable representation for the identities. Therefore, facilicating the differentiation the inter-class decision boundary and improving the matching accuracy among query and gallery instances.

Declaration

I hereby declare that this thesis has been composed by myself and that it describes my work. It has not been submitted, either in the same or different form, to this or any other university for a degree. All verbatim extracts are distinguished by quotation marks, and all sources of information have been acknowledged. Some parts of the work have previously been published or in submission as:

Chapter 3

- **Qilei Li**, Jiabo Huang, and Shaogang Gong. *Local-Global Associative Frame Assemble in Video ReID.* In Proc. British Machine Vision Conference (BMVC), Online, May 2021.

Chapter 4

- **Qilei Li**, Shaogang Gong. *Mitigate Domain Shift by Primary-Auxiliary Objectives Association for Generalizing Person ReID.* In Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Hawaii, United States, January 2024.

Chapter 5

- **Qilei Li**, Jiabo Huang, Jian Hu, and Shaogang Gong. *Feature-Distribution Perturbation and Aalibration for Generalized Person ReID.* In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seoul, Korea, April 2024.

Chapter 6

- **Qilei Li**, Shaogang Gong. *Generalizable Person ReID Attentive to Cross-Domain Style Variations.* Submitted to European Conference on Computer Vision (ECCV), 2024.

Acknowledgements

First of all, I would like to express my deep gratitude to my supervisor Prof. Sean (Shaogang) Gong for his perpetual patience, excellent guidance, and enthusiastic supervision. I was admitted in the year 2020, in the middle of the Covid-19 chaos. Despite the stringent campus restrictions at the time, Sean and I established a consistent schedule for our online weekly meetings throughout the year. Under his expert supervision and thoughtful guidance, I gradually acquired the skills to independently conduct research and generate innovative ideas. Besides, I would like to thank Dr. Qianni Zhang for being my second supervisor and Dr. Changjae Oh for being my independent assessor. I'm grateful for their constructive feedback throughout my PhD study. My warm appreciation goes to all members and visiting researchers at QMUL Vision Group for their friendship and support (in chronological order): Mr. Weitong Cai, Mr. Yu Cao, Mr. Zixu Cheng, Dr. Ke Han, Mr. Jian Hu, Dr. Jiabo Huang, Dr. Wei Li, Mr. Pan Li, Miss Jiayi Lin, Mr. Dezhao Luo, Miss. Shitong Sun, Dr. Guan'an Wang, Dr. Guile Wu, and Dr. Qingze Yin. I had a lot of discussions with them, particularly through the seminars organized by Dr. Wei Li, that initially guided me into the group and get connected with others. I'm honored to have a part-time industry work experience as a research scientist at Vision Semantics Ltd and Veritone Inc, and deeply appreciate my colleagues including Mr. Edward Burnell, Miss Yao Gong, Mr. Tim Kay, Dr. Alessandro Masullo, Dr. James McTavish, Dr. Georgia Rajamanoharan, Mr. Jozsef Szakacs, Mr. Bruce Tuch, and Mr. Thomas Walpole. Their insightful discussions were instrumental in transforming my research findings into practical AI applications. I am thankful to Dr. Da Li at Samsung AI Center Cambridge for the fabulous collaboration. I feel blessed to have enduring love and endless support from my parents. Meanwhile, I am indebted to Miss Yuting Zhou for her immense love, trust, and company all the time. I give sincere thanks to all friendly QMUL administrative and system support staff, especially Mr. Simon Butcher, Mr. Tom Bradford, Mr. Roland Jepsas, Mr. Edward Hoskins, Mr. Branavan Santhakumar, and Ms. Melissa Yeo, for their great help, as well as those who have helped me during my PhD studies but I may not have the fortune of knowing their names.

Contents

List of Figures

List of Tables

List of Abbreviations

BN	Batch Normalization.
CDVM	Cross-Domain Variations Mining.
CE	Cross-Entropy.
CMC	Cumulative Matching Characteristics.
CNN	Convolutional Neural Network.
DA	Domain Adaptation.
DG	Domain Generalization.
DG ReID	Domain Generalizable ReID.
DNNs	Deep Neural Networks.
DTO	Deployment-Time Optimization.
FC	Fully-Connected.
GCM	Global Calibration Module.
GCQ	Global Correlated Quality.
GRL	Gradient Reverse Layer.
ID	Identity.
IID	Independent and Identical Distribution.
IN	Instance Normalization.
KL	Kullback–Leibler.
LAQ	Local Aligned Quality.

LOGA	Local-Global Associative Assembling.
LPM	Local Perturbation Module.
mAP	mean Average Precision.
ML	Machine Learning.
MMD	Maximum Mean Discrepancy.
MSE	mean squared error.
MTL	Multi-Task Learning.
OOD	Out-Of-Distribution.
PAOA	Primary-Auxiliary Objectives Association.
PECA	Feature-Distribution Perturbation and Calibration.
QAN	Quality-Aware Network.
R1	Rank-1.
ReID	Re-Identification.
RNN	Recurrent Neural Network.
RTA	Reinforced Temporal Attention.
SOTA	State-of-the-Art.
TCA	Transfer Component Analysis.
TL	Transfer Learning.
VAE	Variational Autoencoder.

Chapter 1

Introduction

1.1 Scope the Thesis

With the rapid expansion of surveillance multi-camera systems around the world, associating people over space and time becomes an increasingly significant capability for a wide range of applications such as public safety, law enforcement, and forensic search. In this context, Person Re-Identification (ReID) [11, 24, 143] is a fundamental task which aims to retrieve the same pedestrian across non-overlapping camera views by measuring the distances among representations of all the candidates in a pre-constructed discriminative feature space. Person ReID encompasses two main aspects: video-based and image-based methods, and each with its distinct characteristics. Video-based ReID [143, 178] focuses on matching individuals across different camera views by analyzing temporal information, including movements and trajectories over time, to enhance identification accuracy. This approach capitalizes on dynamic behaviors, thereby improving robustness in matching individuals across non-overlapping camera views. Conversely, image-based ReID [1, 179, 104] concentrates on static images captured by surveillance cameras. It aims to identify individuals based on their appearance features extracted from single images. This approach is particularly important in scenarios where only static images are available for matching individuals across camera views. Despite their distinct methodologies, both video-based and image-based Person ReID methods are indispensable for surveillance and security applications in addressing the challenge of identifying individuals across non-overlapping camera views in diverse environments.

Accurate Person ReID relies heavily on extracting discriminative representations from tracklets (clips of frames) and individual images. In deep learning-based methods, this is typically achieved using a feature extractor, which maps raw inputs to high-dimensional feature representations. The discriminative nature of these representations is crucial for achieving high matching accuracy, as it increases the distance between different identities (inter-class distance) while minimizing the distance between instances of the same identity (intra-class distance). Additionally, beyond discrimination, another significant challenge is generalizability. Many networks are designed under the assumption that both training and deployment occur in the same environment, following the Independent and Identical Distribution (IID) assumption, which may not hold in real-world scenarios. To this end, this thesis focuses on enhancing a model's robustness and discrimination power of learned representations, while improving the generalizability of models trained on source domains when applied to unknown target domains, which may suffer from significant distribution shifts. Our ultimate goal is to enable the generalization of discriminative Person ReID models to real-world environments at scale, thus advancing the field towards practical deployment in diverse and challenging scenarios.

1.2 Frame Quality-aware Person Re-Identification

1.2.1 Problem Definition

Given N video tracklets $\mathcal{T} = \{\boldsymbol{T}_i\}_{i=1}^{N}$ with each containing L frames $\boldsymbol{T}_i = \{\boldsymbol{I}_j^i\}_{j=1}^{L}$ depicting C pedestrians in motion, the objective of video person ReID is to derive a representation model θ from the tracklets data $\mathcal{V}$ which is capable of extracting robust and discriminative feature representations $\boldsymbol{x}$: $f_\theta(\boldsymbol{T}) \rightarrow \boldsymbol{x}$ for ReID matching across disjoint camera views. Considering the diverse and unknown sources of noise commonly exist in surveillance videos which leads to distractions in different frames, it is essential for the model to effectively recognize visual patterns that are specific to each pedestrian to selectively assemble frames into a tracklet's representation. This is inherently challenging due to the uncertain nature of noise in tracklets of people in motion against backgrounds of visually similar distractors.

1.2.2 Challenges and Solutions

Challenges: Video-based person ReID necessitates the analysis and aggregation of information across a sequence of video frames within each tracklet to construct a more discriminative and

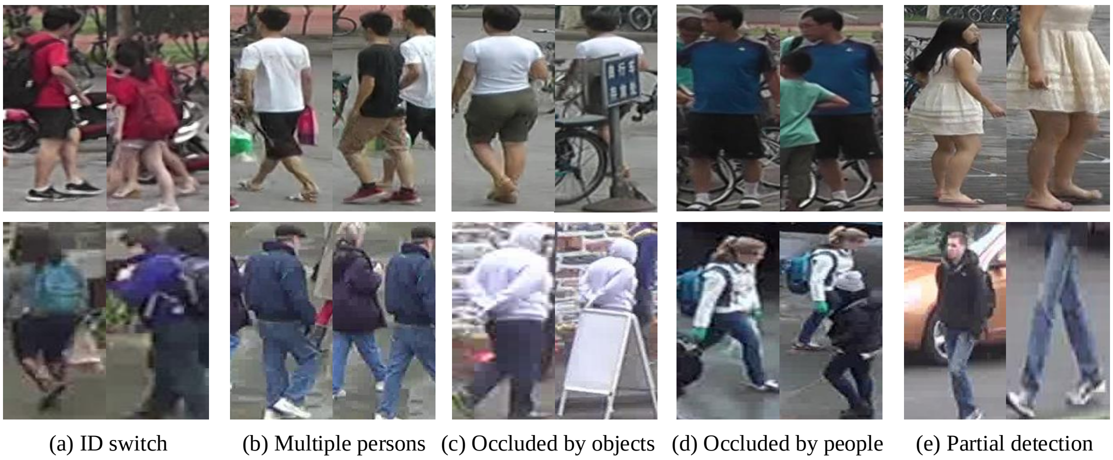

(a) ID switch (b) Multiple persons (c) Occluded by objects (d) Occluded by people (e) Partial detection

Figure 1.1: An illustration of various types of frame quality degradations.

robust representation of pedestrians in motion. However, tracklets often contain poor-quality frames resulting from occlusion, illumination changes, and identity switches [96, 157, 11, 25, 164, 50], as demonstrated in Figuer 1.1. Traditional video ReID methods typically extract per-frame representations, followed by global average pooling to obtain tracklet-level representations. These methods assume uniform frame quality, thereby neglecting the diverse qualities present and making them susceptible to various types of noise. Consequently, the discriminative power of the representation will be compromised, as illustrated in Figure 1.2. Numerous methods have been proposed to address this issue by selectively assembling high-quality frames, either through local alignment or global correlations. However, both approaches have drawbacks. Local alignment-based methods are fragile when detected pedestrians are not well-aligned, while global appearance-based approaches are spatially insensitive and may lead to misaligned patterns, especially in complex backgrounds.

Solution: To address the challenges posed by low-quality frames, we propose an approach for video person ReID termed Local-Global Associative Assembling (LOGA). LOGA dynamically assembles video frames within the same tracklet using two key modules: Local Aligned Quality (LAQ) module and Global Correlated Quality (GCQ) module. These modules assess the importance and relevance of frames based on their alignments in local regions and global appearance correlations, as well as their mutual reinforcement. Additionally, LOGA model constructs a local-assembled global appearance prototype to leverage both types of information and foster mutual complementarity by learning their consensus. Unlike most existing spatial-temporal

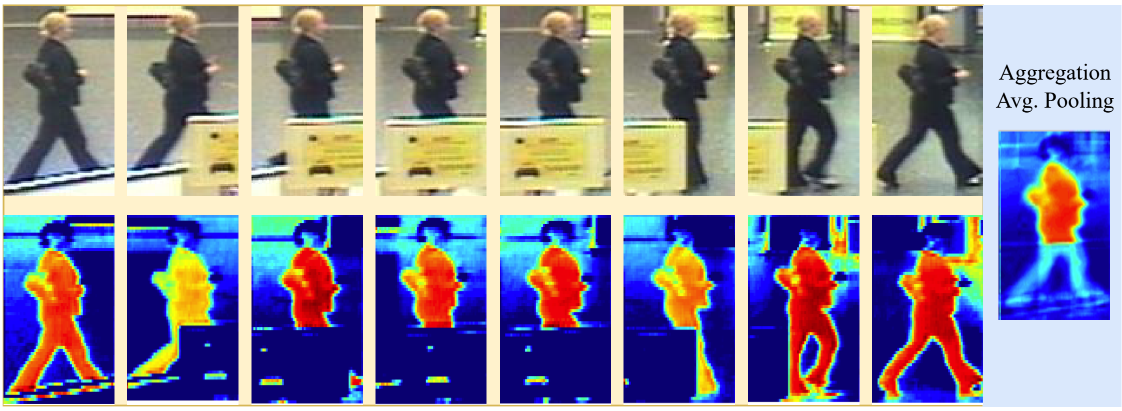


Figure 1.2: An illustration of object occlusion and the heatmap from global average pooling on the frames within a tracklet [128].

attentive methods, which focus on integrating temporal information with intraframe spatial attention, LOGA aims to more effectively exploit inter-frame complements. This approach is distinct and stands to benefit from advancements in per-frame learning. Furthermore, to harness the benefits of both local and global information and exploit their mutual advantages, we define the tracklet's representation assembled by the LAQ module as a prototype, which is compared with global visual features in the GCQ module. Through this association, we encourage the two modules to find a trade-off between local and global knowledge to cope with different types of noise more reliably.

1.3 Cross-domain Generalizable Person Re-Identification

1.3.1 Problem Definition

Despite the great progress made over recent years, most existing ReID methods [155, 95, 181, 174] are built upon the fragile Independent and Identical Distribution (IID) assumption. The performance degrades significantly when deployed on a new test domain due to the covariant shift. This refers to the situation where the input distributions of training and testing data differ, causing a discrepancy in the learned model's performance. This phenomenon poses a significant challenge in real-world scenarios where the target domain during testing may vary from the source domain during training. To illustrate this challenge, a few samples from different domains are shown in Figure 1.3, from which we can observe significant domain shifts caused by various factors. Unlike traditional ReID methods, Domain Generalizable ReID (DG ReID) assuming the absence of target domains during training, aims to learn a generalizable model which can extract discriminative representations in any new environment. It is naturally challenging but practical and has attracted increasing attention. Despite its inherent difficulty, DG ReID holds practical significance and has garnered increasing attention within the research community. By focusing on learning domain-agnostic representations, DG ReID aims to enhance the generalizability and adaptability of ReID models, making them more applicable in real-world settings with diverse environmental conditions. Mathematically, the problem of DG ReID can be formalized as follows: Let $\mathcal{X}_s$ and $\mathcal{Y}_s$ denote the source domain data and corresponding labels, respectively, where $\mathcal{X}_s = \{x_1^s, x_2^s, ..., x_n^s\}$ and $\mathcal{Y}_s = \{y_1^s, y_2^s, ..., y_n^s\}$ with n samples. Similarly, let $\mathcal{X}$ represent the target domain data, where $\mathcal{X}_t = \{x_1^t, x_2^t, ..., x_m^t\}$ with m samples. The goal of DG ReID is to learn a domain-agnostic feature extractor f_θ parameterized by learnable weight θ that maps input data x to a discriminative feature space $\mathcal{Z}$, *i.e* $f : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is the feature space shared across different domains. The objective function is

$$\underset{\theta}{\text{minimize}} \; \mathcal{L}_{\text{reid}}(f(\mathcal{X}_s), \mathcal{Y}_s) + \mathcal{L}_{\text{feat}}(f(\mathcal{X}_s)) \tag{1.1}$$

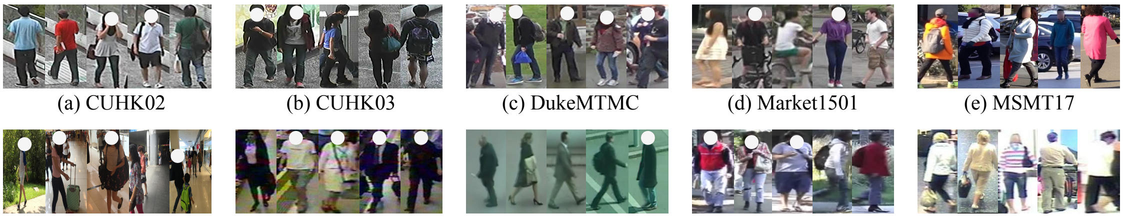
where $\mathcal{L}_{\text{reid}}$ denotes the ReID loss function and $\mathcal{L}_{\text{feat}}$ represents the regularization term, such as KL divergence, which encourages domain-invariant representations extracted by the extractor. To further leverage the diverse training samples from multiple source domains, current research is increasingly focusing on multi-source Domain Generalization for ReID. In this context, the objective function can be reformulated as follows:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{k} \mathcal{L}_{\text{reid}}(f(\mathcal{X}s_i), \mathcal{Y}s_i) + \hat{f}(\mathcal{X}t)\mathcal{L}_{\text{feat}}(f(\mathcal{X}s_i)), \tag{1.2}$$

where k denotes the number of source domains. Recognizing the practical value of multi-source domain generalization in real-world scenarios, this thesis primarily focuses on exploring and advancing this area.

1.3.2 Challenges and Solutions

Challenges: Existing Domain Generalizable ReID methods are typically classified into four groups. The first group focuses on feature disentanglement, aiming to identify explanatory and independent factors by separating domain-invariant components from identity representations. Notably, feature normalization techniques like Instance Normalization (IN) have been extensively studied to reduce style discrepancies among normalized representations [64, 112]. How-



(a) CUHK02 (b) CUHK03 (c) DukeMTMC (d) Market1501 (e) MSMT17

(f) CUHK-SYSU (g) GRID (h) PRID (i) VIPeR (j) iLIDS

Figure 1.3: Visualization on a few identity samples from different domains. Significant domain gaps are caused by the variation on nationality, illumination, viewpoints, resolution, scenario, etc.

ever, while these methods explicitly reduce domain-invariant components, they often compromise the discriminative capability of the acquired representations due to the limited information retained in the disentangled features. Another approach involves aligning the target domain with the Batch Normalization (BN) statistics calculated over the source domain. Additionally, meta-learning has been widely explored to simulate the training-testing discrepancy and enable domain-agnostic feature extraction [177, 17, 4]. Ensemble learning-based techniques represent another group, aggregating descriptors from multiple experts to construct a more robust representation [162, 156, 18]. Alternatively, some methods aim to leverage diverse training data through augmentation. Despite their demonstrated effectiveness, both strategies have limitations in effectively managing cross-domain conflicts and exploring cross-domain correlations. Despite the advancements made by State-of-the-Art (SOTA) models, significant room for improvement remains, as evidenced by low mean Average Precision (mAP) scores, such as less than 20% on MSMT17 and less than 40% on CUHK03. This is primarily attributed to domain-specific interference in the source domain, which hinders the learning of a domain-invariant model.

Solution 1: The first solution is to minimize domain-specific contextual interference in model learning by focusing more on the domain-invariant person's unique characteristics. This is achieved by introducing the association of learning the primary instance classification objective function with an auxiliary weakly labeled/supervised pedestrian saliency detection objective function. Specifically, by two steps : (1) Additionally train a pedestrian saliency detection head with an auxiliary supervision to assist in focusing the primary ReID discriminative learning task on more domain-invariant feature characteristics. (2) Eliminate the interference attributed to inaccurate saliency labels by calibrating the gradients of the shared feature extractor raised from the weakly labeled auxiliary learning task towards that of the primary task as a reference when they are in conflict [122]. This association mechanism helps ensure the ReID model learns to attentively focus on generic yet discriminative pedestrian information whilst both learning tasks are harmoniously trained.

Solution 2: The second solution is to to diversify the feature distribution based on a perturbing factor estimated per domain, which enables the model to be more invariant to distribution shifts, and simultaneously calibrate the feature distributions across all the source domains, so to eliminate the domain-specific data characteristics in feature representations that are potentially caused by identity-irrelevant redundancy. Both local perturbation and global calibration can reinforce the same purpose of regularizing the model training, but they are devised in different hierarchies and complementary to each other, further to promote the model in learning domain-agnostic representations.

Solution 3: The third solution is to enhance the diversity of per-identity instances through the introduction of cross-view style variations across different domains. The objective is to expand the cross-view style inherent to individual identity to learn a generalizable ReID representation that is more robust under the presence of such cross-view style variations. Specifically, we first learn a domain-agnostic (generalizable) identity prototype by exploiting the consensus of identities regardless of their specific domain annotations. Secondly, we enhance the model's robustness by mitigating the covariance from cross-view style variations. This involves augmenting the prototype with cross-domain variations through multi-view augmentation, to simulate the style discrepancy for one identity between query and gallery views. Thirdly, we highlight person-specific attributes to increase the feature discrimination while maintaining the overall consistency across all pedestrians.

1.4 Contributions

The contributions made in this thesis are summarised as follows:

1. **Chapter 3:** We explore the association and mutual promotion of frame's local part alignments and global appearance correlations in assembling a sequence descriptor so to improve the model's robustness to noisy frames and inter-frame ID-switch in video ReID. The association and mutual promotion of frame's local part alignments and global appearance correlations are explored in assembling a sequence descriptor to improve the robustness of

the model to noisy frames and inter-frame ID-switch in video ReID. A video person ReID model termed Local-Global Associative Assembling (LOGA) is proposed, which learns a discriminative and reliable representation for video tracklets by adaptively assembling frames of diverse qualities. A local-assembled global appearance prototype is introduced to associate the local and global visual information by exploiting their mutual agreements to facilitate the learning of a discriminative tracklet representation.

2. **Chapter 4:** We introduce the idea of optimizing a more domain-generic ReID learning task that emphasizes domain-invariant pedestrian characteristics by associating the ReID instance discriminative learning objective to an auxiliary pedestrian saliency detection objective in a way that does not create conflicts or hinder the effectiveness of the primary objective. A regularization called Primary-Auxiliary Objectives Association (PAOA) is formulated to implement the proposed association learning. It jointly trains the primary and auxiliary tasks with referenced gradient calibration to resolve the conflicting optimization criteria between the two learning objectives and promote the learning of a more domain-generic ReID model. The target domain test data characteristics are further explored by incorporating the PAOA regularization into a deployment-time model online optimization process. To that end, we formulate a PAOA+ mechanism for on-the-fly target-aware model optimization and show its performance benefit.

3. **Chapter 5:** We design Feature-Distribution Perturbation and Calibration (PECA) to exploit jointly the local feature-distribution perturbation and the global feature-distribution calibration to improve the model's generalizability to arbitrary unseen domains while maintaining its discrimination. A Local Perturbation Module (LPM) is formulated to diversify per-domain feature distribution, thereby preventing the model from over-fitting to each source domain. Additionally, a Global Calibration Module (GCM) is introduced to further eliminate domain bias by aligning the distribution of multiple source domains. We simultaneously regularize both to strike the optimal balance between these two competing objectives.

4. **Chapter 6:** We propose Cross-Domain Variations Mining (CDVM) to pioneer cross-domain variations to implicitly explore per-identity multi-view augmentation, thereby encouraging the model to learn and maximize invariant representations subject to cross-

camera identity retrieval. A CDVM mechanism is formulated to learn a context-aware generalizable ReID model sensitive to domain-specific cross-camera person-wise variations. This mechanism optimizes jointly two competing criteria of generalizability and specificity. The proposed new model outperforms existing state-of-the-art methods by a large margin on a wide range of benchmarks.

1.5 Thesis Outline

The remaining chapters of this thesis are organized as follows:

(**1**) **Chapter 2** presents a review of extant literature pertinent to the principal components of this thesis.

(**2**) **Chapter 3** proposes a tracklet frame assembling approach called LOGA for video person ReID. LOGA aims to adaptively assemble video frames in the same tracklets by a Local Aligned Quality (LAQ) module and a Global Correlated Quality (GCQ) module to assess importance/relevance of the frames by associatively their alignments in local part and global appearance correlations, as well as their mutual reinforcements.

(**3**) **Chapter 4** introduces a PAOA model that guides model learning of the primary ReID instance classification objective by a concurrent auxiliary learning objective. To solve the problem of conflicting optimization criteria between the two learning objectives, PAOA calibrates the loss gradients of the auxiliary task towards the gradients of the primary task.

(**4**) **Chapter 5** presents a PECA model to accomplish generalized ReID by regularization respectively on the perturbed and the original distributions. These local perturbation and global calibration are conducted simultaneously, with the objective of learning more generalizable discriminative representations for model deployment to unseen target domains.

(**5**) **Chapter 6** introduces a CDVM model to enhance the diversity of per-identity instances through the cross-view style variations across different domains. The objective is to expand the cross-view style inherent to individual identity to learn a generalizable ReID representation that is more robust under the presence of such cross-view style variations.

(**6**) **Chapter 7** provides the conclusion and various research problems and directions to be pursued as further work.

Chapter 2

Literature Review

2.1 Person Re-Identification – Zero-shot Person Retrieval

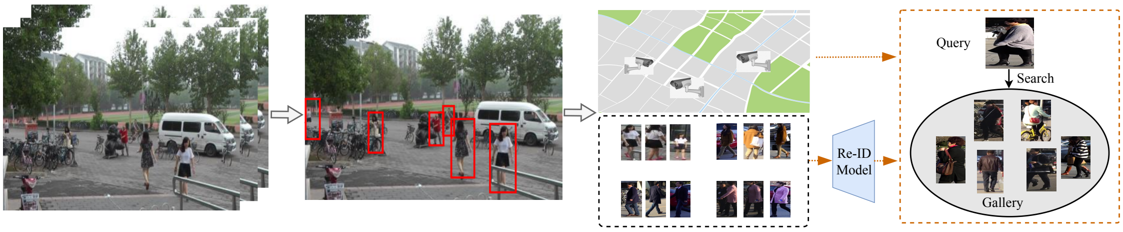


Figure 2.1: Pipeline of a practical person search system, involving five main steps: 1) Video data collection, 2) Bounding box annotation, 3) Tracklet segmentation, 4) ReID model training, 5) Identity retrieval.

Re-Identification (ReID) [11, 24, 143] is a critical task in computer vision that aims to accurately identify and retrieve pedestrians across different camera views. As illustrated in Figure 2.1, it's a critical component of a practical person search system that operates by associating a query person with individuals from a gallery. The query person can be represented through various mediums, such as an image [1, 179, 104], a video sequence [143, 178], or even a text description [161, 83]. Unlike traditional person classification tasks, ReID is considered a **zero-shot** learning task because the identities present in the training data are strictly non-overlap with those in the test data. In other words, they have absolutely disjoint label space. This fundamental characteristic adds a layer of complexity to the task, making ReID a highly challenging representation learning problem. In order to achieve accurate and reliable person re-identification, the learned representations must exhibit two crucial qualities: generalizability and discriminativeness. Generalizability in ReID refers to the ability of the learned representations to effectively capture the inherent characteristics of pedestrians across diverse environments and conditions. This is essential because the appearance of individuals can vary significantly due to factors such as changes in illumination, viewpoint, pose, occlusion, and clothing. Therefore, the learned representations should be robust enough to handle such variations and accurately match individuals across different camera views and scenarios. On the other hand, discrimination ability is equally important in ReID, as it ensures that the learned representations can effectively distinguish between different individuals, even when they share similar visual attributes. The representations should encode unique features that are specific to each individual, which enables reliable identification and retrieval, even in challenging scenarios where there may be significant visual similarities between different pedestrians.

The query person can be represented through various mediums, such as an image [1, 179, 104], a video sequence [143, 178], or even a text description [161, 83]. Due to the urgent demand for public safety and the increasing number of surveillance cameras, person ReID has become a critical component of intelligent surveillance systems, and offers significant research impact and practical value. Early ReID studies focus on exploring appearance patterns unique per identity [27, 87, 160], which has shown remarkable discrimination capacity. However, these methods often rely on the assumption of meticulously curated data with complete identity information, a premise that greatly limits their applicability in real-world scenarios where unpredictable environments are commonplace and video data are collected [80, 97]. Video-based person ReID [96, 157, 11, 25, 164, 50] extends beyond still images by analyzing and integrating information across a series of video frames. This process aims to construct a more discerning and resilient representation of individuals in motion, thereby reducing the impact of substandard frames and ID switches.

2.1.1 Video-based Person Re-Identification

Video person ReID is a popular subtopic [46] of ReID, where each person is depicted by a video sequence with multiple frames. The matching process of video-based ReID is shown in Figure 2.2. A video tracklet is a sequence of frames that captures rich variations of the same person. These video tracklets inherently constitute rich and informative data sources for ReID.

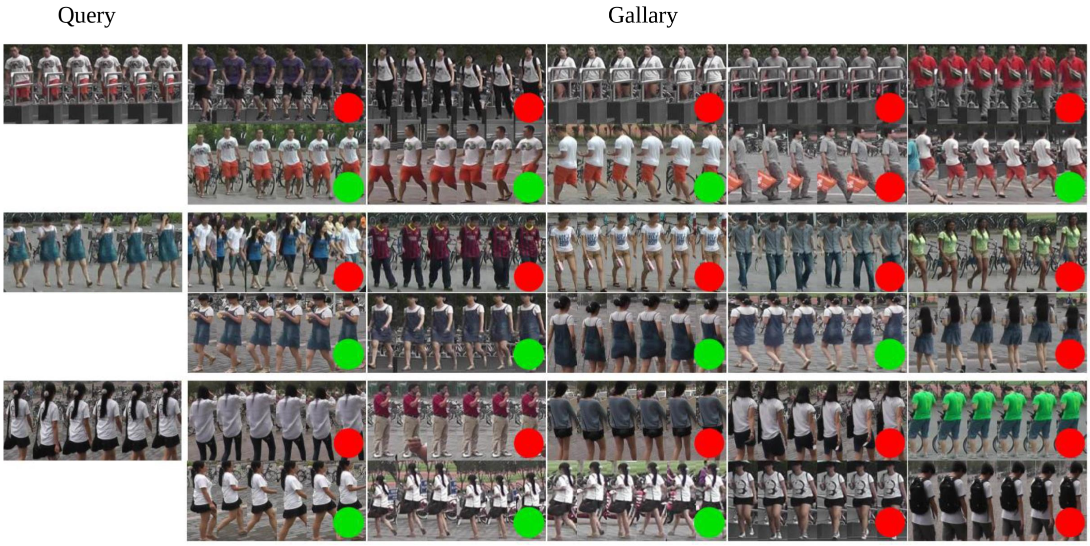


Figure 2.2: An illustration of video personal Re-Identification. Each clip refers to a video sequence for a pedestrian. The given query is searched across the gallery for matching. Green denotes positive matching and red denotes negative matching.

Due to its rich appearance and temporal information, it has gained increasing interest in the ReID community. Moreover, the video person ReID brings in additional challenges in video feature representation learning with multiple images.

Related Works. Niall *et al* [106] devise a deep neural network incorporating pooling and recirculation mechanisms, amalgamating temporal data into a unified feature vector. Inspired by the developments in 3D convolutional neural networks [9, 60], Li *et al* [76] develop a method that uses 3D convolutional automatic learning to explore relationships along spatial and temporal dimensions, transitioning from low-level to high-level features for the first time. Gu *et al* [37] insert the APM module prior to 3*D* convolution to address the feature alignment problem. Moreover, Zhao *et al* [176] introduce an attribute-based approach for feature-weighted frames and entanglement resolution. This methodology divides single-frame features into distinct categories of sub-features, each representing specific semantic information. The attention mechanism is becoming increasingly vital in person ReID. Instead of processing frames separately, certain studies [126, 82] utilize attention mechanisms to focus on identity-revealing regions. Li *et al* [82] employ the interaction of multiple spatial attention modules to emphasize crucial spatial regions across different frames. Spatial features can be aggregated using learnable temporal attention mechanisms. Zhang *et al* [173] introduce an attention mechanism with global reference, facilitating the learning of attention in regions pertinent to the global context. Liu *et al* [91] utilize a non-local self-attention mechanism, which has gained popularity in Convolutional Neural Network (CNN) backbone networks. Song *et al* [127] propose a mask-directed network, integrating masks with character images to reduce background interference. Chen *et al* [13] capture temporal and spatial features and computed attention value maps to specify the importance of different components of the person. The Concentrated Multi-grained Multi-Attention Network (CMMANet) is proposed in [54] to manage multi-scale features and extract details at multiple granularities, featuring a multi-attention module in each block for adaptive region retrieval within frame sequences. Hou *et al* [49] propose a computationally less complex bilateral complementary network, which preserves the spatial features of the original image. Additionally, some recent work has started to address computer vision problems using self-attention mechanisms [144, 6]. Self-attention is typically a non-local network devoid of spatial encoding and featuring multiple attention heads initially devised to address video classification challenges [144]. The Transformer architecture, which employs non-local attention as a key component, has achieved notable progress in video-based person ReID. He *et al* [42] introduce a hybrid interactive learning architecture that combines CNN with attention mechanisms for a video-based person ReID. Zhang *et al* [169] introduce the inaugural Transformer and a data pre-training technique to alleviate overfitting in re-id tasks. However, the substantial computational complexity associated with traditional self-attention results in notable computational overheads. To mitigate this issue, axial attention has been introduced in [48]. By breaking down operations, axial attention can substantially decrease computational costs. Shen *et al* [124] devise an unsupervised algorithm that aligns the ranking mechanism with the ReID approach.

Video-ReID Datasets. The statistics of commonly employed benchmarks are shown in Table 2.1. The DukeMTMC dataset [183] comprisess video sequences captured by eight different

Table 2.1: The statistics of commonly employed video person ReID datasets.

Dataset	#identity	#sequence	#boxes	#frame	#indoor cam.	#outdoor cam.	Detector	Evaluation
DukeMTMC [183]	1,404	4,832	815,420	168	0	8	Hand	CMC + mAP
Duke-SI [80]	1778	4,832	815,420	168	0	8	SSD	CMC + mAP
MARS [178]	1,261	20,715	1,067,516	58	0	6	DPM	CMC + mAP
PRID [47]	200	400	4,003,331	100	0	2	Hand	CMC
iLIDS-VID [180]	300	600	4,246,031	73	2	0	Hand	CMC
LS-VID [75]	3,772	14,943	2,982,685	200	3	12	Faster R-CNN	CMC + mAP

cameras. It offers diverse variations in pedestrian posture, movement, perspective, and lighting conditions across cameras. This diversity demands that algorithms exhibit robustness and an ability to generalize effectively. The Duke-SI dataset [80] is a fully auto-generated version of DukeMTMC without manual frames selection, thus, more practical and challenging. The MARS dataset [178] stands out as the largest and most widely employed dataset for video person ReID. It's captured by six cameras and has significant complexity, characterized by extensive overlaps of pedestrians, prevalent occlusions, and varying viewing angles among the cameras. This complexity presents a challenging environment for video analysis tasks. The iLIDS-VID dataset [180] comprises video sequences captured by two cameras, presenting challenges such as illumination changes, attitude variations, and pedestrian overlaps, aligning closely with real-world scenario requirements for the task. The PRID dataset [47] comprises video sequences obtained from two distinct cameras, featuring variations in pedestrian posture, perspective changes, background interference, and other factors. This dataset serves as a valuable resource for enhancing the accuracy and real-time performance of pedestrian re-identification systems. The LS-VID dataset [75] is captured by 15 cameras and consists of a total of 14,943 video sequences. Notably, this dataset offers more precise pedestrian trajectories, rendering it valuable for research and development in pedestrian tracking and analysis.

2.1.2 Image-based Person Re-Identification

Image-based person ReID is an essential research area within the field of person ReID. The Image-based person ReID tasks still encounter numerous challenges. These encompass perspective shifts, lighting fluctuations, alterations in appearance, and instances where pedestrians are obstructed by other objects.

Related Works. In recent years, numerous studies [133, 52, 10, 36] have shifted their focus towards investigating image-based person ReID tasks. Currently, lightweight network architectures have garnered increasing interest among researchers. Zhou *et al* [187] devise a network that is compact and comprehensive, capable of discerning diverse spatial scales and facilitating multi-scale cooperation. Li *et al* [72] introduce a fusion depth space approach to emphasize the amalgamation of pattern information inherent in pedestrian images. It realizes a cost-effective search method, enhancing the model's generalization capability, recognition accuracy, and feature representation. Gu *et al* [36] employe a novel twin comparison mechanism to explore effective lightweight architectures. They introduced a multi-scale interaction space, offering a rational approach for interacting with multi-scale features. Furthermore, certain studies enhance model performance by refining the loss function, thereby achieving superior results in person ReID tasks. Gu *et al* [35] introduce AutoLoss-GMS, a method designed to search for an optimized loss function within the loss function space, with the aim to achieve efficient and effective person ReID. Chen *et al* [14] develop a quadruplet loss function and proposed a quadruplet deep network, integrating online hard negative mining to enhance the model's generalization capability. Yan *et al* [158] introduces a paired loss function as an alternative to the conventional triplet loss. This novel approach is tailored to dynamically apply exponential penalties to images exhibiting minor differences, while imposing bounded penalties on those with substantial distinctions, facilitating learning fine-grained features. Dong *et al* [20] investigate human partial masks and human poses to enhance feature extraction from the human body. Meanwhile, Zheng *et al* [182] endeavors to map 2D images into 3D spaces to conduct person searches within these 3D spaces. Karianakis *et al* [65] propose a neural network architecture termed Reinforced Temporal Attention (RTA). Besides, some studies [133, 134] have explored scenarios involving image occlusion. Tan *et al* [133] propose a multi-head self-attention network designed to eliminate irrelevant information and capture crucial local features, particularly to tackle occlusion challenges. The RFCNet [52] integrates spatial and temporal RFC to predict features within obscured regions, thereby enabling the network to leverage both images and videos as sources of information. Tan *et al* [134] propose a dynamic prototype mask based on two prior knowledge to bridge the domain gap between the auxiliary model and the ReID dataset, effectively improving the performance of the model.

Image-ReID Datasets. The statistics of commonly utilized image person ReID datasets are shown in Table 2.2. The Market-1501 dataset [179], introduced in 2015, is captured by five high-resolution cameras and one low-resolution camera. It employs DPM pedestrian detector to identify pedestrian bounding boxes and contain a total of 32,668 images. This dataset represents a large-scale resource for person ReID studies. The VIPeR dataset [33] stands as the pioneering small-scale person dataset, featuring manually annotated pedestrians across 1,264 images. Acquired from two cameras, the VIPeR incorporates varying viewing angles, as well as changes in posture and lighting. It continues to be recognized as one of the most formidable datasets for person ReID tasks. The MSMT17 dataset [147], introduced in 2018, represents a large-scale person ReID dataset. Employing the Faster R-CNN [119] pedestrian detector, it automatically identifies

Table 2.2: The statistics of commonly employed image person ReID datasets.

Datasets	#ID	#Track(#Bbox)	#cam.	Label	Res.	Eval.
VIPeR [34]	632	1,264	2	hand	fixed	CMC
iLIDS [180]	119	476	2	hand	vary	CMC
GRID [99]	250	1,275	8	hand	vary	CMC
PRID2011 [16]	200	1,134	2	hand	fixed	CMC
CUHK01 [85]	971	3,884	2	hand	fixed	CMC
CUHK02 [84]	1,816	7,264	10	hand	fixed	CMC
CUHK03 [86]	1,467	13,164	2	both	vary	CMC
Market-1501 [179]	1,501	32,668	6	both	fixed	CMC+mAP
DukeMTMC [183]	1,404	36,411	8	both	fixed	CMC+mAP
MSMT17 [147]	4,101	126,441	15	auto	vary	CMC+mAP

labeled frames, resulting in a compilation of 126,441 images. Captured by 15 campus cameras, MSMT17 is designed to encompass a wider range of scenes. Notably, within a single scene, there are minimal variations in lighting conditions. The DukeMTMC-reID dataset [183], collected at Duke University, is captured by eight stationary high-definition cameras. It includes a rich set of 16,522 training images, 2,228 query images, and an extensive gallery, forming a specialized subset of the MTMCT dataset, specifically the DukeMTMC [120]. The CUHK01 dataset [85] encompasses 971 individuals and 3,884 manually cropped images. Each individual is represented by a minimum of two images, captured from two disjointed camera views. The CUHK02 dataset [84] comprises 1,816 individuals and 7,264 manually cropped images. In contrast to the CUHK01 dataset, CUHK02 offers a more extensive array of identity and camera views, facilitating greater variability in pedestrian image configurations. The CUHK03 dataset [86] is a comprehensive person ReID dataset gathered in Hong Kong. It encompasses 1,360 distinct pedestrians, represented by a total of 13,164 images.

2.1.3 Quality-aware Assembling for Video ReID

Video-based person ReID methods aim to learn an expressive appearance feature or distance metric from a sequence of frames, *i.e*, a video tracklet, by taking advantage of the additional temporal information and complementary spatial information intrinsically available in video tracklets, However, there can be low-quality frames over movement caused by occlusion and scale variations, as shown in Figure 2.3. To reduce the interference caused by these low-quality frames,

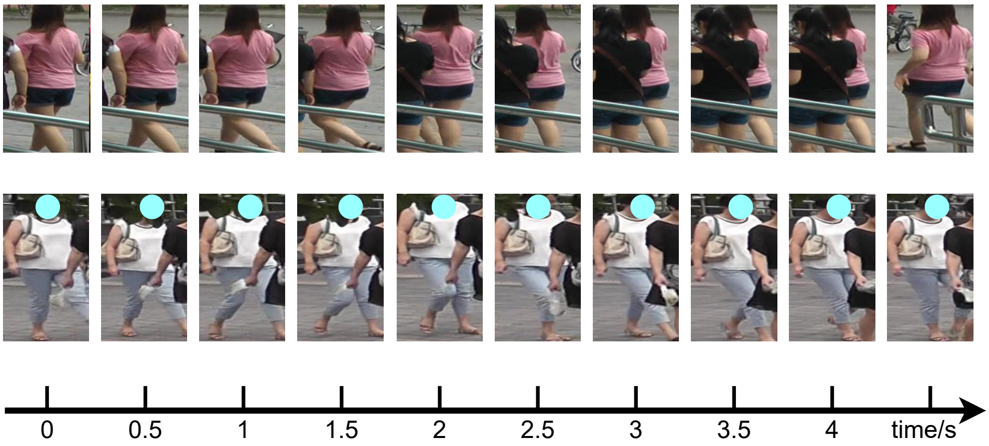


Figure 2.3: An illustration of frame quality variation within a tracklet over a period of time.

existing approaches explore either local part alignments [128, 5, 173, 51, 50] or global appearance correlations [97, 75, 168, 78, 89, 105, 95, 145] to assemble the per-frame representations with high robustness to their diverse qualities.

Local Part Alignments. Song *et al* [128] introduced a region-based quality estimation network that employs a training mechanism for extracting region-based complementary information across various frames. The network architecture is illustrated in Figure 2.4. It utlizes landmarks to denote significant points on the human body, and processes them through Fully-Connected (FC) layers to generate intermediate representation. Then, the representation is divided into distinct regions based on the identified key points. A region-based quality predictor is used to predict image quality, and produces a fixed-dimensional feature representation with a sequence size through weighted aggregation of all frames. Considering the consistent body structure shared among human, It is intuitive to differentiate images/frames of pedestrians regarding their visual similarity in different parts. Local-parts assembling approaches [128, 5, 173, 51, 50] apply per-part comparisons of video frames in the same tracklets to identify outliers that are misaligned with others in most local parts, and to restore the corrupted parts of frames with the complements of others [51, 50], or degrade their importance in frame assembling [128, 5, 173]. However, due to unreliable auto-generated person bounding boxes this assumption that a pedestrian detected in different video frames being mostly well-aligned is often invalid, *e.g* the importance of a noise-free video frame might be underestimated. In this work, we further consider the holistic visual similarity of video frames in the assessment of their quality, thereby helping refrain from inaccurate assessments caused by part misalignments.

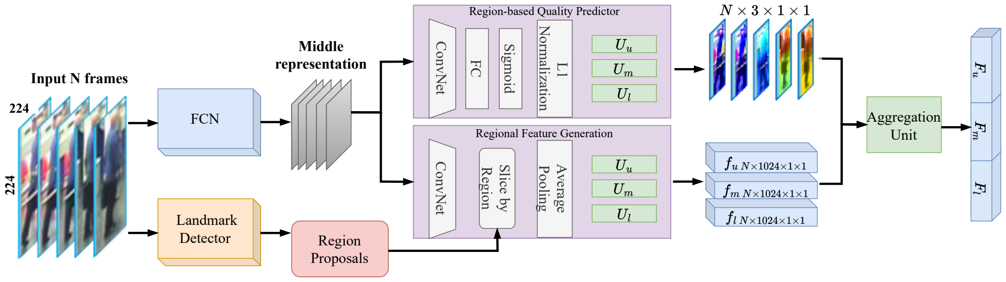


Figure 2.4: An illustration of a local part alignment-based method [128].

Global Appearance Correlations. Liu *et al* [97] introduced a method termed Quality-Aware Network (QAN). It is designed for learning the metric between two sets of images, assuming that each collection contains images from the same identity. The architecture of the network is illustrated in Figure 2.5. QAN comprises two branches: one branch predicts the quality score for each sample, while the other branch extracts appearance feature embeddings from each sample. These features and quality scores of all samples within the set are aggregated to produce the final feature embeddings. Compared to local-parts approaches, methods based on global-appearance [97, 75, 168, 78, 89, 105, 95, 145] take advantage of the strong representational power of CNN [30, 70] to learn correlations between video frames holistically so that the irrelevant frames, which are likely of low quality, are suppressed in frame assembling. However, the CNN features can be insensitive to a spatial shift resulting in potential mis-correlations of visually similar but irrelevant parts, *e.g* the ID-switch issue is shown in Figure 2.3. Detecting the subtle differences in the outfits of the two pedestrians is challenging. This may lead to mis-assembling of frames to represent a tracklet. To address this problem, we propose enhancing global-appearance methods by jointly exploring holistic visual correlations among frames and aligning their local parts through inter-frame spatial relations.

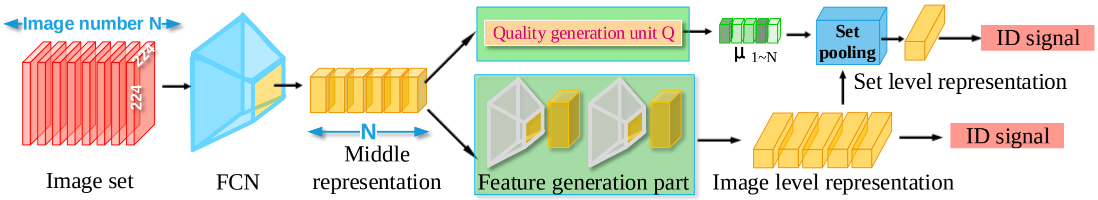


Figure 2.5: An illustration of a global appearance correlation-based method [97].

Local-Global Joint. Beyond the temporal assembling approaches discussed above, spatial attention [148] is also popular in both image and video person ReID [87, 184, 153, 25, 151]. By exploring different parts within a single frame, the spatial attention mechanism can dynamically focus on the more discriminative parts. In contrast, our LAQ module investigates the alignments of the same part across different video frames, focusing on exploiting complementary inter-frame information in a tracklet. Chen *et al* [15] explored both local and global information for frame assembling in video ReID. However, it learns from these two types of information separately through a dual-branch network without considering their synergy (Figure 2.6). We demonstrate the superiority of the proposed LOGA over FGRA in both performance evaluation and ablation analysis in Chapter 3.

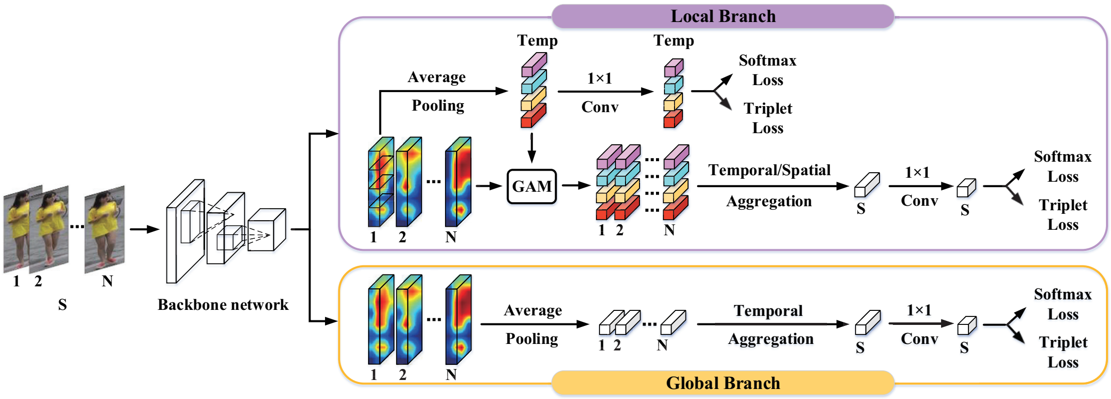


Figure 2.6: An illustration of FGRA [15], which jointly exploits local and global information *separately* in a dual-branch network.

2.1.4 Cross-domain Generalizable Re-Identification

To address this problem, Domain Generalizable ReID (DG ReID) [171, 22, 192, 186, 188, 103, 64, 112] has garnered increasing attention as a potential solution. In contrast to domain adaptive ReID, which adjusts learned representations using unlabeled samples from the target domain. Domain generalizable ReID operates without prior knowledge of the target domain, thus presenting a practical yet challenging task. Domain-adaptive ReID focuses on aligning representations between a labeled source domain and an unlabeled target domain, while DG ReID aims to learn robust representations transferable across diverse domains without specific adaptation. The comparison between Domain Adaptation (DA) and Domain Generalization (DG) approaches is illustrated in Figure 2.7, where DA-based methods may achieve higher performance with target domain data, while DG methods offer practicality in scenarios where access to target domain data is limited or unavailable. Existing DG ReID methods generally designed with the following principles: (1) To benefit the model from the diverse training data achieved by augmentation. (2) To align the target domain with the Batch Normalization (BN) statistics calculated over the source domain. (3) To mimic the train/test discrepancy with meta-learning. The first group of methods [171, 22, 192] revolves around utilizing feature disentanglement to explore explanatory and independent factors by decoupling domain-invariant components from an identity representation. Notably, feature normalization techniques, such as Instance Normalization (IN) have been extensively researched to minimize style discrepancy among the normalized representations [64, 112]. However, while these methods can explicitly reduce domain-invariant components, they inevitably diminish the discriminative capability of the acquired representations due to limited information being retained in the disentangled feature. Furthermore, Meta-learning which imitates the training-testing discrepancy has been widely studied to enable the extracted features to be domain-agnostic [177, 17, 4]. On the other hand, ensemble learning-based techniques often aggregate descriptors derived from multiple experts to assemble a more resilient representation [162, 156, 18]. Despite the improvement obtained by these SOTA models, these strategies have limitations in effectively managing cross-domain conflicts and exploring cross-domain correlations, and leave significant room for improvement, as indicated by the low mAP

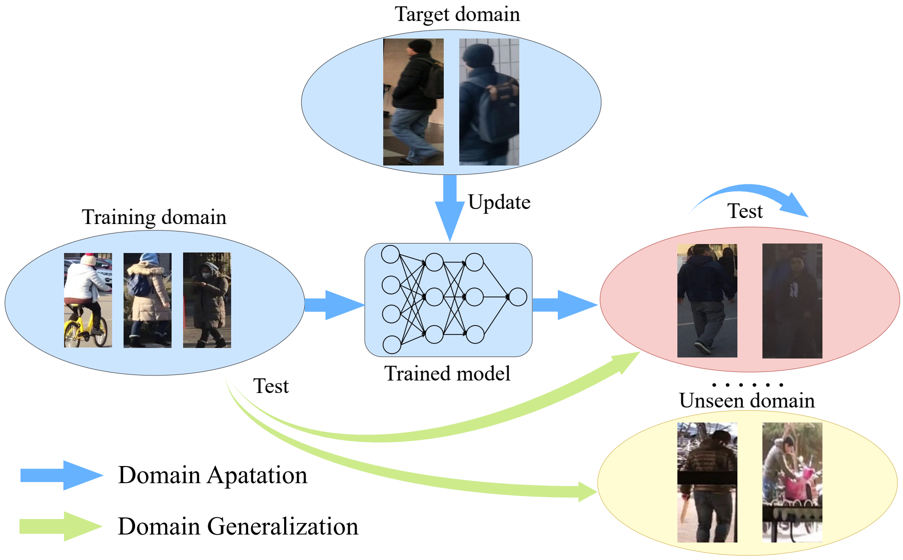


Figure 2.7: Comparison on Domain Adaptation and Domain Generalization for person ReID.

scores, *e.g* less than 20% on MSMT17 and less than 40% on CUHK03. This is attributed to the domain-specific interference in the source domain that limits the learning of a domain-invariant model. In Chapter 6, we formulate a new generalizable ReID model termed CDVM pioneering the incorporation of cross-domain variations to simulate the style shifts for one identity captured by disjoint cameras. This innovation is intended to enhance the model's robustness against domain-shifts and to extract discriminative representations. In Chapter 4, we aim to tackle this issue by guiding the model to focus on the discriminative pedestrian area with the tailored auxiliary task, and propose the PAOA regularization for that end. To obtain a robust model in achieving "out-of-the-box" deployment, recent approaches focus on generalizing ReID to mitigate overfitting in the source domains. This is achieved through either local domain manipulation or global cross-domain alignment methods, which facilitate the extraction of domain-invariant features less prone to bias.

2.2 Learning Representation Adaptable to Target Domain

2.2.1 Domain Adaptation

Domain Adaptation (DA) is a subset of Transfer Learning (TL) that offers solutions to real-world challenges, particularly those encountered in uncontrolled environments. It achieves this by leveraging a model trained on a source dataset to perform testing on a target domain with different distributions. DA [141, 117] has gained popularity in recent years. Deep Neural Networks (DNNs) trained on one dataset (referred to as the source domain) often fail to perform well on another dataset (the target domain), even if the latter shares similar properties with the former. DA aims to mitigate this issue and shows significant potential for various applications in practical settings, real-world scenarios, and industrial domains, among others. It's achieved by transferring relevant knowledge during training. Unlike DG, where the model cannot access data from the target domain during training, DA has access to such **unlabeled** target domain data. Pan *et al* [114] introduce Transfer Component Analysis (TCA), a dimensionality reduction method for domain adaptation. The TCA aims to identify transfer components that minimize distributional differences between domains while preserving data properties. TCA addresses the domain distance reduction problem. It outperforms previous methods like Maximum Mean Discrepancy Embedding (MMDE) in terms of effectiveness and computational cost. Tzeng *et al* [135] present Adversarial Discriminative Domain Adaptation (ADDA), which is a framework that combines discriminative modeling with GAN-based loss for unsupervised domain adaptation. ADDA utilizes asymmetric mapping and domain adversarial training to mitigate domain bias and improve the generalization of datasets and tasks. Ganin *et al* [28] propose a generic domain adaptation approach, termed Domain-Adversarial Neural Network (DANN), that trains neural networks on labeled source domain data and unlabeled target domain data with similar distributions. This method promotes discriminative and domain-invariant features to enhance adaptation performance on various classification tasks. The DANN is a representative DA method. Its general

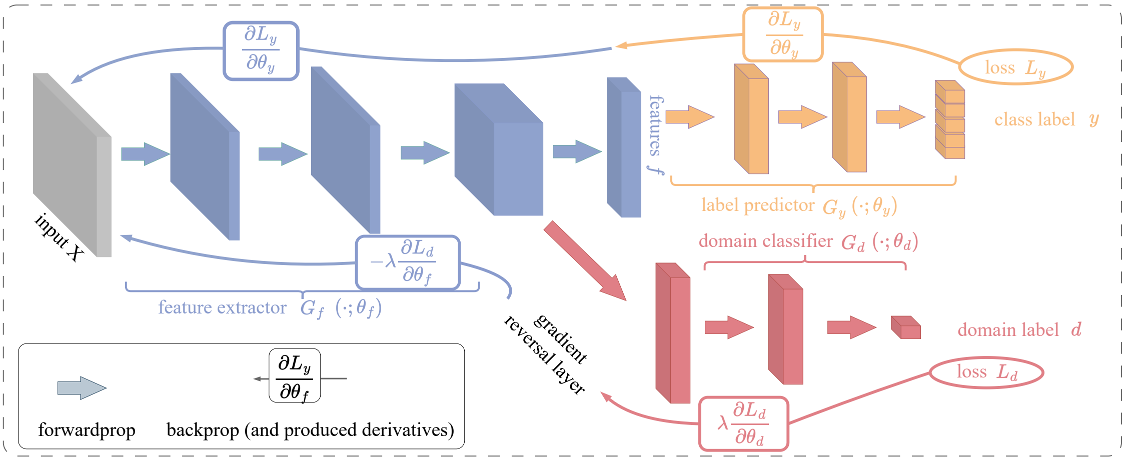

$\frac{\partial L_y}{\partial \theta_y}$
$\frac{\partial L_y}{\partial \theta_y}$
loss L_y
class label y
features f
input X
label predictor $G_y(\cdot;\theta_y)$
domain classifier $G_d(\cdot;\theta_d)$
$-\lambda\frac{\partial L_d}{\partial \theta_f}$
feature extractor $G_f(\cdot;\theta_f)$
gradient reversal layer
domain label d
loss L_d
$\lambda\frac{\partial L_d}{\partial \theta_d}$
$\frac{\partial L_y}{\partial \theta_f}$
forwardprop
backprop (and produced derivatives)

Figure 2.8: An illustration of a representative Domain Adaptation model: DANN [28]

idea is illustrated in Figure 2.8. It enables models to adapt from a labeled source domain to an unlabeled target domain. The core idea behind DANN is to learn feature representations that are not only useful for making accurate predictions in the source domain but also indistinguishable between the source and target domains. This is achieved through a two-part training process: 1. Task Prediction on Source Domain: The model is trained to accurately predict labels for samples in the source domain. This is accomplished by minimizing a classification loss, which encourages the model to learn features that are predictive of the correct labels in the source domain. 2. Domain Classification: Simultaneously, the model is trained with a domain classifier (or regressor) component, which aims to distinguish between source and target domain samples. However, the key twist is that the model is trained to *fail* at this task; that is, to make the domain classifier unable to reliably tell whether a given feature representation comes from the source or the target domain. This is realized by minimizing a domain classification loss, which uses known domain labels (source or target) for both labeled source samples and unlabeled target samples.

The overall objective combines these two aspects, with an emphasis on making the learned feature representations domain-invariant. Mathematically, this can be simplified to an optimization problem for the entire training set, rather than individual samples:

$$\min_{\theta} \left[\frac{1}{N_s} \sum_{i=1}^{N_s} \mathcal{L}_y(\mathbf{x}_i^s, y_i) + \frac{\lambda}{N_s + N_t} \sum_{j=1}^{N_s+N_t} \mathcal{L}_d(\mathbf{x}_j, d_j) \right], \tag{2.1}$$

where $\mathcal{L}_y$ denotes the classification loss on the source domain, $\mathcal{L}_d$ represents the domain classification loss, $\mathbf{x}_i^s$ and y_i are the input features and labels of the source samples, $\mathbf{x}_j$ are the input features of both source and target samples, and d_j indicates their domain labels. N_s and N_t are the numbers of source and target domain samples, respectively, and λ is a hyper-parameter balancing the two losses. In essence, DANN trains a model that excels in its primary task on the source domain while simultaneously learning to generate feature representations that a domain classifier cannot differentiate between domains. Therefore, this approach results in a model whose performance on the target domain is improved, as the features it has learned are domain-agnostic by focusing solely on the aspects relevant to the task at hand.

2.2.2 Deployment-Time Optimization

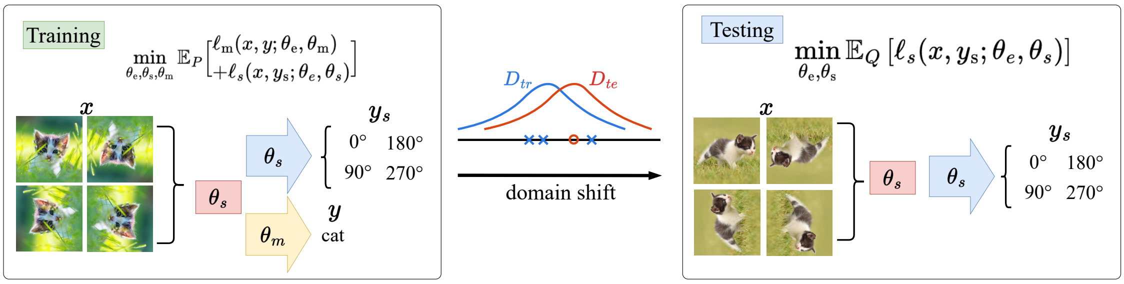

Training
$\min_{\theta_e,\theta_s,\theta_m} \mathbb{E}_P \begin{bmatrix} \ell_m(x,y;\theta_e,\theta_m) \\ +\ell_s(x,y_s;\theta_e,\theta_s) \end{bmatrix}$
x
θ_s
θ_s
θ_m
y_s
0° 180°
90° 270°
y
cat
D_{tr}
D_{te}
domain shift
Testing
$\min_{\theta_e,\theta_s} \mathbb{E}_Q [\ell_s(x,y_s;\theta_e,\theta_s)]$
x
θ_s
θ_s
y_s
0° 180°
90° 270°

Figure 2.9: An illustration of Deployment-Time Optimization with Multi-Task Learning

Deployment-Time Optimization (DTO) is an emerging paradigm to tackle distribution shifts between training and testing environments. The key idea is to perform post-training model optimization given the test samples during deployment. Several recent works[142, 139, 59, 21] propose to optimize the model parameters by providing proper supervision, such as batch-norm statistics, entropy minimization, and pseudo-labeling. This line of methods is not directly applicable to ReID due to the zero-shot nature, in which the soft-max logits are in different label space during training and testing. In contrastive, another line of works [132, 98] jointly trains additional self-supervised auxiliary tasks, in a Multi-Task Learning (MTL) design. These auxiliary tasks can be used to guide the model optimization during testing. This does not involve any assumptions about the output and is more generic. In this realm, the objective is often to optimize a model in such a way that it can efficiently learn from multiple tasks simultaneously. This not only enhances the model's performance on the main task but also leverages auxiliary tasks to guide the model's optimization during testing, especially when labels for the main task are unavailable. This concept is illustrated in Figure 2.9, and formalized through the following set of equations, Considering the conventional single-task learning scenario depicted by:

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell_m(x_i, y_i; \boldsymbol{\theta}), \tag{2.2}$$

where $\boldsymbol{\theta}$ represents the parameters of the model, n is the number of samples, and $\ell_m(x_i, y_i; \boldsymbol{\theta})$ denotes the loss function for the main task, with (x_i, y_i) being the input-label pairs for the training samples. This equation represents the goal of minimizing the average loss across all samples for a single main task. In the context of Deployment-Time Optimization that includes an auxiliary task, the objective extends to:

$$\min_{\boldsymbol{\theta}_e, \boldsymbol{\theta}_m, \boldsymbol{\theta}_s} \frac{1}{n} \sum_{i=1}^{n} \ell_m(x_i, y_i; \boldsymbol{\theta}_m, \boldsymbol{\theta}_e) + \ell_s(x_i, y_s; \boldsymbol{\theta}_s, \boldsymbol{\theta}_e), \tag{2.3}$$

with $\boldsymbol{\theta}_e$, $\boldsymbol{\theta}_m$, and $\boldsymbol{\theta}_s$ representing the parameters of the feature extractor, main task model, and auxiliary task model, respectively. ℓ_m and ℓ_s are the loss functions for the main and self-supervised auxiliary tasks. This framework aims to minimize the combined losses of the main and auxiliary tasks, allowing the model to learn from both simultaneously. At test time, in the absence of main task labels, optimization is focused solely on the auxiliary task:

$$\min_{\boldsymbol{\theta}_e} \ell_s(x, y_s; \boldsymbol{\theta}_s, \boldsymbol{\theta}_e). \tag{2.4}$$

In such a way, the model can adapt to new domains or scenarios with the auxiliary task by optimizing the feature extractor ($\boldsymbol{\theta}_e$) to ensure versatility and effectiveness. DTO has also been applied to Re-Identification (ReID) [39] by considering self-supervised learning tasks for updating BN statistics. In Chapter 4, we formulate the Primary-Auxiliary Objectives Association (PAOA)+ method by incorporating the proposed PAOA regularization into the DTO framework to seek further improvement. With the tailored auxiliary objective as the optimization supervision, PAOA+ effectively exploits the underlying target domain characteristic and exhibits boosted performance on all the benchmarks.

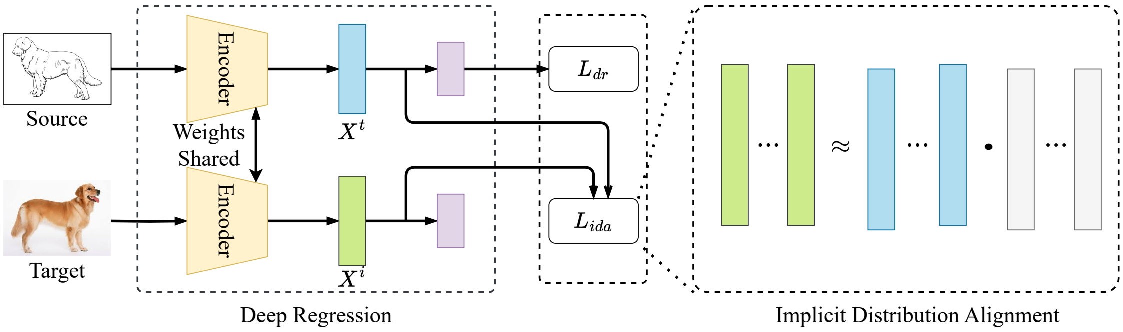


Figure 2.10: The schematic diagram of distribution alignment.

2.3 Learning Representation Generalizable to Unseen Domain

Domain Generalization (DG) [186] is a machine learning strategy designed to overcome the challenge of models underperforming on new, unseen domains. The goal of DG is to empower a model with the ability to achieve robust performance across different domains without requiring retraining for each new domain encountered. This is accomplished by training the model on a variety of domains, thereby to encourage the development of more adaptable feature representations. As a result, the model's dependence on domain-specific data is diminished, so as to enhance its ability to generalize to new domains. By broadening the model's exposure to diverse data during the training phase, DG helps ensure that the model remains effective and reliable even when introduced to unfamiliar data environments. To achieve Domain Generalization, numerous studies have been conducted from the following aspects.

2.3.1 Distribution Alignment

The idea of distribution alignment aims to minimize the feature discrepancy between source and target domains. The schematic diagram of distribution alignment operation is shown in Figure 2.10. It generally designs a dedicated optimization objective to constrain the learned feature distribution, such as by implicitly assume the feature distribution in different domains are are linearly correlated. It has been widely adopted by DA models to align the distribution of per domain learned representation to the target one. However, for DG, it is inherently incapable of directly conducting such a "target-oriented" alignment due to the absence of target data during model training. With a straightforward assumption that features which are invariant to the source domain shift should also be invariant to any unseen target domain [73], DG approaches share the spirit to minimize the discrepancy among source domains to achieve distribution alignment. There are a wide variety of statistical metrics available for minimizing, such as Euclidean distance and f-divergences. In this regard, Li *et al* [74] propose to minimize the Kullback–Leibler (KL) divergence of source domain features with a Gaussian distribution. Several works achieve distribution alignment by minimizing a single moment (mean or variance) [109, 55] or joint moments [23, 29] calculated over a batch of source domain samples through either a projection matrix [29] or a non-linear deep network [63]. Moreover, minimizing contrastive loss offers another avenue for mitigating distribution discrepancies [108, 103]. These methods leverage semantic labels to pull together the anchor and positive groups while simultaneously pushing the anchor away from negative groups. Li *et al* [73] minimize the Maximum Mean Discrepancy (MMD) distance by aligning the source domain feature distributions with a prior distribution via adversarial training [31]. Unlike explicit distance metrics MMD, adversarial learning formulates the problem of distribution minimization through a minimax two-player game. Warde *et al* [146] justify that generative adversarial learning is tantamount to minimize the Jensen-Shannon divergence between the real and generated distributions. Alternatively, Variational Autoencoder (VAE) [58] can be employed to model a normal distribution and create a shared space, thus aligning the learned features. The loss function of VAE consists of two terms: the reconstruction loss and the regularization term. The reconstruction loss measures the discrepancy between the decoder's output and the input data. It is typically measured using reconstruction error metrics, such as mean squared error (MSE) or Cross-Entropy (CE). The regularization term penalizes the deviation of the latent representation, ensuring that the sample distribution in the latent space is close to the unit normal distribution. The alignment achieved by VAE is formulated as:

$$\begin{aligned}\ell(\phi,\theta,x) &= \ell_{\text{recon}} + \ell_{\text{KL}},\\ &= \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{X}_i)^2 + KL[G(Z_\mu, Z_\sigma), \mathcal{N}(0,1)],\end{aligned} \tag{2.5}$$

where ℓ_{recon} represents the reconstruction loss, and KL stands for KL divergence. The diagram of VAE alignment is shown in Figure 2.11. In chapter 5, the proposed global distribution calibration operates on the same principle to align the source domains in learning a domain-agnostic model. In contrast, we tailor the alignment objective for person ReID considering that all samples are depicting pedestrians, rather than predefine a deterministic distribution to align, *e.g* Gaussian or Laplace distributions. Specifically, we constructed a common feature space upon the ID prototypical representations stored in a global memory bank to eliminate domain-biased information.

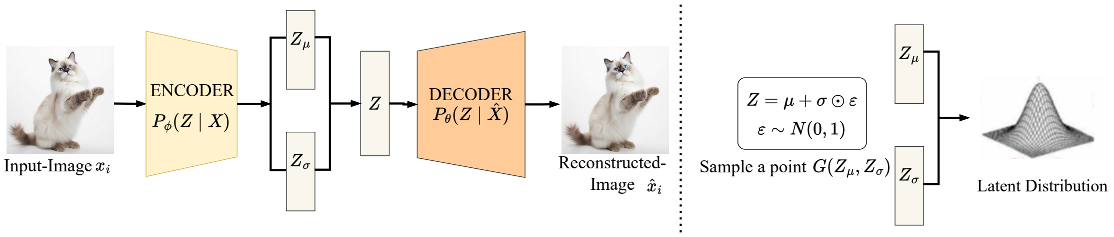

Input-Image x_i
ENCODER
$P_\phi(Z \mid X)$
Z_μ
Z_σ
Z
DECODER
$P_\theta(Z \mid \hat{X})$
Reconstructed-Image $\hat{x}_i$
$Z = \mu + \sigma \odot \varepsilon$
$\varepsilon \sim N(0,1)$
Sample a point $G(Z_\mu, Z_\sigma)$
Z_μ
Z_σ
Latent Distribution

Figure 2.11: Structure of a Variational Autoencoder that aligns the latent vector with a normal distribution.

Disentangled Representation Learning. The objective of disentanglement learning is to explore the distinct and explanatory components, and decouples a representation into domain-invariant and domain-specific parts. It is generally achieved through adversarial training [28] where the aim is to deceive a domain discriminator, to enable the learned features to be domain-agnostic. For example, Wang *et al* [140] employs an adversarial learning approach to train two independent encoders. These encoders were designed to capture identity and domain information separately, and encoders aim to address cross-domain facial anti-spoofing. Disentangled Representation Learning has also found widespread application in style and content disentanglement. This method aims to separate the style (*e.g* color, texture, brushstrokes) of an image from its content (*e.g* objects, structure, layout), enabling independent manipulation of both. The schematic diagram of different feature disentanglement methods is illustrated in Figure 2.12. This separation enables networks to combine the style of one image with the content of another, create images with new styles, or generate different styles of the same content by changing the style parameters. For example, Kotovenko *et al* [69] propose a method that captures the nuances of style and variations within it. The method can separate style from content. Disentanglement has also been studied to generalize person ReID. For instance, EOM [22] designs a disentanglement module incorporating a cycle-consistency constraint, while Zhang [171] *et al* construct a structural causal model to approximate the shifted distribution and pursue the causality between identity-specific factors and identity labels. However, it remains uncertain whether the disentanglement criteria and the model is susceptible to learning less discriminative representations when a significant domain shift occurs [107, 8]. In chapter 6, we design a disentanglement module constrained by maximizing the consensus of domain-shared knowledge to learn an identity prototype that is domain-agnostic.

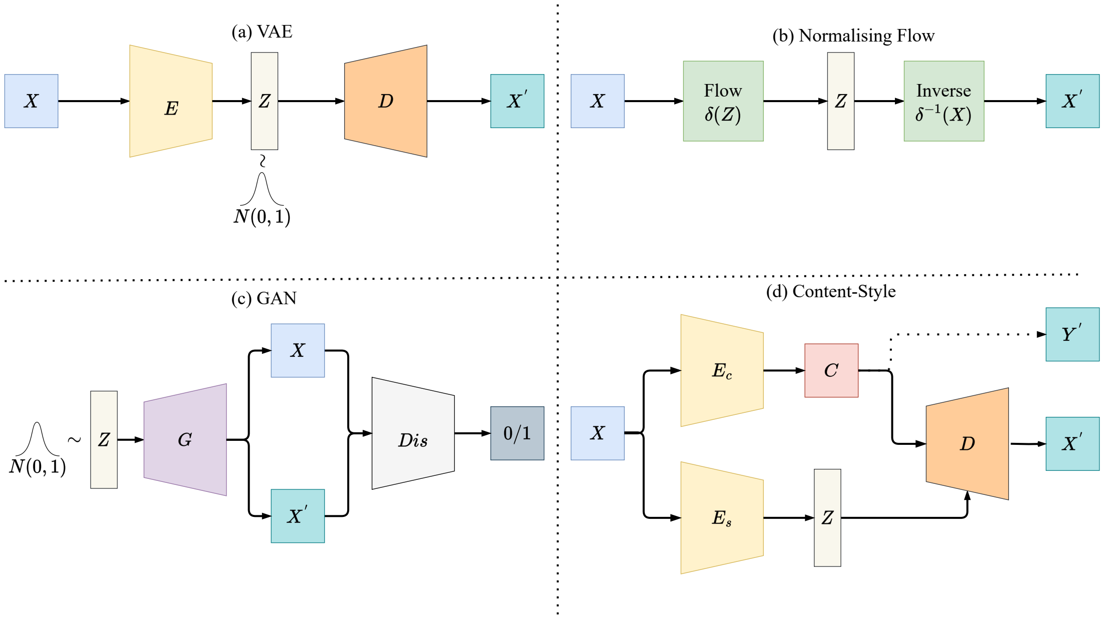


Figure 2.12: Comparison on different feature distribution manipulation methods, including (a) distribution alignment by VAE, (b) distribution normalization by Normalizing Flow. (c) feature disentanglement by GAN. (d) style-content disentanglement by Adversarial Training.

2.3.2 Data Augmentation

Training a neural network with diverse data can improve its generalizability on new, unseen domains [66, 3], and further improve its robustness against spurious correlations. Data augmentation [125] serves as a cost-effective method to enrich data diversity. It is a widely used technique in computer vision and ML to increase the diversity of training data. It involves applying transformations or perturbations to the original data. Data augmentation aims to improve the generalization ability of models, reduce overfitting, and make them more robust to different scenarios and conditions. Traditional data augmentations [165, 38, 44, 159] are most commonly applied within the raw image space, often through geometric transformations or random erasing. A few augmented applied on the raw image space examples are illustrated in Figure 2.13 (a). The emergence GANs [32] has enabled the generation of new, realistic augmented counterparts featuring different contents or styles. The conventional paradigm of data augmentation is to diversify data. Yang *et al* [159] design an image augmentation module which helps the network to learn domain-invariant representation by distilling information learned from the augmented samples to the teacher network. More recently, feature augmentation has emerged as a more

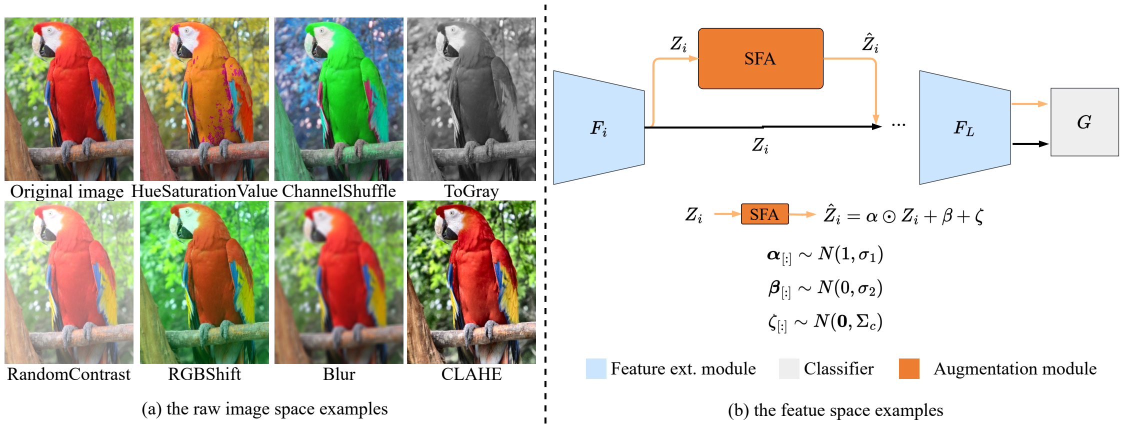
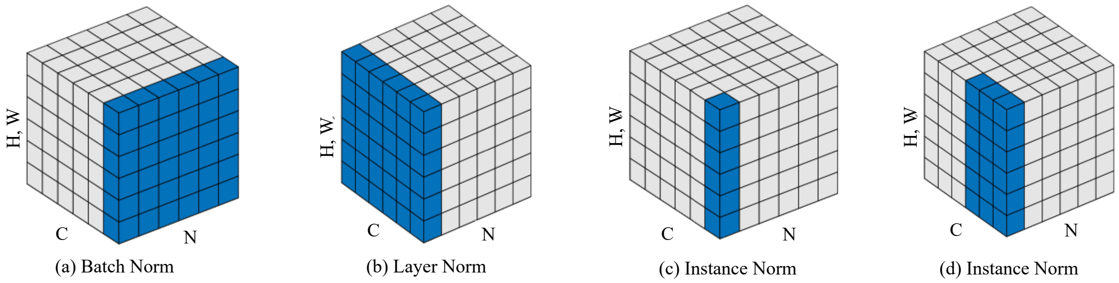

Original image
HueSaturationValue
ChannelShuffle
ToGray
RandomContrast
RGBShift
Blur
CLAHE
(a) the raw image space examples
Z_i
SFA
$\hat{Z}_i$
F_i
Z_i
...
F_L
G
Z_i
SFA
$\hat{Z}_i = \alpha \odot Z_i + \beta + \zeta$
$\alpha_{[:]} \sim N(1, \sigma_1)$
$\beta_{[:]} \sim N(0, \sigma_2)$
$\zeta_{[:]} \sim N(\mathbf{0}, \Sigma_c)$
Feature ext. module
Classifier
Augmentation module
(b) the featue space examples

Figure 2.13: Comparison on augmentation in raw image space and feature space.

effective transformations. Different from augmentation on the raw image, feature augmentation is directly applied on the holistic representation space. The diagram of performing feature distribution augmentation [81] in shown Figure. 2.13 (b). In this regards, DeepAugment [43] perturbs features via stochastic operations by forwarding images through a pre-trained image-to-image model, to generate semantically meaningful and diverse samples. Li *et al* [81] discover that embedding white Gaussian noise in high-dimensional feature space provides substantive statistics reflective of cross-domain variability. Li *et al* [88] propose to model the feature uncertainty with a multivariate Gaussian distribution to perturb hierarchical features to diversify the feature space. In chapter 5, we explore feature distribution augmentation in each source domain to achieve per-domain feature distribution diversification rather than diversifying the data, with the objective of making the model invariant to per-domain holistic shift to avoid model overfitting in each source domain. In Chapter 6, we model the cross-domain style variations and employ them to augment the identity prototype, providing diverse pedestrian styles to achieve per-identity multi-view augmentation. By simulating the identity cross-view discrepancy, the trained model is robust in extracting domain-unbiased representations during testing.

2.3.3 Feature Normalization

Feature normalization, known as feature scaling or feature standardization, is commonly applied to feature maps before feeding it into subsequent layers. It mainly includes Batch Normalization, Instance Normalization, Layer Normalization, and Group Normalization. Their diagrams and comprehensive comparisons are illustrated in Figure 2.14. Among them, Batch Normalization



(a) Batch Norm (b) Layer Norm (c) Instance Norm (d) Instance Norm

Figure 2.14: Comparison of different feature normalization techniques.

(BN) and Instance Normalization (IN) have been widely employed in Domain Generalization. to improve model generalizability to unseen domains. BN [121] is a well-established technique that improves training stability and accelerates convergence. This normalization process involves centering and scaling the activations of each layer within a mini-batch, thereby mitigating the internal covariant shift problem. This arises from the changing distribution of activations as the network learns. BN introduces learnable parameters, termed scale γ and shift β parameters, which allow the network to adaptively adjust the normalized activations. As a result, it contributes to standardizing the training process so that it is less sensitive to initialization choices and enables the use of higher learning rates. Given a mini-batch of activations $X = x_1, x_2, ..., x_n$, where n is the batch size. The process of performing BN is:

$$\begin{aligned} \mu_B &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{N \times H \times W}\sum_{c=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{i,c,h,w}\right), \\ \sigma_B^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{N \times H \times W}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} (x_{i,c,h,w} - \mu_B)^2\right), \end{aligned} \tag{2.6}$$

where μ_B is the mean of the batch activations, σ_B^2 is the variance. Therefore, the activation is shifted as:

$$\begin{aligned} \hat{x}_{i,c,h,w} &= \frac{x_{i,c,h,w} - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \\ y_{i,c,h,w} &= \gamma_c \hat{x}_{i,c,h,w} + \beta_c, \end{aligned} \tag{2.7}$$

where γ_c and β_c are learnable scale and shift parameters for each channel c, and ε is a small constant added for numerical stability to avoid division by zero. IN [57] is a variant aimed at reducing style variation by holistically shifting per-instance activation moments. Compared to BN, Instance Normalization (IN) has been widely used in style transfer and image generation [57]. It can be used to eliminate illumination and color variations in images and makes the model more robust and stable. In style transfer tasks, IN can help to extract the style of one image and apply it to another image, which facilitates style transfer. acrshortin operators as:

$$\begin{aligned} \mu_i &= \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} x_{i,c,h,w}, \\ \sigma_i^2 &= \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} (x_{i,c,h,w} - \mu_i)^2, \end{aligned} \tag{2.8}$$

where H and W represent the height and width of the feature map, respectively. $x_{i,c,h,w}$ denotes the pixel value at position (c,h,w) in feature map i. Then, Instance Normalization scales and shifts the normalized feature values:

$$\begin{aligned} \hat{x}_{i,c,h,w} &= \frac{x_{i,c,h,w} - \mu_i}{\sqrt{\sigma_i^2 + \varepsilon}}, \\ y_{i,c,h,w} &= \gamma_c \hat{x}_{i,c,h,w} + \beta_c. \end{aligned} \tag{2.9}$$

IN [64, 62] in conjunction with BN [111, 17] have also been studied in ReID models to eliminate style information associated with identity. However, considering that a specific identity captured by disjoint cameras showcases distinct styles [156], a model trained solely with normalized features is limited in extracting a discriminative representation against such style variations. Additionally, IN dilutes essential complementary information that is crucial for general visual recognition, therefore it is suboptimal. In Chapter 6, we capture the distinct domain-specific variations by computing the statistical moments and utilize them diversify the style of a singular identity. This approach aims to achieve per-identity multi-view (style) augmentation, to further result in improved model robustness.

2.4 Multi-Task Learning for Inter-task Representation Association

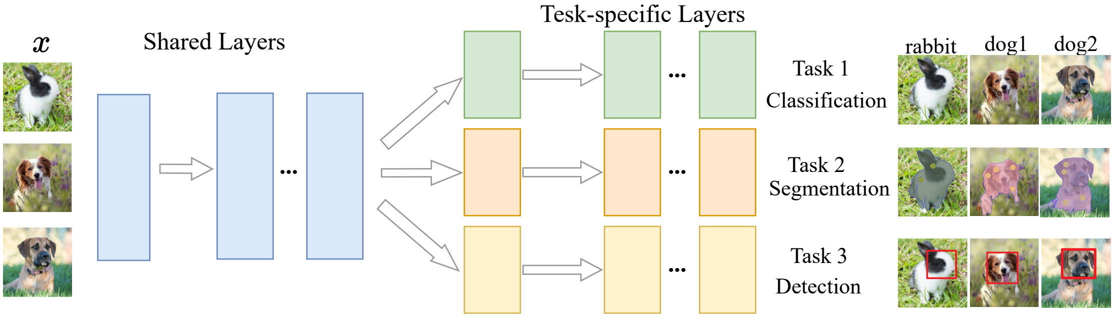


Figure 2.15: An illustration of a Multi-Task Learning framework.

Multi-Task Learning (MTL) [172] is presented as a strategy to train a single model that can be applied across multiple tasks, which offers greater efficiency compared to the training of individual models for each task separately. A Deployment-Time Optimization (DTO) framework [132, 98] usually optimize a MTL architecture during model training. The general design of a MTL framework involves a shared feature extractor and several individual task heads. The idea of MTL is illustrated in Figure 2.15, in which three tasks are involved, including classification, segmentation, and detection. The shared feature extractor is responsible for deriving universal and beneficial representations from the input data, whereas the individual task heads utilize these common representations to perform predictions or classifications for their respective tasks. We denote the output of the shared feature extractor for any given input data x as $f_{\text{shared}}(x)$, which is typically realized through a combination of CNN, RNN, or various other neural network designs. For any given task i, the output of the task-specific head associated with task i is expressed as $f_i(f_{\text{shared}}(x))$, where $f_i(\cdot)$ is a function tailored to adapt the shared features for predictions or classifications relevant to task i. The formulation of the loss function for MTL is outlined as follows:

$$\mathcal{L} = \sum_{i=1}^{N} \lambda_i \mathcal{L}_i(f_i(f_{\text{shared}}(x)), y_i), \tag{2.10}$$

where $\mathcal{L}$ is the total loss function, $\mathcal{L}_i(\cdot)$ is the loss function specific to task i calculated over the predictions from the task-specific head $f_i(\cdot)$ and the corresponding ground truth y_i. The λ_i is a coefficient that controls the importance of task i, and N is the total number of tasks. The $\mathcal{L}_i(\cdot)$ can vary depending on the nature of each task. For example, In classification tasks, $\mathcal{L}_i(\cdot)$ could be cross-entropy loss. In regression tasks, $\mathcal{L}_i(\cdot)$ could be MSE loss. In this diagram, closely related tasks can enhance each other's learning by exchanging information and utilizing shared features. This collaborative learning approach often results in enhanced performance on individual tasks compared to training them in isolation. Nonetheless, recent research [163] has highlighted that conflicting gradients during MTL can degrade performance, potentially due to noisy labels. To break this condition and achieve positive interactions between tasks, research work [163] propose to de-conflict such gradients by altering their directions towards a common orientation. In Chapter 4, we design the PAOA model within a MTL framework, in which we incorporate a complementary pedestrian saliency detection task as an auxiliary task alongside the primary pedestrian classification task, and both tasks are concurrently optimized during training. In contrast to conventional MTL designs where tasks are typically considered within the same hierarchy, we design the auxiliary task specifically to support the main task. To this end, we propose referenced gradient calibration by setting the main task as the reference, and calibrating the auxiliary gradient towards it, so as to ensure the auxiliary task can be harmoniously trained alongside the main task, and provide supervision for the primary model objective.

Salient Object Detection [7] aims to identify objects or regions that are visually more attentive than the surrounding areas. It has been significantly boosted solely by the rapid development of deep learning. Current detection models are usually trained end-to-end and output a fine-grained saliency map at the pixel level. In Chapter 4, we formulate PAOA that incorporate the auxiliary task with a pedestrian saliency detection objective. Instead of exhaustively labeling the pedestrian area manually as the previous work [127], we propose to use weakly labeled data generated by a pretrained salient object detection model on the fly. The recent work GASM [41] shares a similar spirit to ours by employing weakly labeled saliency masks as an additional prior. However, GASM simply trains the saliency detection layers with the classification network while omitting the potential worst-case where the weak label is not accurate and causes potential conflict optimization direction during model training. In contrast, PAOA focuses on the association between instance classification and saliency detection objectives by the proposed referenced gradient calibration mechanism, which promotes the learning of the primary objective while mitigating the conflicts between the primary and auxiliary tasks.

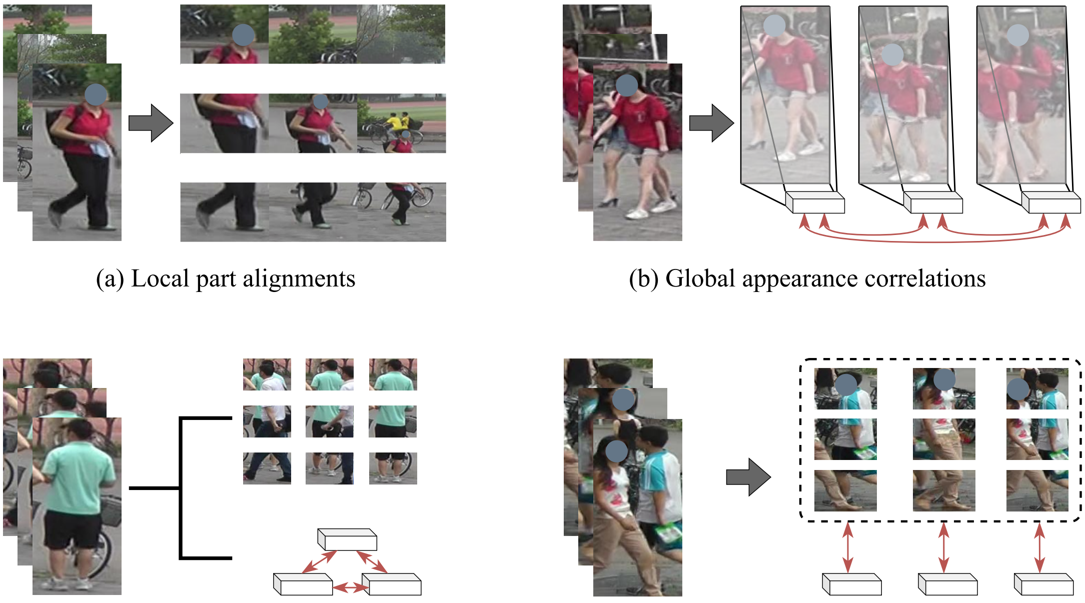
Chapter 3

Local-Global Associative Frame Assemble in Video ReID

3.1 Introduction

Early ReID studies [11, 149, 143, 150] concentrate on exploring appearance patterns unique per identity from still images [27, 87, 160], which has shown remarkable discrimination capacity. However, such methods assume well-curated data and the identity information is preserved in images. This assumption dramatically restricts their scalability and usability to many practical application scenarios when uncontrollable environments are the norm not the exception where video data are captured [80, 97].

In the literature, one of the most commonly adopted techniques for assembling identity information from different video frames is *averaging* by pooling [130, 126]. By assuming all the frames are of equal importance, the pooling method neglects their diverse qualities caused by the constantly changing environments and/or unreliable pedestrian detections. Therefore, the aggregated tracket's representations are likely impacted by various types of noise as shown in Figure 3.1. In order to *selectively* assemble video frames rather than averaging, attention mechanisms [53, 137, 144, 148, 56] have been studied to explore the correlations between the global visual features of frames (Figure 3.1 (b)) so that the common appearance patterns shared among frames in the same tracklet are maintained while removing/ignoring unusual and low-quality frames [89, 105, 95, 145]. In contrast to the global appearance correlations, an alternative approach [173, 51, 50] compares video frames by local parts (Figure 3.1 (a)) so to identify outliers that are significantly misaligned with other frames in a tracklet. Although sharing the same ob-



(a) Local part alignments

(b) Global appearance correlations

(c) Local-global Joint

(d) Local-global Association

Figure 3.1: An illustration of four types of quality assessment strategies for frame assembling.

jective to adaptively assemble only the relevant video frames, these two approaches differ in exploiting information in different granularities. In isolation, both are sub-optimal in different real-world video scenes. The local-parts approach is fragile if the detected pedestrians are not well-aligned while the global-appearance approach is spatially insensitive, tending to miscorrelate patterns of interest in the background. Beyond attentive assembling, Recurrent Neural Network (RNN) [157, 106] has also been exploited for modeling temporal information to represent frame sequences in video tracklets. However, this approach is also vulnerable to noisy frames without careful frame selections [151].

To learn robust and discriminative representation with high-quality frames, we propose a tracklet frame assembling approach to video person ReID termed Local-Global Associative Assembling (LOGA). As shown in Figure 3.1 (d), the LOGA method adaptively assembles video frames in the same tracklets by a Local Aligned Quality (LAQ) and a Global Correlated Quality (GCQ) modules to assess the importance/relevance of the frames by both their alignments in the local part and global appearance correlations as well as their mutual reinforcements. Moreover, the LOGA model constructs a local-assembled global appearance prototype to not only take the advantage of two types of information but also complement each other mutually by learning their consensus. Whilst the focus of most existing spatial-temporal attentive methods is on collaborat-

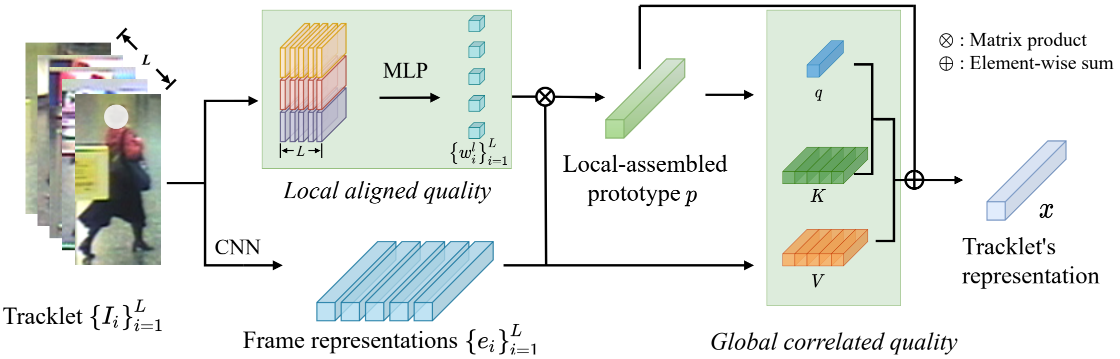

⊗ : Matrix product
⊕ : Element-wise sum
L
MLP
$|\leftarrow L \rightarrow|$
$\{w_i^l\}_{i=1}^L$
Local aligned quality
CNN
Local-assembled prototype p
q
K
V
Global correlated quality
x
Tracklet's representation
Tracklet $\{I_i\}_{i=1}^L$
Frame representations $\{e_i\}_{i=1}^L$

Figure 3.2: An overview of the proposed LOGA model.

ing the temporal information with intra-frame spatial attention, we aim to exploit the inter-frame complements more effectively, which is different and ready to benefit from the advancing per-frame learning. Specifically, the LAQ module divides all video frames in a tracklet into the same set of spatial parts and assesses each frame's quality by their part-wise alignment to the other frames so to measure both inter-frame visual similarity and spatial alignment. On the other hand, the GCQ module is applied on the holistic feature representation of each frame to consider inter-frame global appearance correlations, which is more robust to local part misalignment but spatially insensitive so less reliable from mis-correlation of information, *e.g* irrelevant patterns in the background. Furthermore, to associate the local and global information and exploit their mutual benefits, we take the tracklet's representation assembled by the LAQ as its prototype and compare the global visual feature of frames with it in the GCQ module so that the two modules are encouraged to find a trade-off between the local and global information to cope with different types of noise more reliably.

Extensive experiments show the performance advantages and superior robustness of the proposed LOGA model over the State-of-the-Art (SOTA) video ReID models on four video ReID benchmarks MARS [178], Duke-Video [120, 152], Duke-SI [80], and iLIDS-VID [143].

3.2 Methodology

3.2.1 An overview

To learn robust and discriminative representation from high-quality frames, we propose a LOGA model to selectively exploit information from video frames in the same tracklets according to both their local part alignments and global appearance correlations as well as the synergy and mutual promotion of these two types of information. For notation clarity, in the following, we focus on the formulation of assembling frames $\{\boldsymbol{I}_i\}_{i=1}^L$ in a single video tracklet $\boldsymbol{T}$ and ignore its tracklet index. As shown in Figure 3.2, the video tracklet is first fed into a LAQ module to assess the quality of frames regarding their part-wise alignment:

$$\{w_i^l\}_{i=1}^L = f_{\theta_l}(\{\boldsymbol{I}\}_{i=1}^L). \tag{3.1}$$

The θ_l in Eq. (3.1) is the learnable parameters of the LAQ and w_i^l denotes the importance of frames $\boldsymbol{I}_i$ determined by its alignments with other frames in local parts. Then, a GCQ module is devised which is applied to the D-dim holistic visual representation $\boldsymbol{E} = \{\boldsymbol{e}_i\}_{i=1}^L \in \mathbb{R}^{D\times L}$ of frames to determine their global appearance correlations. Instead of focusing on only the global visual features that are prone to spatial-insensitive miscorrelation, we explore the mutual synergy between local and global information by associating LAQ and GCQ through a prototypical descriptor $\boldsymbol{p}$. This assembles a frame's global features by their local-parts quality in GCQ for correlation exploration:

$$\boldsymbol{p} = \sum_{i=1}^{L} w_i^l \boldsymbol{e}_i, \tag{3.2}$$

$$\boldsymbol{x} = f_{\theta_g}(\{\boldsymbol{e}_i\}_{i=1}^L | \boldsymbol{p}), \tag{3.3}$$

where $f_{\theta_g}(\cdot)$ denotes global-appearance quality assessment on $\boldsymbol{E}$, and $\boldsymbol{x}$ is the representation of a tracklet $\boldsymbol{T}$ assembled by associating LAQ and GCQ through $\boldsymbol{p}$. With the tracklet-level representations, a generic distance metric (*e.g* cosine distance) is used to measure the pairwise visual similarity of tracklets for video ReID matching.

3.2.2 Local Aligned Quality

To explore the visual similarity of frames in terms of their local alignments, we separate them uniformly into M non-overlapping patches (parts) and apply patch-wise cross-frame convolution to recognize the aligned local patterns. This is accomplished by first flatten the 2D frames $\{\boldsymbol{I}_i\}_{i=1}^L$ then stacking them in the channel dimension as the raw representation of the tracklet $\boldsymbol{T}$ maintaining the inter-frames spatial correspondence. An 1D convolution is then applied on $\boldsymbol{T}$ to explore the per-part visual patterns,

$$\tilde{\boldsymbol{w}}^l = \boldsymbol{F} * \boldsymbol{T}, \quad \boldsymbol{F} \in \mathbb{R}^{S\times L\times L}, \tag{3.4}$$

where $*$ denotes the 1D convolution function and $\boldsymbol{F}$ is a trainable kernel. The size S of kernel $\boldsymbol{F}$ is determined by the granularity of the spatial separation, *i.e*, $S = \frac{H\times W}{M}$ where H and W are

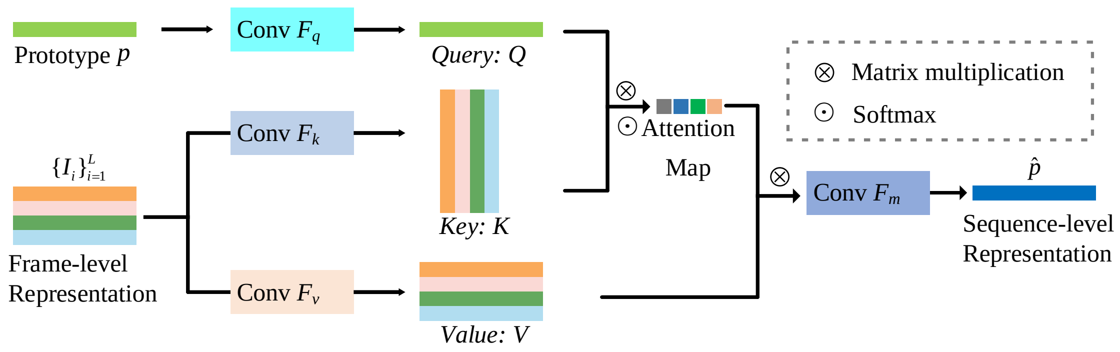


Figure 3.3: An overview of the proposed GCQ module.

the height and width of frames, respectively. The computed results $\tilde{\boldsymbol{w}}^l \in \mathbb{R}^{M\times L}$ encode the part-wise importance of every frame, which is then aggregated by pooling followed by a multi-layer perceptron (MLP) to obtain the per-frame scores:

$$\boldsymbol{w}^l = \mathrm{Softmax}(\mathrm{MLP}(\mathrm{Pooling}(\tilde{\boldsymbol{w}}^l))) \in (0,1)^{L\times 1}. \tag{3.5}$$

The Pooling$(\cdot)$ in Eq. (3.5) is a frame-wise mean pooling function and the MLP$(\cdot)$ stands for a single layer MLP activated by a ReLU function. The resulted scores are then normalized by softmax function as the indication $\boldsymbol{w}^l$ of per-frame importance to the tracklet $\boldsymbol{T}$. In this way, the LAQ learns to assess the frame's quality by its local part alignments to other frames, so as to identify the misaligned outlier frames and suppress them from representing a tracklet.

3.2.3 Global Correlated Quality

The GCQ module, as demonstrated in Figure 3.3, is formulated to explore the inter-frame correlations according to their global appearances. However, the spatial invariant characteristic of the CNN features tends to miscorrelate patterns of interests with potential noise in the background, *i.e* completely ignoring the spatial part's alignment. In this case, we propose to establish the GCQ on the results yielded by LAQ so to associate them by their synergy. Specifically, given the frame's importance $\boldsymbol{w}^l$ computed by Eq. (3.5) regarding their local part alignments, we first assemble their visual features accordingly in Eq. (3.2), which serves as the appearance prototype $\boldsymbol{p}$ of a tracklet. Then, the global-appearance quality of a frame is estimated according to the correlation between their global features and the prototype:

$$\begin{aligned} \boldsymbol{q} = f_{\theta_q}(\boldsymbol{p}) \in \mathbb{R}^{D\times 1}, \quad \boldsymbol{K} = f_{\theta_k}(\boldsymbol{E}) \in \mathbb{R}^{D\times L} \\ \boldsymbol{w}^g = \text{Softmax}(\boldsymbol{K}^\top \boldsymbol{q}) \in (0,1)^{L\times 1}. \end{aligned} \tag{3.6}$$

The f_{θ_q} and f_{θ_k} functions in Eq. (3.6) are to linearly transform respectively the prototype and frame's features. Both are followed by batch normalization. Given the global appearance quality of frames, their visual features can be selectively aggregated by:

$$\boldsymbol{V} = f_{\theta_v}(\boldsymbol{E}) \in \mathbb{R}^{D\times L}, \quad \hat{\boldsymbol{p}} = \boldsymbol{V}\boldsymbol{w}^g \in \mathbb{R}^{D\times 1}, \tag{3.7}$$

where f_{θ_v} is identical to f_{θ_q} and f_{θ_k} in Eq. (3.6) with independent parameters θ_v. Rather than taking $\hat{\boldsymbol{p}}$ as the final representation of the tracklet $\boldsymbol{T}$, in light of the residual learning [40], we distill the complementary information from global appearance correlations of frames to enhance the prototype computed by local-parts quality so to minimize representational error from identity-irrelevant part misalignments. To that end, we further learn the residual of $\boldsymbol{p}$ from $\hat{\boldsymbol{p}}$ and obtain the visual feature representation of $\boldsymbol{T}$ by:

$$\boldsymbol{x} = \boldsymbol{p} + \text{FC}(\hat{\boldsymbol{p}}) \in \mathbb{R}^{D\times 1}. \tag{3.8}$$

This design of the GCQ module not only explores the global features of frames but also considers their local part alignments for optimizing a discriminative tracklet representation.

3.2.4 Model Training

Given the formulations of LAQ and GCQ, the proposed LOGA model can benefit from conventional learning supervision. Specifically, the LOGA model is jointly trained with a softmax Cross-Entropy (CE) loss $\mathcal{L}_{\text{id}}$ and a triplet ranking loss $\mathcal{L}_{\text{trip}}$ [45]. The softmax CE loss $\mathcal{L}_{\text{id}}$ is employed to optimize identity classification:

$$\tilde{\boldsymbol{y}}_i = \text{Softmax}(\text{FC}(\boldsymbol{x}_i)), \quad \mathcal{L}_{\text{id}}(\boldsymbol{T}_i) = -\sum_{j=1}^{C} y_{i,j} \log \tilde{y}_{i,j}. \tag{3.9}$$

The $\boldsymbol{y}_i$ in Eq. (3.9) is an one-hot indicator of the ground-truth identity of tracklet $\boldsymbol{T}_i$ and the FC$(\cdot)$ serves as a linear classifier which maps the tracklet's representation $\boldsymbol{x}_i$ into an identity prediction distribution $\tilde{\boldsymbol{y}}_i$ while C is the total number of identities. Moreover, the triplet ranking loss $\mathcal{L}_{\text{trip}}$ explicitly draws the features of a positive tracklet pair sharing the same identity closer in the learned latent space while pushing the negative pairs apart:

$$\mathcal{L}_{\text{trip}}(\boldsymbol{T}_i) = \max(0, \Delta + \mathcal{D}(\boldsymbol{x}_i, \boldsymbol{x}_i^+) - \mathcal{D}(\boldsymbol{x}_i, \boldsymbol{x}_i^-)), \tag{3.10}$$

where $\boldsymbol{x}_i^+$ and $\boldsymbol{x}_i^-$ are the representations of two randomly sampled tracklets with the same and different ground-truth labels as $\boldsymbol{x}_i$ in respective, $\mathcal{D}(\cdot,\cdot)$ measures the distance of two features and Δ is a predefined margin. The overall optimization objective of a batch of tracklets is then formulated by combining the two losses as:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(\mathcal{L}_{\text{id}}(\boldsymbol{T}_i) + \mathcal{L}_{\text{trip}}(\boldsymbol{T}_i)), \tag{3.11}$$

where n is the size of a mini-batch. Since the objective function Eq. (3.11) is differentiable, the LOGA model can be trained end-to-end by the conventional stochastic gradient descent algorithm in a batch-wise manner. The overall training process is depicted in Algorithm 1.

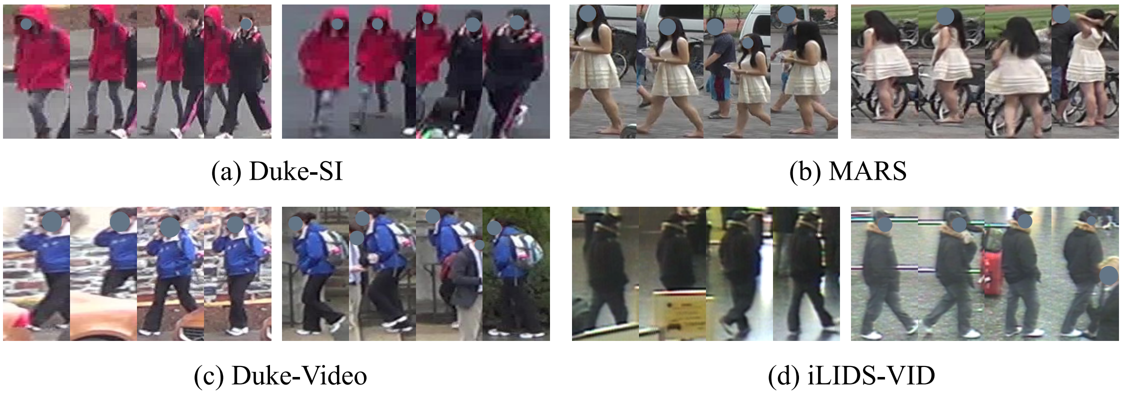
Algorithm 1 Local-Global Associative Assembling (LOGA).

```
Input: Video tracklets $\mathcal{T}$, Identity labels $\mathcal{Y}$.
Output: A deep CNN model for video person ReID.
for $i = 1$ to $max\_iter$ do
  Randomly sample a mini-batch of video tracklets from $\mathcal{T}$ and their identity labels from $\mathcal{Y}$.
  Compute the local-aligned per-frame importance scores (Eq. (3.5)).
  Feed the tracklets into the backbone network to obtain their holistic visual features $\boldsymbol{E}$.
  Compute the local-assembled global appearance prototype (Eq. (3.2)).
  Compute the global-correlated per-frame importance scores (Eq. (3.6)).
  Compute the tracklet-level representations (Eq. (3.7) and Eq. (3.8)).
  Compute the objective losses and update the network by back-propagation (Eq. (3.11)).
end for
```

3.3 Experiments

3.3.1 Experimental Settings

Datasets and protocols. The proposed LOGA is evaluated on four video-based ReID datasets: MARS [178], Duke-Video [120, 152], Duke-SI [80], iLIDS-VID [143]. Example tracklets are shown in Figure 3.4. The MARS has 20,478 tracklets of 1,261 persons captured from a camera network with 6 near-synchronized cameras. Duke-Video is a newly released large-scale benchmark of 1,812 person identities with 4,832 tracklets. Duke-SI is a fully auto-generated version of Duke-Video without manual frame selection, thus, more practical and challenging. The iLIDS-VID dataset is a relatively small-scale including 600 video tracklets of 300 persons captured



(a) Duke-SI

(b) MARS

(c) Duke-Video

(d) iLIDS-VID

Figure 3.4: Example pairwise tracklets for the same identity. Various noises are caused by illumination, viewpoints, resolution, occlusion, background clutter, etc.

by two disjoint cameras in an airport arrival hall. To evaluate the effectiveness of the proposed LOGA model, we adopted two commonly used performance metrics in person re-id including Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) [179].

Implementation Details. For fair comparisons, we took a ResNet50 [40] as the backbone network for global visual feature extraction [37]. Given that the video tracklets are composed of an arbitrary number of frames, we split each tracklet into several clips with a fixed length of 10. We randomly sampled 4 identity instances each with 8 clips to construct a mini-batch in model training. All the frames were resized to 256×128 and augmented by random horizontal flip. We used Adam [67] with a weight decay of $5e-4$ for model optimization. The margin Δ in Eq. (3.10) is set to 0.3, and the dimension D of representations is set to 2048 following [37, 101]. The kernel size S for the 1D convolution in Eq. (3.4) is set to 10. The model was trained on two P100 GPUs for 240 epochs, and the learning rate is initialized to $3e-4$ which linearly decayed with a factor of 0.1 per 60 training epochs. During the testing stage, the tracklet-level representation was obtained by averaging and pooling the learned representations of their clips. Cosine distance was then used to measure the distances between a query and every probed tracklet in the gallery for ReID.

3.3.2 Comparative Evaluations

In Table 3.1, we compared the proposed LOGA model with a wide range of SOTA video person ReID methods. The LOGA model yielded the best results across the board, which suggests

Methods	Duke-Video				Duke-SI				MARS				iLIDS-VID		
	mAP	R1	R5	R20	mAP	R1	R5	R20	mAP	R1	R5	R20	R1	R5	R20
TAUDL† [79]	-	-	-	-	20.8	26.1	42.0	57.2	29.1	43.8	59.9	72.8	26.7	51.3	82.0
EUG [152]	78.3	83.6	94.6	97.6	-	-	-	-	67.4	80.8	92.1	96.1	-	-	-
Snippet [11]	-	-	-	-	-	-	-	-	76.1	86.3	94.7	98.2	85.4	96.7	99.5
VRSTC [51]	93.5	95.0	99.1	99.4	-	-	-	-	82.3	88.5	96.5	-	83.4	95.5	99.5
GLTP [75]	93.7	96.3	99.3	99.7	-	-	-	-	78.5	87.0	95.8	98.2	86.0	98.0	-
UTAL† [80]	-	-	-	-	36.6	43.8	62.8	76.5	35.2	49.9	66.4	77.8	35.1	59.0	83.8
STMP [96]	-	-	-	-	-	-	-	-	72.7	84.4	93.2	96.3	84.3	96.8	99.5
STA [25]	94.9	96.2	99.3	99.6	-	-	-	-	80.8	86.3	95.7	98.1	-	-	-
STAR [151]	93.4	94.0	99.0	99.7	-	-	-	-	76.0	85.4	95.4	97.3	85.9	97.1	99.7
FGRA [15]	-	-	-	-	-	-	-	-	81.2	87.3	96.0	98.1	88.0	96.7	99.3
MG-RAFA [173]	-	-	-	-	-	-	-	-	**85.9**	88.8	**97.0**	**98.5**	88.6	98.0	99.7
AP3D [37]	95.6	96.3	-	-	74.7	79.3	91.7	97.4	85.1	**90.1**	-	-	86.7	-	-
LOGA	**96.6**	**97.0**	**99.4**	**99.9**	**76.6**	**81.0**	**92.8**	**97.8**	84.1	89.5	96.3	97.9	**91.3**	**99.3**	**100.0**

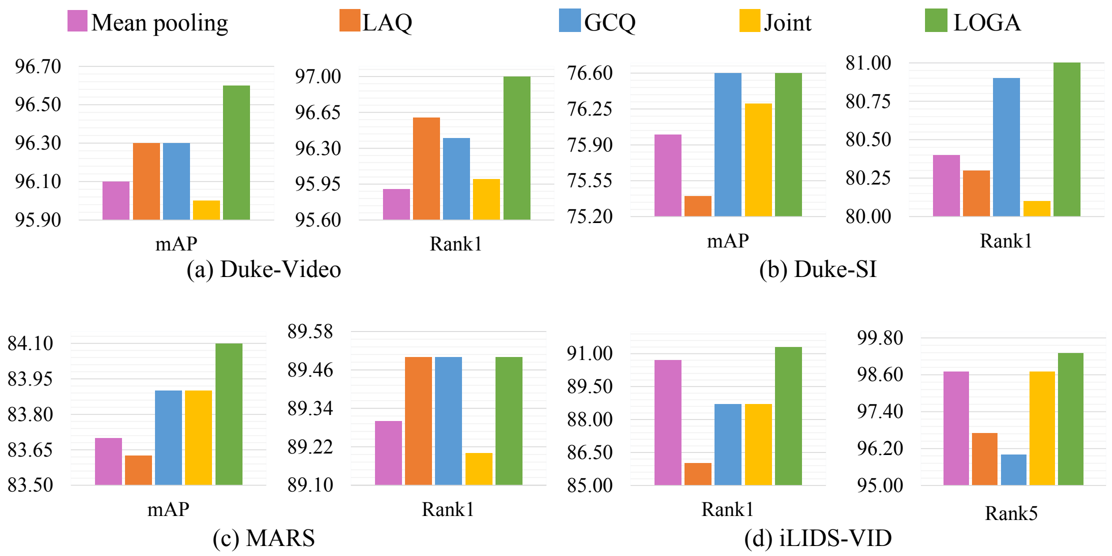
Table 3.1: Performance comparisons of LOGA in video person ReID. Results of the prior methods are from the original papers beside AP3D on Duke-SI which was reproduced by their released code. The 1st/2nd best results are in **bold**/underlined. '†': unsupervised learning based methods.

the efficacy of associativity exploring local part alignments and global appearance correlation in assembling a discriminative representation of a tracklet. Whilst maintaining its competitiveness on the large-scale MARS and the well-curated Duke-Video datasets, the LOGA model achieved compelling improvements over the other methods on iLIDS-VID and its performance advantage is more significant on the automatically detected and segmented Duke-SI, in which case LOGA outperformed the others by 1.9%∼55%, 1.7%∼54.9% and 1.1%∼50% on mAP, rank-1 and rank-5, respectively.

3.3.3 Ablation Study

We conducted further studies to experimentally investigate the effectiveness of exploring the complementary local and global information by solely considering one while ablating another, and also demonstrated the superiority of our associative assembling over the dual-branch strategy [15] which used both local and global information separately. We also provided comprehensive visualization for intuitive understanding.

Components analysis. We started by examining the role of local part alignments by introducing



(a) Duke-Video

(b) Duke-SI

(c) MARS

(d) iLIDS-VID

Figure 3.5: Ablation studies on components and an alternative 'Joint' model, learned with local and global information independently by a dual-branch network [15].

LAQ for frame assembling. Figure 3.5 (pink *v.s* orange) shows that both metrics on most datasets are decreased. This is caused by the unrealistic assumption that local regions of all the frames are well-aligned. Such an assumption is shown to be unreliable due to uncontrollable environment and fragile detection/segmentation. We further examined the importance of global appearance by solely employing GCQ for frame assembling. The unsatisfying performance as reported in Figure 3.5 (pink *v.s* gray) suggests assessing the quality of frames in accordance with solely the unobstructed global appearance is unreliable owing to the fine-grained details being ignored. In contrast, when both LAQ and GCQ are adopted, LOGA exhibits remarkable advantage over all other counterparts (green *v.s* others). This demonstrates the indispensable of both LAQ and GCQ.

Effects of assembling strategy. We further studied the effects of different strategies to join the local and global information in frames assembling: (1) separately assembling by two individual branches learned in parallel according to the two kinds of information [15]. (2) directly connecting local and global information by rescaling the per-frame visual features $\boldsymbol{E}$ according to their normalized local alignment scores (Eq. (3.5)) then explore their global correlations by the conventional self-attention on the rescaled features. (3) associatively assembling by combining the local-assembled prototype and global-assembled residual (Eq. (3.8)) to exploit their synergy. The comparison given in Figure 3.6 (green *v.s* others) shows a noticeable advantage of LOGA

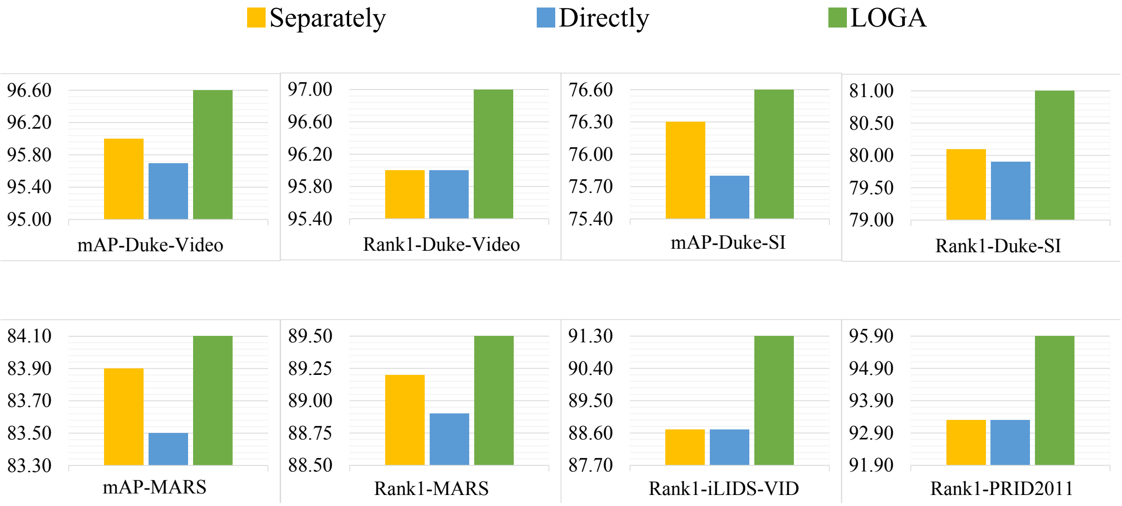


Figure 3.6: Impacts of assembling strategies.

over the dual-branch or direct-connecting counterpart, which demonstrates the effectiveness of the proposed associative assembling strategy.

Effects of local part size. We study the effects of local part size by varying the kernel size of the 1D convolution in Eq. (3.4) and experimented on iLIDS-VID. The experimental results shown in Figure 3.7 indicates our model's robustness to this hyper-parameter within a wide range of values thanks to the subsequent GCQ module which helps refine the local alignment scores according to global correlations. Given that improving S doesn't benefit the performance but increase the model's complexity, we set $S = 10$ in practice.

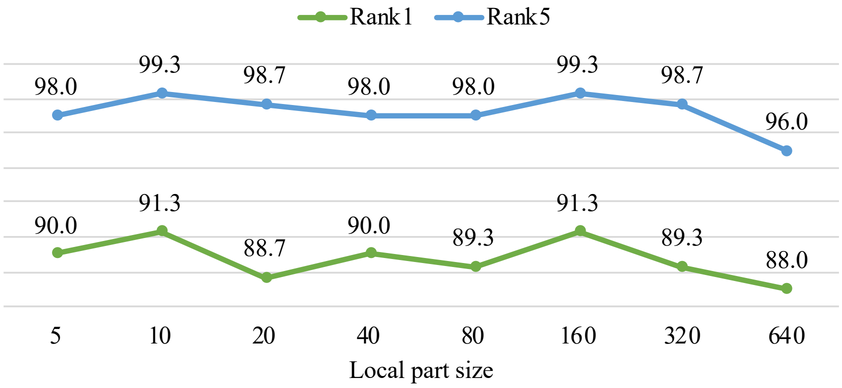
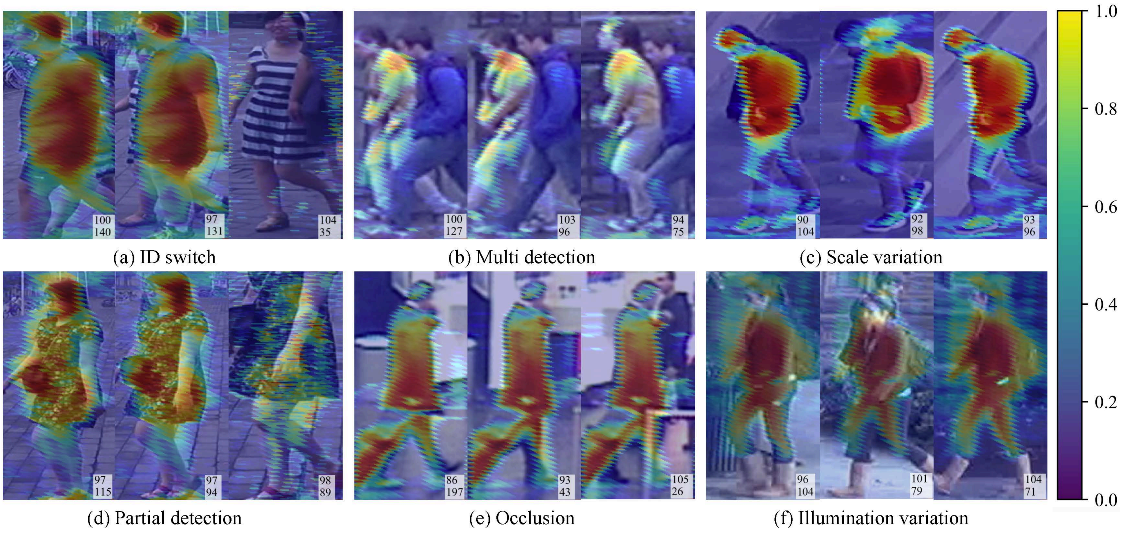


Figure 3.7: Impacts of local part size in LAQ module.

Qualitative studies. Figure 3.8 shows several video clips stacked with their activation maps generated according to the quality of their local parts. Each frame's local-aligned score (upper, Eq. (3.5)) and global-correlated score (lower, Eq. (3.6)) are attached at their bottom-right corner. As exhibited, LOGA is robust to various kinds of noise by providing a faithful importance score for assembling a discriminative representation. The activation maps accurately reveal the critical regions for ReID. The global-correlated scores are obtained with the complementary appearance information so can reliably adjust the biased local-aligned scores. For instance, as shown in Figure 3.8, LAQ enables the network to focus on the target instead of the switched ID or the irreverent multi-detected ID as shown in the activation maps. For the low-quality frames caused by partial-detection, scale-variation and occlusion, etc. LAQ can faithfully assess the local quality. The suitable importance score revealed by the association of LAQ and GAQ efficiently guides LOGA to learn the representation from the most discriminative region in the most discriminative frames.



(a) ID switch

(b) Multi detection

(c) Scale variation

(d) Partial detection

(e) Occlusion

(f) Illumination variation

Figure 3.8: Visualisations of video clips suffering from various noise. The local-alignment scores and global-correlation scores are shown at the upper and lower parts of each frame's right-bottom corner, respectively. Their corresponding importance in assembling are shown at the right-bottom corner of each frame with the local-alignment scores at the top and the global-correlation scores beneath (amplified by 1,000 times).

3.4 Summary

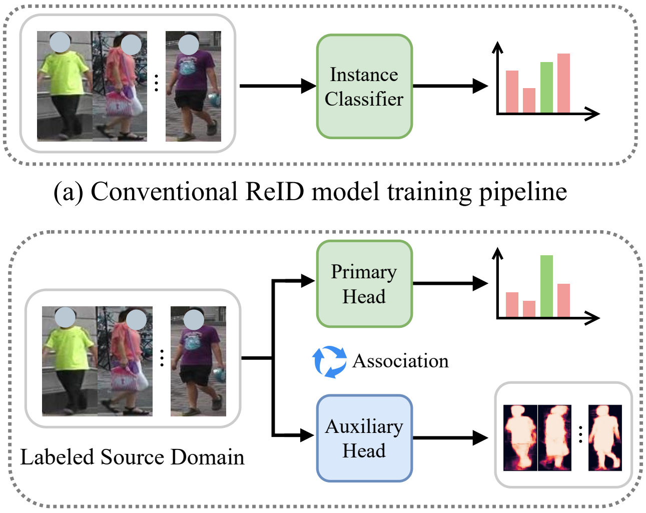
In this chapter, we presented LOGA method for video person ReID through selectively assembling video frames of diverse qualities to derive a more reliable and discriminative representation of a video tracklet. This is accomplished by assessing the frame's quality according to both their local part alignments and global appearance correlation so to refrain from integrating undesired visual information into tracklet's representation causing identity mismatch. Different from existing approaches which explore either local or global information separately, our LOGA method constructs a local-assembled global appearance prototype of a tracklet so to alleviate biased quality assessment caused by either identity-irrelevant misalignment or spatial-insensitive appearance miscorrelation. Extensive experiments on four benchmark datasets show the performance advantages of LOGA over a wide range of the SOTA video ReID methods. Detailed ablation studies are also conducted to provide in-depth discussions about the rationale and essence of different components in our model design.

Chapter 4

Primary-Auxiliary Objectives Association

4.1 Introduction

Current Re-Identification (ReID) techniques are built based on an intrinsic assumption of (Independent and Identical Distribution (IID)) between training and test data. The IID assumption becomes mostly invalid across different domains when training and test data are not from the same environment. As a result, most contemporary ReID models suffer from dramatic degradation when applied to a new domain [100, 17, 147]. Domain Generalization (DG) methods [186, 188, 103], which aim to learn a generalizable model between a source and a target domain have been explored by recent studies to address this problem. Several Domain Generalizable ReID (DG ReID) methods have been developed to mitigate performance degradation caused by domain shift between training (source) data and test (target) data. They can be broadly categorized into three main groups: (1) Learning from diversified training samples [64, 2], (2) Aligning the distribution of source domains by data statistics [191, 190, 61], (3) Exploiting meta-learning [17, 188, 18, 177] to mimic source-target distribution discrepancies. The first category confers advantages to a model through the utilization of a diversified training dataset by either image sample augmentation or feature distribution expansion. The second category aims to learn a source-invariant model by aligning the training data, and expecting it to be invariant for the target domain. The third category focuses on simulating the training/testing discrepancy. Despite some performance improvement from these methods, their overall performances across domains remain poor, *e.g* the latest State-of-the-Art (SOTA) models [17, 177] can only achieve below 20%



(a) Conventional ReID model training pipeline

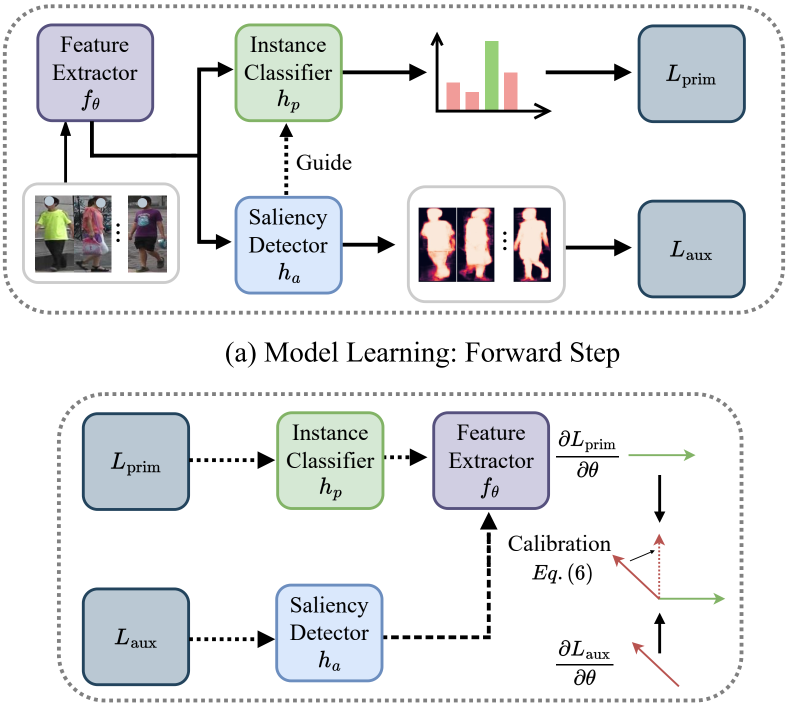
(b) The proposed Primary-Auxiliary Objective Association

Figure 4.1: Comparison on conventional DG ReID model and PAOA model. A typical ReID model is typically trained by optimizing an instance classification objective, which can suffer from overfitting to domain-specific characteristics, *e.g* luminance, background, scale, and viewpoint. The PAOA model considers learning jointly a weakly labeled/supervised auxiliary saliency detection task concurrently with the primary task of the discriminative person ReID. This is achieved by calibrating the gradient of the auxiliary task against that of the primary objective as its reference.

mean Average Precision (mAP) on the MSMT17 benchmark. This highlights the limitation of overfitting in the current Domain Generalizable ReID models and their inability to learn a more generalizable cross-domain model representation. We consider this is due to the not-insignificant interference of domain-specific contextual scene characteristics such as background, viewpoint, and object distances to a camera (scale), which are identity-irrelevant but can change significantly across different domains. Contemporary Domain Generalizable ReID models are mostly trained by an instance-wise classification objective function, indirectly learning person foreground attention selection (Figure 4.1(a)). They are sensitive to such domain-specific but identity-irrelevant contextual information, resulting in the misrepresentation of person foreground attention and leading to less discriminative ReID representation. This likely causes notable ReID performance degradation on models trained and deployed in different domains. To mitigate the impact of domain-specific contextual attributes, an intuitive solution is to isolate the pedestrian object to acquire a domain-invariant representation. Several endeavors [41, 189, 127] have been made to guide the person identification network focusing on the pedestrian with the human saliency prior, which can point out the attentive region relevant to the human subject. These methods have certain limitations, either relying on exhaustive manual masking [127] or lacking an appropriate training objective [41, 189] to ensure the accuracy of the generated segmentation mask. Besides this, it is crucial to note that these methods fail to consider the potential worst-case scenario in which the saliency attention prior may be inaccurate, further leading to negative impacts on identification rather than improvement.

In this chapter, we address this problem by introducing a novel model learning regularization method called Primary-Auxiliary Objectives Association (PAOA). Our aim is to minimize domain-specific contextual interference in model learning by focusing more on the domain-invariant person's unique characteristics. This is achieved by introducing the association of learning the primary instance classification objective function with an auxiliary weakly labeled/supervised pedestrian saliency detection objective function, the idea is illustrated in Figure 4.1(b). Specifically, PAOA is realized in two parts: (1) Additionally train a pedestrian saliency detection head with an auxiliary supervision to assist in focusing the primary ReID discriminative learning task on more domain-invariant feature characteristics. (2) Eliminate the interference attributed to inaccurate saliency labels by calibrating the gradients of the shared feature extractor raised from the weakly-labeled auxiliary learning task towards that of the primary task as a reference when they are in conflict [122]. This association mechanism helps ensure the ReID model learns to attentively focus on generic yet discriminative pedestrian information whilst both learning tasks are harmoniously trained.

Our contributions are: (1) We introduce the idea of optimizing a more domain-generic ReID learning task that emphasizes domain-invariant pedestrian characteristics by associating the ReID instance discriminative learning objective to an auxiliary pedestrian saliency detection objective in a way that does not create conflicts or hinder the effectiveness of the primary objective. (2) We formulate a novel regularization called PAOA to implement the proposed association learning. It jointly trains the primary and auxiliary tasks with referenced gradient calibration to solve the conflicting optimization criteria between the two learning objectives, and promote the learning of a more domain-generic ReID model. (3) We further explore the target domain test data characteristics by incorporating the PAOA regularization into a deployment-time model online optimization process. To that end, we formulate a PAOA+ mechanism for on-the-fly target-aware model optimization and show its performance benefit.


Feature Extractor f_θ
Instance Classifier h_p
L_{prim}
Guide
Saliency Detector h_a
L_{aux}
L_{prim}
Instance Classifier h_p
Feature Extractor f_θ
$\frac{\partial L_{\text{prim}}}{\partial \theta}$
Calibration
$Eq.$ (6)
L_{aux}
Saliency Detector h_a
$\frac{\partial L_{\text{aux}}}{\partial \theta}$

(a) Model Learning: Forward Step

(b) Model Learning: Backward Step

Figure 4.2: An overview of the proposed PAOA model. It aims to derive generic feature representations by guiding the network to attentively focus on pedestrian information and mitigate the interference of domain-specific knowledge, which is achieved by the PAOA regularization of a primary classification objective and an auxiliary pedestrian saliency detection objective: (a) The auxiliary task is jointly trained to provide hard-coded spatial attention to the pedestrian region. (b) The primary task is used as a reference to calibrate the gradients of the auxiliary objective when they are conflicting.

4.2 Methodology

In this chapter, we consider the problem of generalizing a ReID model to any new deployment target environment subject to unknown domain bias between the training and the test domains, where there is no labeled training data from the test domain. To that end, we propose a PAOA regularization method to enable the model to be more attentive to learning universal identity generative information that is applicable in any domain whilst concurrently maximizing ReID discriminative information from the domain labeled data. Figure 4.2 shows an overview of PAOA in model training with two associative steps: (1) Guiding the ReID model to focus on discriminative pedestrian information with an additional auxiliary task dedicated to visual saliency detection. (2) Calibrate the gradients of the auxiliary task when it conflicts with the primary instance classification objective. To further boost the performance, we build PAOA+ to utilize the available samples in deployment time by minimizing the proposed auxiliary objective, and demonstrate the plug-and-play merit of our design.

4.2.1 Joint Primary-Auxiliary Objectives Learning

The primary and auxiliary objectives are jointly trained in a multitask learning architecture, which is composed of a shared feature extractor f_θ, and two dedicated heads h_p and h_a respectively for the primary and auxiliary tasks.

Primary Objective: Person ReID Learning a strong instance classification network is fundamentally important for training a discriminative ReID model. Given a labeled training set $\mathcal{D} = \{(x_i, y_i^{(p)})\}_{i\in\{1,\cdots,N\}}$, where x_i is a person image and $y_i^{(p)}$ is the corresponding instance category label, the primary instance classification task is trained with a softmax Cross-Entropy (CE) loss $\mathcal{L}_{\text{id}}$ and a triplet loss $\mathcal{L}_{\text{tri}}$:

$$\mathcal{L}_{\text{id}} = -\sum_{i=1}^{N}\sum_{j=1}^{C} p_i^j \log \hat{p}_i^j, \tag{4.1}$$

where p_i is one-hot vector activated at $y_i^{(p)}$, and $\hat{p}_i^j$ is the probability for categorized into the jth class that calculated from the classifier. The additional triplet loss constrains the distance between positive (same identity) and negative (different identities) sample pairs, which is formulated as

$$\mathcal{L}_{\text{tri}} = \sum_{i=1}^{N}[d_p - d_n + \alpha]_+, \tag{4.2}$$

where d_p and d_n respectively denote the Euclidean distances for the positive and negative pairs in feature space. α is the margin that controls the sensitivity and $[s]_+$ is $\max(s,0)$. The overall loss function for the primary task is as follows:

$$\mathcal{L}_{\text{prim}} = \mathcal{L}_{\text{id}} + \mathcal{L}_{\text{tri}}. \tag{4.3}$$

Auxiliary Objective: Pedestrian Saliency Detection As illustrated in [132], an auxiliary task closely aligned with the primary task can substantially prompt the learning of the primary objective. Inspired by this, we formulated the auxiliary task as pedestrian saliency detection to perform pixel-level pedestrian localization within the cropped pedestrian bounding boxes. Such an auxiliary task is complementary to the primary task by providing pixel-level hard-coded spatial attention to guide the ReID model to focus on the pedestrian region. Instead of exhaustively manually annotating the pedestrian region, we benefit from the large-scale trained model [175] and perform feed-forward inference to get the weakly labeled samples. Specifically, given a trained saliency model $\mathcal{G}$, we feed the sample to obtain the weak label as $y_i^{(a)} = \mathcal{G}(x_i)$, which is a 2D map to indicate the saliency area. The auxiliary task is essentially a regression task in the pixel level. To that end, the auxiliary head h_a is designed as a lightweight module composed of cascaded 2D CNN layers to predict the saliency map. It is optimized by minimizing a conventional $L1$ loss on the predicted salient label $\hat{y}_k^{(a)}$:

$$\mathcal{L}_{\text{aux}} = \sum_{k=1}^{N_k} |y_k^{(a)} - \hat{y}_k^{(a)}|. \tag{4.4}$$

Joint Multi-task Learning: To build a joint multitask learning pipeline, we formulate the overall objective function by combining both L_{prim} and L_{aux} as

$$\mathcal{L}_{\text{train}} = \frac{1}{N}\sum_{1}^{N} \mathcal{L}_{\text{prim}}(x_i, y_i^{(p)}; f_\theta, h_p) + \lambda \mathcal{L}_{\text{aux}}(x_i, y_i^{(a)}; f_\theta, h_a), \tag{4.5}$$

where λ is the balancing hyperparameter.

Limitation: Despite the auxiliary objective essentially providing hard-coded spatial attention to guide the network being focused on the salient pedestrian object, this pipeline is intrinsically limited. This is due to the inherent noise in the weak label of the auxiliary task that brings a detrimental impact on the primary task and distracts the shared feature extractor from focusing on the pedestrian region. This has further resulted in a divergent gradient descent direction, reflected by the conflicting gradients. Hence, it becomes necessary to perform a post-operation that resolves the conflicts between the learning objectives.

4.2.2 Association: Referenced Gradient Calibration

During the model training, the learnable parameter θ of the shared feature extractor f_θ is updated based on two loss gradients: $\boldsymbol{g_p} = \frac{\partial L_{\text{prim}}}{\partial \theta}$ from the primary objective and $\boldsymbol{g_a} = \frac{\partial L_{\text{aux}}}{\partial \theta}$ from the auxiliary objective. However, when $\boldsymbol{g_p}$ and $\boldsymbol{g_a}$ are in conflict as reflected by a negative inner product, *i.e* $(\boldsymbol{g_a} \cdot \boldsymbol{g_p}) < 0$, their joint effort cannot provide the network with an informative direction on which to perform the gradient descent to optimize the parameters. Therefore, collectively they bring significant difficulty in model convergence and can even lead to destructive interference [163]. To address this fundamental limitation, we propose to break through the dilemma by calibrating the conflicting gradient yield by the auxiliary objective with that from the primary objective as a reference. Specifically, when $\boldsymbol{g_a}$ is conflicting with $\boldsymbol{g_p}$, we consider $\boldsymbol{g_p}$ as a reference and manually alter the direction of $\boldsymbol{g_a}$ by mapping it to the normal plane of $\boldsymbol{g_p}$ to get the calibrated gradient $\boldsymbol{g_a^c}$ as

$$\boldsymbol{g_a^c} = \boldsymbol{g_a} - \frac{\boldsymbol{g_a} \cdot \boldsymbol{g_p}}{\|\boldsymbol{g_p}\|^2} \boldsymbol{g_a}, \qquad \text{subject to } (\boldsymbol{g_a} \cdot \boldsymbol{g_p}) < 0, \tag{4.6}$$

Remark: This procedure changes the direction of the conflicting gradient to ensure it does not conflict with the primary task. With the calibrated gradient, the model can consider the partial guidance of the auxiliary objective, ensuring the joint effort is non-conflicting with the primary objective. It is effective in minimizing the side effects caused by the inaccurate labeling of the auxiliary task while still performing conventional first-order gradient descent to optimize the model.

4.2.3 Deployment-Time Optimization

We further formulate the PAOA+ to exploit the data characteristic of the target domain and perform deployment time optimization with the available samples during testing. Considering that the proposed PAOA is composed of a shared feature encoder f_θ and two separate task heads h_p and h_a that are optimized jointly during model training. When the trained model is deployed in a new environment, given a batch of identity-unknown samples $\{x'_i\}_{i \in \{1, \cdots, B'\}}$, with the corresponding weakly labels $\{y_i^{\prime(a)}\}$ generated by the pre-trained saliency detection model $\mathcal{G}$, the shared feature extract f_θ can be further optimized on the auxiliary task by minimizing the following loss

$$\mathcal{L}_{\text{test}} = \frac{1}{B} \sum_{1}^{B} \mathcal{L}_{\text{aux}}(x'_i, y_i^{\prime(a)}; f_\theta). \tag{4.7}$$

So that f_θ can be swiftly adapted by considering the data distribution of the new environment, further to yield improved performance on the main task. Note the difference from domain adaptation-based method which assume the test sample is available during the training phase for explicit distribution alignment, PAOA+ only requires a batch of samples with arbitrary numbers for on-the-fly updates, allowing it to seamlessly adapt to new data distributions.

4.2.4 Model Training

Training stage: Given the formulation of the primary and auxiliary tasks, the PAOA model is designed in multitask learning architecture and can benefit from the conventional learning supervision by jointly minimizing the primary and auxiliary losses. The parameters are iteratively optimized with the training loss (Eq. (4.5)). As the feature extractor parameterized by θ is shared by both the primary and auxiliary tasks, it will be jointly updated with two gradients: $\boldsymbol{g_p}$ for the primary task and $\boldsymbol{g_a}$ for the auxiliary task. To seek positive interactions between tasks, the direction of $\boldsymbol{g_a}$ will be calibrated only if it conflicts with $\boldsymbol{g_p}$ by Eq. (4.6). Note that the CE loss provides stronger supervision for person classification, therefore we use its gradients as the reference to calibrate that of the auxiliary task. This calibrated gradient ensures the auxiliary task is harmoniously trained with the primary task by back-propagation and thereby brings benefits to facilitate the deployment-time optimization. The overall training procedure is depicted in Algorithm 2.

Algorithm 2 Model Training with PAOA regularization

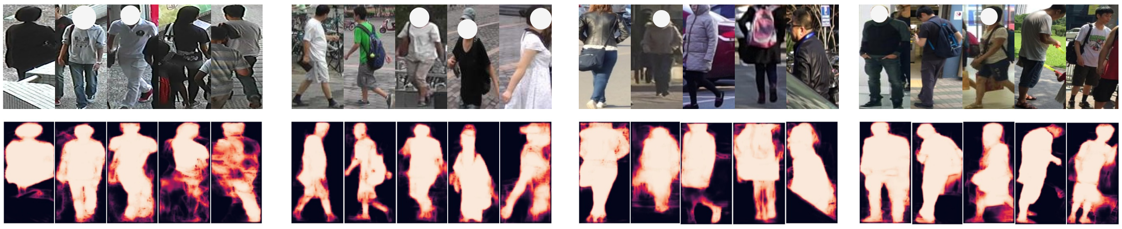
```
**Input:** Labeled dataset $\mathcal{D} = \{(x_i, y_i^{(p)})\}$ for primary task, weak label generator $\mathcal{G}$ for auxiliary task, shared feature extractor $f_\theta$, head modules $h_p/h_a$ for primary/auxiliary tasks.
**Output:** Trained $f_\theta$, $h_p$ and $h_a$.
for $i = 1$ to $max\_iter$ do
    Randomly sample a mini-batch $\{(x_i, y_i^{(p)})\}_{i \in \{1,\cdots,N_\mathrm{B}\}}$ from source dataset $\mathcal{D}$.
    Generate the weak label for the auxiliary task by $\{y_i^{(a)} = \mathcal{G}(x_i)\}_{i \in \{1,\cdots,N_\mathrm{B}\}}$.
    Compute the training loss (Eq. (4.5)) and calculate the gradients.
    Calibrate the conflicting gradients (Eq. (4.6)).
    Update the network by gradient descent.
end for
```

Deployment stage: To make a consistent comparison with Domain Generalizable ReID methods, we can directly apply the trained PAOA model for identity representation extraction. Additionally, the improved PAOA+ model further performs deployment time optimization during the testing stage to mitigate the domain shift between the training and testing domains. Given the identity representations, subsequent identity retrieval is performed by a general distance metric.

4.3 Experiments

4.3.1 Experimental Settings

Implementation Details We used PFAN [175] as the wake label generator for the auxiliary task. The shared feature extractor is a ResNet50 [40] pre-trained on ImageNet [19] to bootstrap the feature discrimination. The balancing hyper-parameter in Eq. (4.5) was set to 0.1. The batch size was set to 64, including 4 images for 16 randomly sampled identities. All images were resized to 128 × 256. The model was trained for 200 epochs with the Adam optimizer [67]. The learning rate was set to $3.5e-4$. The dimension of the extracted identity representation was set to 2048. The dimension of the saliency map is 64×32. The learning rate for PAOA+ was set to $1e-6$ and the test batch size was 200. The post-optimization step is set to 1 for balancing performance and efficiency. All the experiments were implemented on PyTorch [116] on a single A100 GPU.



(a) CUHK-SYSU (b) CUHK03 (c) MSMT17 (d) Market1501

Figure 4.3: Examples from different domains and weak labels for the auxiliary task. Significant domain gaps are caused by the variation in nationality, illumination, viewpoints, resolution, scenario, etc. As complementary, the pedestrian saliency label can provide a guide on the most discriminative person area.

Datasets and Evaluation Protocol We conducted multi-source domain generalized ReID on a wide range of benchmarks. including Market1501 (M) [179], MSMT17 (MS) [147], CUHK03 (C3) [86], CUHK-SYSU (CS) [154], CUHK02 (C2) [84], VIPeR [34], PRID [47], GRID [99], and iLIDs [180]. We evaluated the performance of PAOA on the four small-scale datasets following the traditional setting [129, 63, 4, 167]. We also performed leave-one-out evaluations by using three datasets for training and the remaining for the test [177, 17, 90]. Note that the CUHK-SYSU is only for training given all the images are captured by the same camera. To learn a discriminative model benefits from diverse identities, all the identities regardless of the original train/test splits, were used for training. We adopted mAP and R1 of CMC as the evaluation metrics.

4.3.2 Comparative Evaluations

We compared the proposed PAOA against several recent SOTA methods, and the comparison results are shown in Table 4.1 and Table 4.2. Under a fair comparison with existing Domain Generalizable ReID methods, the PAOA model outperforms all the competing methods by a significant margin on both the traditional setting and the large-scale settings across all the evaluation metrics. It shows a clear advantage over the recent SOTA methods. Notably, even trained with

Table 4.1: Performance comparisons of PAOA on traditional evaluation protocol. The best results are shown in **red** and the second-best results are shown in blue.

Source	Method	PRID		GRID		VIPeR		iLIDs		Average	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
M+D+C2 +C3+CS	DIMN [129]	52.0	39.2	41.1	29.3	60.1	51.2	78.4	70.2	57.9	47.5
	SNR [64]	66.5	52.1	47.7	40.2	61.3	52.9	89.9	84.1	66.3	57.3
	DMG-Net [4]	68.4	60.6	56.6	51.0	60.4	53.9	83.9	79.3	67.3	61.2
M+C2+ C3+CS	M3L [177]	64.3	53.1	55.0	44.4	66.2	57.5	81.5	74.0	66.8	57.2
	MetaBIN [17]	70.8	61.2	57.9	50.2	64.3	55.9	82.7	74.7	68.9	60.5
	ACL [167]	73.5	63.0	65.7	55.2	75.1	66.4	86.5	81.8	75.2	66.6
	META [156]	71.7	61.9	60.1	52.4	68.4	61.5	83.5	79.2	70.9	63.8
	PAOA (Ours)	74.0	65.6	67.2	56.3	76.6	66.7	87.1	83.1	76.2	67.9
	PAOA+ (Ours)	**75.1**	**66.5**	**67.8**	**56.9**	**77.2**	**67.7**	**88.0**	**83.9**	**77.0**	**68.8**

fewer datasets compared with [129, 64, 4], the proposed method is still able to extract discriminative features for identity matching. Besides, we extended our analysis to include the results from the test-time optimization variant, PAOA+, which notably improves PAOA consistently across all benchmarks. These results provide additional evidence on the effectiveness of the associative learning strategy, where the auxiliary task can promote the primary ReID objective during test time given the absence of identity labels.

4.3.3 Ablation Study

Component Analysis We investigated the effects of different components in PAOA model design to study their individual contributions. The baseline model is a ResNet50 pre-trained on ImageNet. The comparison results are shown in Figure 4.4, from which we can observe that the auxiliary objective (A) and the gradient calibration (G) strategies can consistently improve performance. With further deployment-time optimization (D), our model can be advanced by benefiting from mining the data characteristics in the target domain. Notably, the variant without gradient calibration can always benefit more from that post-optimization compared with the PAOA+ (B+A+G+D) model, This further illustrates that the referenced calibration mechanism has already enabled the PAOA model to be more attentive to the domain-invariant pedestrian region, and therefore it relies less on on-the-fly optimization.

Influe on Number of Update Interactions We analyzed the effects of update iterations on Deployment-Time Optimization and reported the results in Table 4.3. From which we observed

Table 4.2: Performance comparisons of PAOA on large-scale evaluation protocol. The best results are shown in **red** and the second-best results are shown in blue.

Method	Reference	M+MS+CS→C3		M+CS+C3→MS		MS+CS+C3→M		Average	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1
SNR [64]	CVPR2020	17.5	17.1	7.7	22.0	52.4	77.8	25.9	39.0
QAConv$_{50}$ [90]	ECCV2020	32.9	33.3	17.6	46.6	66.5	85.0	39.0	55.0
M^{3}L [177]	CVPR2021	35.7	36.5	17.4	38.6	62.4	82.7	38.5	52.6
MetaBIN [17]	CVPR2021	43.0	43.1	18.8	41.2	67.2	84.5	43.0	56.3
ACL [167]	ECCV2022	49.4	50.1	21.7	47.3	76.8	90.6	49.3	62.7
META [156]	ECCV2022	47.1	46.2	24.4	52.1	76.5	90.5	49.3	62.9
PAOA	Ours	49.8	50.5	25.1	51.5	77.1	90.8	50.7	64.3
PAOA+	Ours	**50.3**	**50.9**	**26.0**	**52.8**	**77.9**	**91.4**	**51.4**	**65.0**

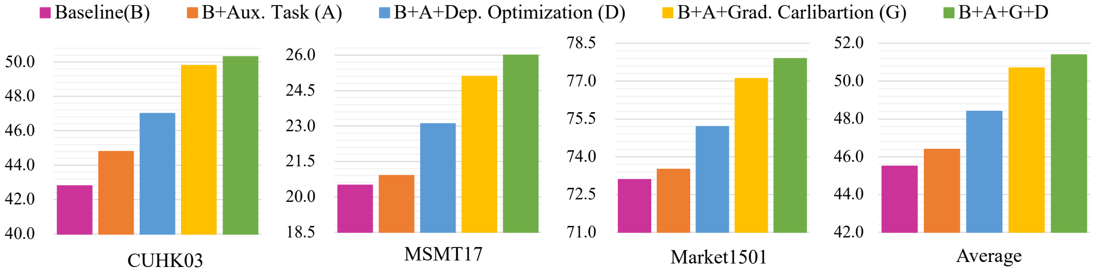


Figure 4.4: Component analysis in PAOA model.

that, a single update iteration consistently enhances model performance across all datasets, with the mAP consistently increasing. This improvement underscores the benefits of feature extractor optimization during deployment, particularly noticeable in datasets C3 and MS, where mAP peaks at the second iteration. However, further iterations tend to slightly degrade performance, which suggests an optimal balance between model update frequency and maintaining core task efficacy. This trend is evident as the mAP averages peak at the second iteration before declining, which indicats that excessive updates may introduce a bias towards the auxiliary task, thereby diminishing the model's discriminative power for its primary objective. Consequently, we strategicly limit of one update iteration by default, so as to ensure the model remains adaptable and generalizable across varying inputs.

Gradient Calibration Designs We adopted a primary-referenced design for the gradient calibration between the primary and auxiliary objectives. This was based on the fact that the primary

Table 4.3: Effects of update iterations during deployment optimization on mAP (%).

Dataset	0	1	2	3	4
C3	49.8	50.3	50.5	50.6	50.3
MS	25.1	26.0	26.5	26.0	25.0
M	77.1	77.9	77.5	77.0	76.2
Avg.	50.7	51.4	51.5	51.2	50.5

instance classification objective provides stronger supervision to identify pedestrians, while the auxiliary objective is to guide the instance classifier to attentively focus on the pedestrian area and ignore the domain-specific interference. It's weakly labeled and therefore is intricately noisy which can lead to a negative influence on the primary objective, reflected by the conflicting gradient. We examined the effect of the calibration design by additionally testing three more formulations as demonstrated in Figure 4.5. Table 4.4 shows the auxiliary-referenced design yielded the worst performance, given the gradients of the auxiliary objective is noisy and unreliable, using it as a reference is harmful to the learning of the primary objective. By contrast, the mutually referenced calibration design includes the primary gradients as referenced on top of the auxiliary-referenced design, which alleviates the fallout caused by the gradient destruction, despite it's still inferior to the baseline. In comparison, the primary-referenced design consistently obtained improved performance which supports the design of the proposed primary referenced gradient calibration.

Table 4.4: Comparison of different gradient calibration designs by mAP (%). Refer to Figure 4.5 for the corresponding design.

Design	C3	MS	M	Avg.
a	44.8	20.9	73.5	46.4
b	44.1	21.7	74.7	46.8
c	47.3	23.0	75.3	48.5
d	**49.8**	**25.1**	**77.1**	**50.7**

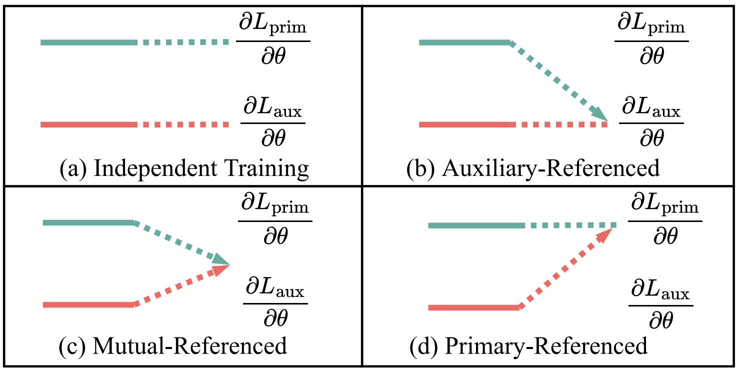

$\frac{\partial L_{\text{prim}}}{\partial \theta}$
$\frac{\partial L_{\text{aux}}}{\partial \theta}$
(a) Independent Training
(b) Auxiliary-Referenced
(c) Mutual-Referenced
(d) Primary-Referenced

Figure 4.5: An illustration of various gradient calibration designs. (a) No gradient calibration as [127]. (b) Gradients of the primary objective are calibrated with the auxiliary objective as a reference. (c) Gradients are calibrated in relation to each other as a reference, as designed in [163]. (d) Gradients of the auxiliary objective are calibrated with the primary objective as a reference.

4.4 Summary

In this chapter, we introduced a novel PAOA regularization to learn a generalizable ReID model for extracting domain-unbiased representations more generalizable to unseen novel domains for person ReID. PAOA encourages the model to get rid of the interference of domain-specific knowledge and to learn from discriminative pedestrian information by the association of learning an auxiliary pedestrian detection objective with a primary instance classification objective. To mitigate the fallout caused by the noisy auxiliary labels, we further derive a referenced-gradient calibration strategy to alter the gradient of the auxiliary object when it's conflicting with the primary object. The PAOA framework is task-agnostic, making it readily adaptable to other tasks through the incorporation of a close auxiliary task and a shared learning module.

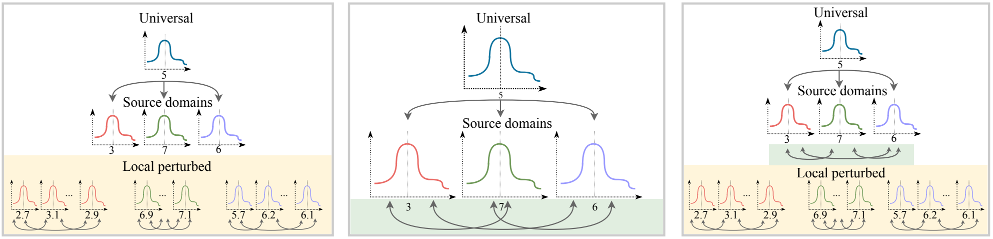
Chapter 5

Feature-Distribution Perturbation and Calibration

5.1 Introduction

Person Re-Identification (ReID) aims to identify the images of the same pedestrians captured by non-overlapping cameras at different times and locations. It has achieved remarkable success when both training and testing are performed in the same domains [87, 181, 174]. However, the widely held IID assumption does not always hold in real-world ReID scenarios due to significantly diverse viewing conditions at different locations of biased distributions at different camera views, and more generally across different application domains. As a result, a well-trained model will degrade significantly when applied to unseen new target domains [100, 17, 147]. To that end, Domain Generalization (DG) [186, 188, 103], which aims at learning a domain-agnostic model, has drawn increasing attention in the ReID community. It is a more practical and challenging problem, which requires no prior knowledge about the target test domain to achieve "out-of-the-box" deployment.

Recent attempts on generalized ReID aim to prevent models from overfitting to the training data in source domains from either a local perspective by manipulating the data distribution of each domain, or in a global view to represent the samples of all domains in a common representational space. The local-based methods [18, 162, 64, 61] are usually implemented by feature perturbation and/or normalization, as shown in Figure 5.1 (a). However, the perturbed distributions constructed from the original data of a single source domain is subject to subtle distribution shift and also domain bias. On the other hand, the global-based approaches [17, 2, 188, 171] aim



(a) Local perturbed training (b) Global aligned training (c) Local-global regularized training

Figure 5.1: An illustration of three training schemes in domain generalized ReID. The 'Universal' is ideally the distribution for any new target domains. The source domains are differentiated by different colors, and the perturbed distributions share the same color with the corresponding original. The number indicates the characteristics of that distribution, and similar value means a smaller domain gap, vice versa. The proposed Feature-Distribution Perturbation and Calibration (PECA) model simultaneously conducts local perturbation and global calibration to eliminate domain bias for learning a domain-agonistic representation.

to align the feature distributions of multiple domains so that the per-domain data characteristic (*i.e* mean and variance of the data distribution which is assumed to be a Gaussian distribution) is ignored when representing images of different domains, as illustrated in Figure 5.1 (b). They often explicitly pre-define a target distribution to be aligned toward, or implicitly learn a global consensus by training a single model with data from all the source domains. However, even the domain gap is reduced by such a global regularization from restricted 'true' distributions, the learned representations are inherently domain-biased toward the consensus of the multiple seen training domains rather than the desired universal distribution scalable to unseen target domains given the number of domains available for training is always limited.

Local domain data manipulation. It is easy to train separate local models with labeled samples respective to each source domain, with subsequently a model aggregation [18, 162]. However, these local models would overfit to the corresponding source domain, while losing generalizability to others. A natural way is either to diverse the local training samples for learning a knowledgeable model, or to eliminate biased information within each source domain for learning a domain-unbiased model. Both solutions fall into the category of local domain manipulation, which alters the data distribution in a per-domain manner. For data diversification, the most intuitive approach is to perform augmentation, on either the raw image [12] or feature spaces [26]. To eliminate local domain bias, the methods-based normalization have been widely studied recently.

Jin *et al* [64] introduced IN for restituting the style component out of an ID representation. Jia *et al* [61] combined BN with instance normalization in a unified architecture to achieve content and style unification. However, these local diagrams consider only per-domain information during feature perturbation, and is still subject to subtle distribution shifts. In this chapter, we propose to associate the local per-domain feature distribution perturbation with global cross-domain feature distribution alignment, to empower the model to be agnostic against holistic domain shift. The complementary regularization provided by the global distribution calibration remedy helps the learned model being invariant against both perturbed distribution shift and real domain gap, so to extract generic yet discriminative representation for any unseen domain.

Global distribution calibration. In contrast to the local approaches, some methods based on global distribution calibration consider the cross-domain association by learning a shared representational space for all domains. These methods are built based on a straightforward assumption that source invariant features are also invariant to any unseen target domains [73]. In this spirit, DEX [2] dynamically performed the space expansion towards the direction of a zero-mean normal distribution with a covariance matrix estimated from the corresponding domain. Recent works [177, 17] took the idea of meta-learning with the aim of "learning to generalize" by randomly splitting available source domains into meta-training and meta-testing sets, to mimic real-world deployment scenarios. Such a scheme implicitly aligns the cross-domain feature distributions to a shared space by randomly setting the alignment target, *i.e* the meta-testing set. Zhang [171] *et al* propose learning causal invariant feature by disentangling ID-specific and domain-specific factors for all the training samples from all the source domains, which enables the disentangled feature to well-preserved ID information while sharing the same feature space for all the domains. However, even aligning among multiple 'real' source domains can reduce the domain gap, the learned representations are still biased towards the consensus of the limited seen training domains, instead of the desired universal distribution scalable to unseen target domains.

In this chapter, we present a PECA model to accomplish generalized ReID with the objective to learn more generalizable discriminative representations for model deployment to unseen target domains. This is achieved by regularizing model training simultaneously with local distribution perturbation and global distribution calibration, as depicted in Figure 5.1(c). Specifically, on the one hand, as each source domain usually depicts limited numbers of pedestrians under certain scenarios, simply training from such data will lead to overfitting to the domain-specific inherently domain biased distribution, which harms the model's generalizability. To address this issue, we introduce the local perturbation module to diversify the feature distribution based on a perturbing factor estimated per domain, which enables the model to be more invariant to distribution shifts. On the other hand, despite the unpredictable distribution gaps between different ReID data domains due to the undesirable scenario-sensitive information embedded in images specific to each domain, *e.g* the background, we consider that the features derived from different independent domains should share a high proportion of information as the universally applicable explanatory factors for domain-independent identity discrimination. In this regard, we propose to simultaneously calibrate the feature distributions across all the source domains, so to eliminate the domain-specific data characteristics in feature representations that are potentially caused by identity-irrelevant redundancy. Both the proposed local perturbation and global calibration modules reinforce the same purpose of regularizing the model training, but they are devised in different hierarchies and complementary to each other. Different from the existing methods which consider only partially from the local or global perspectives, our method handles both to promote the model in learning domain-agnostic representations.

5.2 Methodology

In this chapter, we propose a PECA model to derive domain-agnostic yet discriminative ID representations. It regularizes the model training to satisfy simultaneously both local perturbation and global calibration. The local regularization is built to perform per-domain *feature-distribution* diversification, and the global calibration is designed to achieve cross-domain *feature-distribution* alignment, as shown in Figure 5.2. During training, for each source domain D^k, a batch of samples (x^k, y^k) is fed into the network backbone to extract the feature map e^k. Then we perform per-domain diversification with Local Perturbation Module (LPM) as

$$\{\hat{e}^k\}_{k=1}^K = \{l(e^k)\}_{k=1}^K, \tag{5.1}$$

where $l(\cdot)$ is the function of LPM to enable the local model to be invariant against per-domain shifts by training with the perturbed features $\{\hat{e}^k\}_{k=1}^K$.

The balancing Global Calibration Module (GCM) further regularizes the model learning by aligning the holistic representation (the input feature of the classifier) into a common feature space constructed from $\mathcal{M}$ regardless of the domain label. To distinguish the holistic representation from the intermediate representation e^k, we note it as $v^k \in \mathbb{R}^{B\times d}$ and its perturbed counterpart

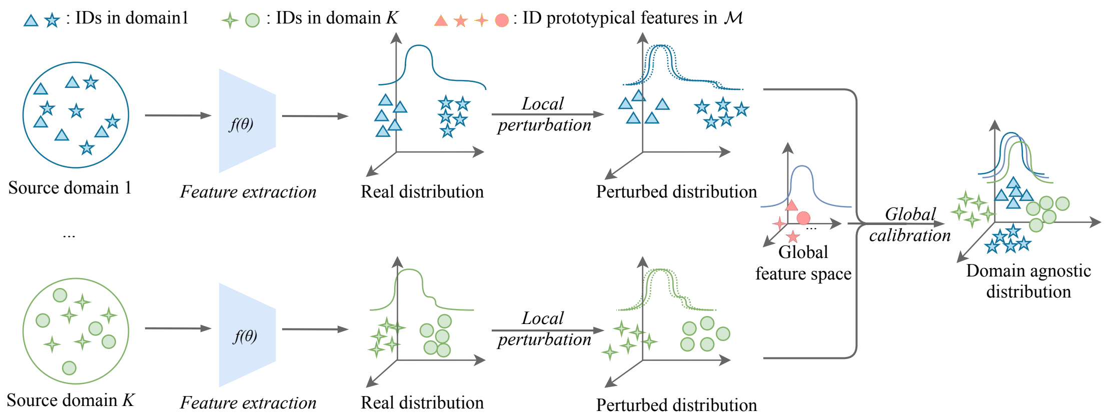


Figure 5.2: An overview of the proposed PECA model. The overall objective is to derive a generic feature representation by avoiding model overfitting to the source domains, which is achieved by Local Perturbation Module to enforce the learned feature invariant to per-domain distribution shifts caused by perturbation, and Global Calibration Module to align cross-domain distribution regardless of domain annotations.

as $\hat{v}^k$ correspondingly, where d is a hyperparameter to the representation dimension. This global regularization is mathematically formulated as

$$\mathcal{L}_{\mathrm{g}}(\hat{v}^k, \mathcal{M}) = ||\mathrm{dist}(\hat{v}^k), \mathrm{dist}(\mathcal{M})||_1, \tag{5.2}$$

where $\mathcal{L}_g(\cdot)$ is the global regularization term aiming to align the distribution of holistic ID representations $\mathrm{dist}(\hat{v}^k)$ with the global distribution $\mathrm{dist}(\mathcal{M})$.

As complementary to the LPM, GCM focus on cross-domain regularization by pulling representations into a domain-agnostic space, thus empowering the generalizability of the ReID model for any unseen novel domain. With the collaboration of LPM and GCM, the PECA model can be trained with arbitrarily conventional ReID objectives in an end-to-end manner. When deployed to an unseen novel domain, a generic distance metric (*e.g* Euclidean or Cosine distance) is used to measure the pairwise representational similarity between the query image against the galleries for identity retrieval.

5.2.1 Local Feature-Distribution Perturbation

Given an intermediate feature representation $e_i^k \in \mathbb{R}^{B \times C \times H \times W}$ extracted from the source domain D^k at i-th layer, the objective of LPM is to perturb per-domain features to avoid local-domain overfitting. For notation clarity, we omit the layer index i in the following formulations. Inspired by feature augmentation [81] and Instance Normalization (IN) [57, 88], LPM performs perturbation by randomly substituting the transformation factors of IN. Specifically, we first calculate the channel-wise moments $\mu(e^k) \in \mathbb{R}^{B\times C}$ and $\sigma(e^k) \in \mathbb{R}^{B\times C}$ for IN as

$$\mu(e^k) = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} e^k_{h,w}, \quad \sigma^2(e^k) = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(e^k_{h,w} - \mu(e^k))^2. \tag{5.3}$$

As suggested by [64], these statistical moments encode not only style information but also certain task-relevant information dedicated to ReID. Instead of discarding all of them for style bias reduction as adopted in [61, 177], we propose to maintain the discrimination while increasing the local-domain data diversity by holistically shifting its distribution. This is achieved by perturbing the per-domain instance moments as

$$\hat{\mu}(e^k) = \mu(e^k) + \varepsilon_\mu \mathrm{h}(\mu(e^k)), \quad \hat{\sigma}(e^k) = \sigma(e^k) + \varepsilon_\sigma \mathrm{h}(\sigma(e^k)), \tag{5.4}$$

where $\mathrm{h}(\cdot)$ calculate the perturbation factors, which are mathematically the standard derivation. They reflect the dispersed level of the local domain, and ensures the perturbation is within a plausible range, so to avoid over-perturbation which causes model collapse, or under-perturbation which cannot provide any benefit in model learning. ε_μ and ε_σ varies the perturbation intensity to guarantee the diversity of perturbed features, and both are randomly sampled from a standard normal distribution. We subsequently perform feature transformation by substituting the local-domain moments as

$$\hat{e}^k = \hat{\sigma}(e^k)\frac{e^k - \mu(e^k)}{\sigma(e^k)} + \hat{\mu}(e^k). \tag{5.5}$$

By introducing the perturbed representation $\hat{e}^k$, the per-domain feature becomes more diverse so to improve the model's generalizability against the per-domain shift.

5.2.2 Global Feature-Distribution Calibration

The GCM is complementary to LPM by aligning the distribution of cross-domain features into a common feature space. GCM considers the association between the perturbed holistic representation $\hat{v}^k$ and a global memory bank $\mathcal{M}$. Specifically, we calculate the global statistical moments $\mu_\mathrm{g} \in \mathbb{R}^d$ and $\sigma_\mathrm{g} \in \mathbb{R}^d$ in each training iteration as

$$\mu_\mathrm{g} = \frac{1}{K}\frac{1}{N^k}\sum_{k=1}^{K}\sum_{n=1}^{N^k}\mathcal{M}^k_n, \quad \sigma_\mathrm{g} = \frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N^k}(\mathcal{M}^k_n - \mu_g), \tag{5.6}$$

where $\mathcal{M}_n^k \in \mathbb{R}^d$ is the prototypical feature of the n-th identity in the k-th domain. These global statistical moments depict a feature space shared by the prototypical representations on $\mathcal{M}$ for all the identities. Subsequently, the holistic representations are calibrated into the joint feature space by

$$\mathcal{L}_{\mathrm{g}}(\hat{v}^k, \mathcal{M}) = \frac{1}{K} \sum_{k=1}^{K} (||\mu(\hat{v}^k) - \mu_{\mathrm{g}}||_1 + ||\sigma(\hat{v}^k) - \sigma_{\mathrm{g}}||_1). \tag{5.7}$$

Here, $\mu(\hat{v}^k) \in \mathbb{R}^d$ and $\sigma(\hat{v}^k) \in \mathbb{R}^d$ are the channel-wise mean and standard derivation of the perturbed representation $\hat{v}^k$. GCM enables the extracted features to fall into a domain-invariant space. The hierarchical regularization achieved by LPM and GCM makes the model generic in extracting domain-agnostic representations.

5.2.3 Model Training

Learning objective. Given the formulations of LPM and GCM, the proposed PECA can benefit from conventional learning supervision. Specifically, the PECA model is jointly trained with a softmax CE loss $\mathcal{L}_{\mathrm{id}}$ and the global regularization item $\mathcal{L}_{\mathrm{g}}$ as

$$\mathcal{L} = \mathcal{L}_{\mathrm{id}} + \lambda \mathcal{L}_{\mathrm{g}}, \quad \mathcal{L}_{\mathrm{id}}(x^k, y^k) = -\sum_{j=1}^{C} p_j^k \log \tilde{p}_j^k, \quad \tilde{p}^k = \mathrm{Softmax}(\mathrm{MC}(\hat{v}^k)). \tag{5.8}$$

The notations x^k and y^k are the raw input images sampled from domain D^k and its corresponding ID label, respectively, whilst p^k is a one-hot distribution activated at y^k. The function $\mathrm{MC}(\cdot)$ stands for the memory-based classifier [185, 177], and λ decides the importance of $\mathcal{L}_{\mathrm{g}}$ regarding the identity loss $\mathcal{L}_{\mathrm{id}}$.

Memory bank update. In each training iteration, once the network parameters are updated according to $\mathcal{L}$ (Eq. (5.8)), the memory bank $\mathcal{M}$ is then refreshed by Exponential Moving Average (EMA) as

$$M_{y^k}^k = \beta M_{y^k}^k + (1-\beta)\hat{v}^k, \quad k = \{1, \ldots, K\}, \tag{5.9}$$

in which β is the EMA momentum. The prototypical features in the memory bank $\mathcal{M}$ is iteratively updated with the latest corresponding ID representations. Consequently, a more discriminative feature space will be yielded by $\mathcal{M}$ for global alignment.

5.3 Experiments

5.3.1 Experimental Settings

Datasets and protocols. We conducted multisource domain generalization on a wide range of benchmarks, including Market1501 (M) [179], DukeMTMC (D) [183], MSMT17 (MT) [147], CUHK02 (C2) [84], CUHK03 (C3) [86], CUHK-SYSU (CS) [154], and four small datasets including PRID [47], GRID [99], VIPer [34], and iLIDs [180]. mean Average Precision (mAP) and CMC accuracy on R1 are adopted as evaluation metrics.

Implementation Details. We used ResNet50 [40] pre-trained on ImageNet to bootstrap our feature extractor. The batch size was set to 128, including 16 identities and 8 images for each. All images were resized to 256×128. We randomly augmented the training data by cropping, flipping, and colorjitter. The proposed PECA was trained 60 epochs by Adam optimizor [67], and we adopted the warm-up strategy in the first 10 epochs to stabilize model training. The learning rate was initialized as $3.5e-4$ and multiplied by 0.1 at the 30th and 50th epochs. The momentum for the memory update was set to 0.8. The dimension of extracted representations was conventionally set to 2048. All the experiments were conducted on the PyTorch [116] framework with four A100 GPUs.

5.3.2 Comparative Evaluations

Comparison under the traditional benchmark setting. Under the existing benchmark setting [18, 64, 129], five datasets (M+D+C2+C3+CS) were used as source domains, and the generalizability was evaluated on four *small-scale* datasets of different domains not contributing to training (unseen), which are PRID, GRID, VIPeR, and iLIDs. All the images in the source domains were used for training, without the original training or testing splits. Being consistent with existing performance evaluation protocols [64, 129], we performed 10-trail evaluations by randomly splitting query/gallery sets, and reported the averaged performance in Table 5.1, which shows the considerable superiority of the proposed PECA over the State-of-the-Art (SOTA) competitors.

Comparison under large-scale benchmark setting. We further evaluated our model on four *large-scale* datasets (M+D+C3+MS) with the 'leave-one-out' strategy, namely taking three datasets used as source domains for model training, and one left out as an unseen target domain. Under this setting, The original train splits in the three source domains were used for training, while

Table 5.1: Performance comparisons of PECA on traditional evaluation protocol. The best results are in **bold**.

Method	PRID		GRID		VIPeR		iLIDs		Average	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
AggAlign [170]	25.5	17.2	24.7	15.9	52.9	42.8	74.7	63.8	44.5	34.9
Reptile [113]	26.9	17.9	23.0	16.2	31.3	22.1	67.1	56.0	37.1	28.0
CrossGrad [123]	28.2	18.8	16.0	9.0	30.4	20.9	61.3	49.7	34.0	24.6
Agg_PCB [131]	32.0	21.5	44.7	36.0	45.4	38.1	73.9	66.7	49.0	40.6
MLDG [71]	35.4	24.0	23.6	15.8	33.5	23.5	65.2	53.8	39.4	29.3
PPA [118]	45.3	31.9	38.0	26.9	54.5	45.1	72.7	64.5	52.6	42.1
DIMN [129]	52.0	39.2	41.1	29.3	60.1	51.2	78.4	70.2	57.9	47.5
SNR [64]	66.5	52.1	47.7	40.2	61.3	52.9	89.9	84.1	66.3	57.3
RaMoE [18]	67.3	57.7	54.2	46.8	64.6	56.6	**90.2**	**85.0**	69.1	61.5
PECA (Ours)	**72.2**	**62.7**	**59.4**	**48.4**	**70.1**	**61.2**	85.7	79.8	**71.9**	**63.0**

the test split on the unseen target domain was used for testing, same as in [177]. The evaluation results in Table 5.2 show that PECA outperforms the SOTA competitors by a compelling margin, Specially, on the more challenging datasets CUHK03 and MSMT17 with larger domain gaps than the other datasets, all methods give relatively poorer generalization performances. In comparison, our PECA model gains the greater advantage over the other methods, especially on R1 scores. This suggests PECA's better scalability with greater potential in real-world deployment to different unseen target domains.

Table 5.2: Performance comparisons of PECA on large-scale evaluation protocol.

Method	Market-1501		DukeMTMC		CUHK03		MSMT17		Average	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
QAConv$_{50}$ [90]	39.5	68.6	43.4	64.9	19.2	22.9	10.0	29.9	28.0	46.6
M3L [177]	51.1	76.5	48.2	67.1	30.9	31.9	13.1	32.0	35.8	51.9
M3L(IBN) [177]	52.5	78.3	48.8	67.2	31.4	31.6	15.4	37.1	37.0	53.5
PECA (Ours)	**58.3**	**81.4**	**49.8**	**70.0**	**34.1**	**35.5**	**17.7**	**43.1**	**40.0**	**57.5**

5.3.3 Ablation Study

Components analysis. We investigated the effects of different components in PECA model design to study individual contributions. We trained a baseline model with only identity loss $\mathcal{L}_{\text{id}}$, and then incorporated it with either LPM or GCM as well as both PECA. Table 5.3 shows that both the LPM and GCM are beneficial individually, and the benefits become clearer when they are jointly adopted as in the PECA model. From another perspective, it also verifies that solely considering the local or global regularization is biased, and it is non-trivial that the PECA explores both in a unified framework to learn a more generic representation.

Table 5.3: Components analysis of LPM and GCM. PECA incorporates both in a unified framework.

Setting	Market-1501		DukeMTMC		CUHK03		MSMT17		Average	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
baseline	54.1	78.5	49.0	68.1	31.1	31.9	14.9	38.1	37.3	54.1
+LPM	57.9	80.4	49.4	69.4	32.7	33.2	**17.7**	42.8	39.4	56.5
+GCM	55.0	79.5	49.0	68.5	32.6	33.6	16.1	39.4	38.2	55.2
PECA	**58.3**	**81.4**	**49.8**	**70.0**	**34.1**	**35.5**	**17.7**	**43.1**	**40.0**	**57.5**

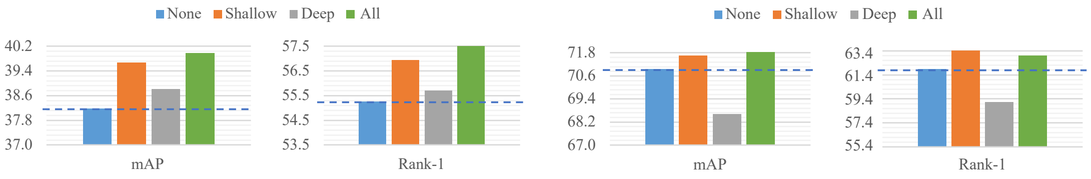
Setting	PRID		GRID		VIPeR		iLIDs		Average	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
baseline	69.1	59.0	59.0	48.4	68.9	60.1	82.5	74.5	69.9	60.5
+LPM	71.5	61.2	58.0	**48.5**	69.7	60.9	85.3	78.7	71.1	62.3
+GCM	69.7	59.5	59.1	**48.5**	69.7	60.7	85.3	78.7	71.0	61.8
PECA	**72.2**	**62.7**	**59.4**	48.4	**70.1**	**61.2**	**85.7**	**79.8**	**71.9**	**63.0**

Discrimination and generalization trade-off. There is a trade-off between being discriminative to the source domains, and being generalized to the target domains [166]. We quantitatively assessed the proposed PECA model in this regard. The results in Table 5.4 indicate the baseline method fails to generalize well to the target domains but yielded compelling discrimination capacity in the source domains, which is likely due to overfitting. As a comparison, our PECA gains notable improvements in generalization ability with only slight performance drops in the source domains. This implies that PECA can effectively balance the generalization and discrimination of feature representations, so to be applied to any novel unseen domains.

Effects of distribution perturbation on different layers. We studied the effects of perturbing the input distributions of various layers in our backbone network, including the 'Shallow' layers

Table 5.4: Local discrimination and global generalization trade-off.

Setting	Source Average		Target: M		Source Average		Target: D	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1
Baseline	**58.6**	**73.8**	54.1	78.5	**63.0**	**77.4**	49.0	68.1
LPGC	57.9	73.0	**58.3**	**81.4**	61.9	76.3	**49.8**	**70.0**
Setting	Source Average		Target: C		Source Average		Target: MS	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1
Baseline	**64.2**	**82.3**	31.1	31.9	**69.9**	**79.3**	14.9	38.1
LPGC	63.9	82.2	**34.1**	**35.5**	68.8	78.7	**17.7**	**43.1**



(a) Averaged performance under large-scale setting. (b) Averaged performance under traditional setting.

Figure 5.3: Effects of distribution perturbation on different layers.

(the first convolution layer and a following residual block), and 'Deep' layers (the last two residual blocks). The results are shown in Figure 5.3. It is not a surprise that perturbing the shallow layers consistently improves the performance under both the traditional and large-scale settings, as perturbations in earlier stages helps enhance the invariance of most layers to distribution shift. However, solely perturbing the deep layers exhibit distinct behaviors under different benchmark settings. This is because the training data under the traditional setting is relatively smaller with restricted diversity and perturbations in later stages tends to affect a limited part of the network that is insufficient to improve the model's generalization ability. Based on these observations, we propose to perturb all the layers to improve the robustness of the PECA model regardless of the dataset scale.

Effects of the global calibration objective. The importance of the global calibration objective for avoiding the model from overfitting to source domains is determined by the hyperparameter λ in Eq. (5.8). By linearly varying λ from 0.1 to 100, we observed from Table 5.5 that moderately applying GCM (*e.g* 0.1 or 1) is beneficial to PECA's generalizability; further increasing λ to a

Table 5.5: Effects of the global calibration objective whose importance is decided by the weight λ in Eq. 5.8. Averaged performances are reported.

Setting	Traditional setting		Large-scale setting	
	mAP	R1	mAP	R1
PECA (**default**, $\lambda = 1$)	**71.9**	**63.0**	**40.0**	**57.5**
PECA w/o GPM	71.0	61.9	37.5	54.3
PECA w/ $\lambda = 0.1$	71.2	62.2	39.4	56.5
PECA w/ $\lambda = 10$	70.8	62.1	39.2	56.5
PECA w/ $\lambda = 100$	33.9	23.9	38.5	55.5

larger value (*e.g* 10 or 100) brings more harm than help. This is because the learning process is dominated by the calibration regularization and the model can barely learn from the identity loss, hence, the resulting feature is less discriminative. We also observed that the traditional setting is relatively more sensitive to λ, as it holds much less training data for learning a robust model, and a similar phenomenon is shown in Figure 5.3. Given the above observations, we set $\lambda = 1$ in practice for our PECA model.

5.4 Summary

In this chapter, we presented a novel PECA model to learn generic yet discriminative representation in multiple source domains generalizable to arbitrary unseen target domains for more accurate unseen domain person ReID. PECA simultaneously conducts model regularization on local per-domain feature-distribution and global cross-domain feature-distribution to learn a better domain-invariant feature space representation. Benefited from the diverse features synthesized by local perturbation, PECA expands per-domain feature distribution to enable more robust to domain shifts. From the global calibration, feature distributions of different domains are represented and holistically referenced in a shared feature space with their domain-specific data characteristics (*i.e* mean and variance of feature distributions) being ignored, resulting in higher model generalizability. Experiments on extensive ReID datasets show the performance advantages of the proposed PECA model over a wide range of SOTA competitors. Extensive ablation studies further provided in-depth analysis of the individual components designed in PECA model.

Chapter 6

Cross-Domain Style Variations Mining

6.1 Introduction

Advances in Convolutional Neural Network (CNN) have notably contributed to enhancing ReID performance, particularly when the training and test data are drawn from the same distribution [87, 181]. However, despite these advancements, a well-trained ReID model can suffer significant degradation when applied to unseen target domains, primarily due to Out-Of-Distribution (OOD) samples resulting from domain shift [22]. Most existing generalizable models are typically designed for a classification task, rather than a ReID task, assuming a universal and homogeneous environment with a joint label space shared between the source (seen) training domain and target (unseen) test domains, as shown in Figure 6.1 (a). In contrast, person ReID is a retrieval task with completely disjoint label spaces in both training and testing. Hence, the direct application of existing domain generalizable models to person ReID is sub-optimal. Recent efforts in the domain of generalizable ReID primarily focus on learning a domain-invariant representation by removing domain-specific information during model training. This is typically achieved by designing a disentanglement module to factorize domain-invariant and domain-specific components from an identity representation [171, 22, 192]. Alternatively, one assumes that the domain gap is mainly caused by style (appearance) variations [136] which can be mitigated with batch or/and instance normalization [64, 112]. Both of these approaches reduce domain-specific characteristics and enable the learned representations to be less domain-biased. However, they are inherently vulnerable in unseen target domain tests for two reasons. Firstly, they inevitably di-

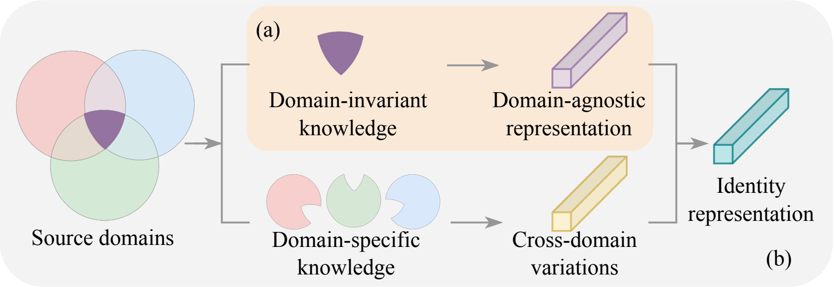

(a)
Source domains
Domain-invariant knowledge
Domain-agnostic representation
Domain-specific knowledge
Cross-domain variations
Identity representation
(b)

Figure 6.1: Comparison of disentanglement learning-based models and CDVM model. (a). Disentangled representation learning-based methods solely utilize domain-invariant knowledge rendering them less robust to cross-view appearance variations unique to different target domains. (b). CDVM explores cross-domain variations to mimic the style discrepancy of an identity captured by different cameras. A model trained with cross-view augmented features further improves model robustness against domain shift in unseen target domains.

minish contextual information and sacrifice the discrimination of the identity representation [17]. Secondly, models trained without accounting for cross-view style variations lack the robustness to extract a generic domain-invariant representation owning to subtle distribution shifts in the test environment [8]. To construct a model with the capacity to learn representations that are simultaneously context-aware discriminative and domain-agnostic generic, a straightforward solution is to collect more cross-camera pairwise samples for each identity, and from more people. However, this is not only too expensive to be realistic but also intrinsically prohibitive due to privacy concerns. Another solution is to increase training data by augmentation, such as random perturbation [81] or adversarial diversification [138]. However, current data augmentation methods lack the assurance of diversified per-identity cross-camera style variations, and may lead to the deterioration of pedestrian-specific information following augmentation.

In this chapter, we introduce a new CDVM model to overcome these limitations. The idea of CDVM is shown in Figure 6.1 (b). The central concept behind the CDVM approach is to enhance the diversity of per-identity instances through the introduction of cross-view style variations across different domains. The objective is to expand the cross-view style inherent to individual identity to learn a generalizable ReID representation that is more robust under the presence of such cross-view style variations. Specifically, we first learn a domain-agnostic (generalizable) identity prototype by exploiting the consensus of identities regardless of their specific domain annotations. Secondly, we enhance the model's robustness by mitigating the covariance

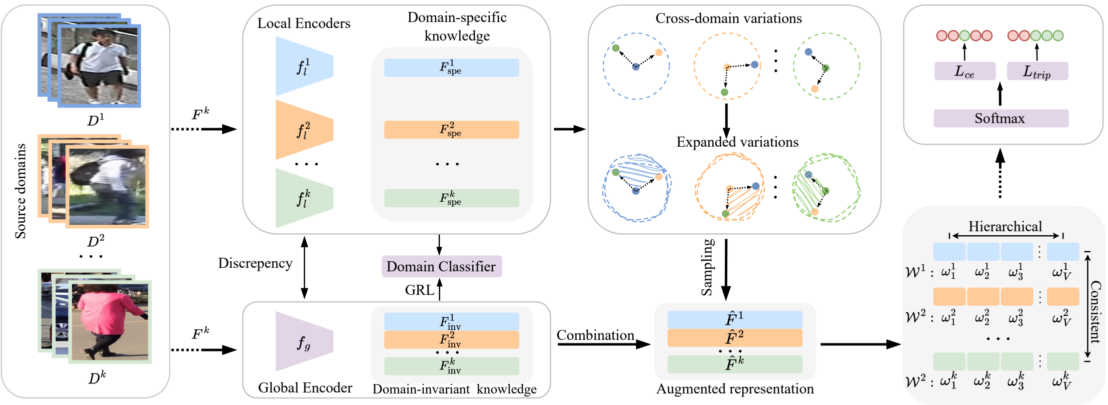


Figure 6.2: An overview of the CDVM model. The overall objective is to enhance the model's robustness against domain shift when applied to an unseen environment. This is achieved by exploring cross-domain variations to mimic the style discrepancy of an identity as captured by different cameras. Throughout model training, an identity representation undergoes progress refinement as follows: **(a)** Learning disentangled representations via the coordinated global and local encoders, supervised by the disentanglement objective (Eq. (6.5)). **(b)** Employing cross-domain variations on the domain-invariant ID representation to facilitate multi-view augmentation (Eq. (6.8)). **(c)** Estimating the significance of the augmented representation and imposing constraints on the intra-domain discrimination and cross-domain consistency (Eq. (6.11)). The block index s is omitted for the sake of simplicity in notation.

stemming from cross-view style variations. This involves augmenting the prototype with cross-domain variations through multi-view augmentation, to simulate the style discrepancy for one identity between query and gallery views. Thirdly, we highlight person-specific attributes to increase the feature discrimination while maintaining the overall consistency across all pedestrians. Our contributions are: 1. To our knowledge, our method pioneers the use of cross-domain variations to implicitly explore per-identity multi-view augmentation, so to encourage model learning to maximize invariant representations subject to cross-camera identity retrieval. 2. We formulate a principled mechanism CDVM to learn a context-aware generalizable ReID model sensitive to domain-specific cross-camera person-wise variations, optimizing jointly two competing criteria of generalizability and specificity. 3. The proposed new model outperforms existing SOTA methods by a large margin on a wide range of benchmarks.

6.2 Methodology

Assuming that there is a sample space that $\mathcal{D} = \{D^k\}_{k=1}^{K}$ available for training, each domain is composed of numerous labeled image pairs $D^k = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$, where N_k is the number of images in the domain D^k. The label space for the domain D^k is formulated by all the identity labels as $\mathcal{Y}^k = \{1, 2, \cdots, M_k\}$, where M_k is the number of identities in D^k. The label spaces for different domains are completely disjoint, *e.g* $\mathcal{Y}^i \neq \mathcal{Y}^j$ if $i \neq j$. The objective of generalizable ReID is to learn a feature extraction model f_θ that is capable of extracting discriminative representations $f_\theta(x)$ for retrieval among different identities. This inherently constitutes a heterogeneous zero-shot task due to the disjoint label spaces in training and testing.

6.2.1 Domain-invariant Knowledge Disentanglement

One premise of a generalizable model is the capacity to extract a domain-agnostic representation for the pedestrian image captured under arbitrary conditions. We fulfill this premise by deriving a global-local correlation module to explore the applicable consensus of identities among different domains. Specifically, the global-local module consists of numerous parallel branches: a global encoder applied for all the domains to learn the domain-invariant representation, and K local encoders specific to each domain to model domain-specific knowledge. The global-local encoders are designed as a plug-and-play component and incorporated into the feature extractor by replacing the final layer in each block. To ensure the local and global encoders learn distinct information, we propose to maximize the discrepancy of the parameter spaces. Specifically, as illustrated in Figure 6.2, assuming an intermediate feature $F_s^k \in \mathcal{R}^{B_k,C,H,W}$ extracted in the block s for the minibatch of samples $\mathcal{X}^k$ from domain D^k, it is fed into simultaneously two branches to disentangle the domain-invariant and domain-specific knowledge as

$$F_{\text{inv},s}^k = f_{(g,s)}(F_s^k), \quad F_{\text{spe},s}^k = f_{(l,s)}^k(F_s^k). \tag{6.1}$$

where $f_{(g,s)}$ and $f_{(l,s)}^k$ are the functionalities corresponding to the global and local branches at s-th block. The global and local branches are in the same structure with different parameter initialization. The global-local encoders are trained with the following constraints: (1) To disentangle the intermediate representation, they are constrained to learn distinct information by maximizing their discrepancy as

$$\mathcal{L}_{\text{ws}} = \frac{1}{KS} \sum_{k=1}^{K} \sum_{s=1}^{S} \text{cosine}(\theta_{(g,s)}, \theta_{(l,s)}^k), \tag{6.2}$$

where $\theta_{(g,s)}$ and $\theta^k_{(l,s)}$ are the learnable parameters respectively for $f_{(g,s)}$ and $f^k_{(l,s)}$. (2) To ensure the disentangled output from the global branch is domain-agnostic, we adopt adversarial training with a domain discriminator to maximize the likelihood of domain label from the latent representation $F^k_{\text{inv, s}}$, while $f_{\text{(g,s)}}$ aims to learn domain-invariant feature to fool the discriminator. This is performed by solving the following min-max game:

$$\mathcal{L}_{\text{inv}} = -\frac{1}{S}\sum_{s=1}^{S} \min_{\theta_{D_{(c,s)}}} \max_{\theta_{(g,s)}} k \log D_{(c,s)}(f_{(g,s)}(F^k_{\text{inv},s})), \tag{6.3}$$

where $D_{(c,s)}$ is the domain classifier parameterized by $\theta_{D_{(c,s)}}$. This is realized by applying a Gradient Reverse Layer (GRL) [28] on the domain-invariant knowledge to fool the domain classifier. (3) To ensure the disentangled output from the local branch is domain-specific, we maximize the probability predictions of domain label from $F^k_{\text{spe},s}$ by optimizing the following objective:

$$\mathcal{L}_{\text{spe}} = -\frac{1}{S}\sum_{s=1}^{S} k \log D_{(c,s)}(f^k_{(l,s)}(F^k_{\text{spe},s})). \tag{6.4}$$

The final disentanglement objective is formulated as

$$\mathcal{L}_{\text{dise}} = \mathcal{L}_{\text{ws}} + \mathcal{L}_{\text{inv}} + \mathcal{L}_{\text{spe}}. \tag{6.5}$$

The global encoder is specifically designed to extract a disentangled domain-agnostic prototype $F^k_{\text{inv,s}}$, as defined by Eq. (6.1). However, considering the potential for significant style shifts in testing samples, the prototype might exhibit limited robustness and discrimination, thereby restricting its ability to effectively represent a single identity in this case. Normalization techniques, *e.g* BN and IN, have been widely employed in recent generalizable ReID models to mitigate per-identity style disparities. Nevertheless, we contend that such approaches overlook the diverse style variations offered by samples from other domains which could be beneficial in learning a more robust model. Instead, we perform per-identity multi-view style augmentation using cross-domain normalization statistics. This is equivalent to introducing additional instances for one identity but in different styles mimicking the style variations present in query and gallery views.

Specifically, given the domain-specific knowledge $F^k_{\text{inv,s}}$, we derive the per-domain style characteristic by pooling the instance normalization statistics [64] over the current minibatch as:

$$\begin{aligned}\mu^k_s &= \frac{1}{B_k HW}\sum_{b=1}^{B_k}\sum_{h=1}^{H}\sum_{w=1}^{W} F^k_s(b,h,w),\\ (\sigma^k_s)^2 &= \frac{1}{B_k HW}\sum_{b=1}^{B_k}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(F^k_s(b,h,w)-\mu^k_s\right)^2,\end{aligned} \tag{6.6}$$

where B_k represents the number of samples in a mini-batch drawn from domain D^k. The statistical moments μ_s^k and $(\sigma_s^k)^2$ encode the characteristics of domain D^k. These statistics are modeled at a mini-batch level rather than instance-level, to offset the potential disruptions caused by outliers, *e.g* an image without a person. Instead of treating each of them as a determined point, to consider the randomness of the combinations, we further build a Gaussian distribution $\mathcal{N}(\hat{\mu_s^k}, \hat{\sigma_s^k}^2)$ with the $\hat{\mu_s^k}$ to indicate the expansion direction and $\hat{\sigma_s^k}$ for the intensity as

$$\begin{aligned}\hat{\mu_s^k} &= \frac{1}{K-1}\sum_{i=1,i\neq k}^{K} \mu_s^k + \varepsilon_\mu \delta(\mu_s^k),\\ \hat{\sigma_s^k} &= \frac{1}{K-1}\sum_{i=1,i\neq k}^{K} \sigma_s^k + \varepsilon_\sigma \delta(\sigma_s^k).\end{aligned} \tag{6.7}$$

where ε_μ and ε_σ are sampled from the normal distribution to vary the expansion direction and intensity. Moreover, $\delta(\cdot)$ calculates the variance to measure the diversity of the cross-domain statistics. The expanded cross-domain statistics $\hat{\mu_s^k}$ and $\hat{\sigma_s^k}$ encodes diverse style information sampled over disjoint domains. Therefore, the style information in $F^k_{\text{spe, s}}$ is modified by substituting the feature statistics as

$$F^k_{\text{sty,s}} = \hat{\sigma}_s^k \frac{F^k_{\text{spe, s}} - \mu_s^k}{\sigma_s^k} + \hat{\mu}_s^k. \tag{6.8}$$

Subsequently, the modified style information is fused with invariant ID knowledge through a Fully-Connected (FC) layer as

$$\hat{F_s^k} = \text{FC}\left(\text{cat}\left(F^k_{\text{inv},s}, F^k_{\text{sty,s}}\right)\right), \tag{6.9}$$

where cat$(\cdot)$ is the concatenation operator functions on the channel dimension, and FC$(\cdot)$ represents the FC layers which reduces the channel dimension of the concatenated representation from $2C$ to C. Therefore, the identity representation is expanded in various directions to achieve multi-view augmentation.

6.2.2 Local Hierarchical and Global Consistent Constraint

Given the augmented representation $\hat{F_s^k}$, we group it into subspaces, with the assumption that each group corresponds to specific characteristics essential for representing an identity. Intuitively, certain attributes, such as facial appearance and body structure, play a more dominant role in identifying a person compared to others. By emphasizing these influential characteristics, we aim to enhance the discriminative power of the learned representation. To this end, we introduce a subspace constraint that considers two critical aspects: (1) Per-identity local hierarchy: For one identity, the significance of characteristics should be weighted differently so as to emphasize different aspects of the feature that contribute to identity identification. (2) Cross-domain global consistent: Considering the universally applicable explanatory of pedestrians regardless of the domain annotation, the dominant characteristics in one domain should retain their importance when considering any other domain. We implicitly realize this constrain by slicing the augmented representations into subspaces (groups) along the channel dimension, and feeding them into a hyper-network to estimate the significance of each group with a set of predictions $\mathcal{W} = \{w_v\}_{v=1}^{V}$, where V is the number of subspaces. This constrain is mathematically formulated as

$$\begin{aligned}\mathcal{L}_{\mathrm{v}} &= \frac{1}{V \times (V-1)} \sum_{m=1}^{V} \sum_{n \neq m}^{V} [m_1 - \|w_m - w_n\|_2]_{+}^{2}, \\ \mathcal{L}_{\mathrm{c}} &= \frac{1}{B \times (B-1)} \sum_{i=1}^{B} \sum_{j \neq i}^{B} [\|\mathcal{W}_i - \mathcal{W}_j\|_2 - m_2]_{+}^{2},\end{aligned} \tag{6.10}$$

where B is the number of samples in a minibatch, and $\mathcal{W}_i$ is the importance prediction for the pedestrian i. The two hyperparameters m_1 and m_2 are the margins. The final characteristic constraint is the combined as

$$\mathcal{L}_{\mathrm{attr}} = \mathcal{L}_{\mathrm{c}} + \mathcal{L}_{\mathrm{v}}. \tag{6.11}$$

6.2.3 Model Training

Training Objectives The proposed CDVM is jointly trained with various objectives, including the conventional cross-entropy loss $\mathcal{L}_{\mathrm{ce}}$, triplet loss $\mathcal{L}_{\mathrm{tri}}$, center loss $\mathcal{L}_{\mathrm{cent}}$, the feature disentanglement loss $\mathcal{L}_{\mathrm{dise}}$, and the proposed attribute constraint $\mathcal{L}_{\mathrm{attr}}$.

$$\mathcal{L} = \mathcal{L}_{\mathrm{ce}} + \mathcal{L}_{\mathrm{tri}} + \mathcal{L}_{\mathrm{cent}} + \alpha \mathcal{L}_{\mathrm{dise}} + \beta \mathcal{L}_{\mathrm{attr}}, \tag{6.12}$$

where α and β are the hyperparameters to balance the importance of the corresponding learning objective.

Training Pipeline To improve the generalizability of the proposed model, we adopt the meta-learning algorithm as the training strategy to simulate the training-testing discrepancy. Given K source domains available during training, samples in $K-1$ domains are used as the meta-training set and the remaining domain is used as a meta-testing set. The parameters of the entire network are updated by the second-order gradient with respect to the meta-test loss.

6.3 Experiments

6.3.1 Experimental Settings

Datasets and protocols. We conducted multisource domain generalization on a wide range of 9 benchmarks, including five large-scale datasets: Market1501 (M) [179], MSMT17 (MT) [147], CUHK02 (C2) [84], CUHK03 (C3) [86], CUHK-SYSU (CS) [154], and four small-scale datasets: PRID [47], GRID [99], VIPer [34], and iLIDs [180]. For CUHK03, we used the “labeled” subset to keep a fair comparison with the State-of-the-Art (SOTA) competitors [156, 17, 177]. mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) accuracy on R1 are adopted as evaluation metrics.

Table 6.1: Performance comparisons of CDVM on under protocol-1. The best results are in **bold**.

Source	Method	→PRID		→GRID		→VIReR		→iLIDs		Average	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
D+M+C2 +C3+CS	DIMN [129]	52.0	39.2	41.1	29.3	60.1	51.2	78.4	70.2	57.9	47.5
	SNR [64]	66.5	52.1	47.7	40.2	61.3	52.9	89.9	84.1	66.4	57.3
	RaMoE [18]	67.3	57.7	54.2	46.8	64.6	56.6	90.2	85.0	69.1	61.5
	DMG-Net [4]	68.4	60.6	56.6	51.0	60.4	53.9	83.9	79.3	67.3	61.2
Protocol-1: M+C2+ C3+CS	QAConv$_{50}$ [90]	62.2	52.3	57.4	48.6	66.3	57.0	81.9	75.0	67.0	58.2
	M^{3}L [177]	65.3	55.0	50.2	40.0	68.2	60.8	74.3	65.0	64.5	55.2
	MetaBIN [17]	70.8	61.2	57.9	50.2	64.3	55.9	82.7	74.7	68.9	60.5
	META [156]	71.7	61.9	60.1	52.4	68.4	61.5	83.5	79.2	70.9	63.8
	CDVM	**74.1**	**64.8**	**66.1**	**56.0**	**69.6**	**63.6**	**87.7**	**83.1**	**74.4**	**66.9**

Implementation Details. Following the conventional settings [167, 102, 177], we used ResNet50 [40] with IBN [115] pre-trained on ImageNet to bootstrap the feature extractor. The batch size for each domain was set to 64, including 32 randomly sampled identities and 2 images for each identity. All images were resized to 256×128. We augmented the training data by random erase, flipping, and color jitter. The proposed CDVM was trained for 120 epochs with an SGD optimizer [67], and the warm-up strategy was adopted in the first 10 epochs to stabilize model training. The learning rate was initialized as 0.01 and decay to $5e-5$ by Cosine Annealing. The balancing factors λ and β in Eq. (6.12) were both set to 0.5. The margins m_1 and m_2 in Eq (6.11) were set to 0.1. The dimension of the ID representation is conventionally set to 2048. All the experiments were conducted on the PyTorch [116] framework with four A100 GPUs.

6.3.2 Comparative Evaluations

Comparison under Protocol-1. One established evaluation protocol [18, 64, 129], is to train on five large-scale datasets, *i.e* DukeMTMC [183], Market1501, CUHK02, CUHK03, and CUHK-SYSU and test on four small-scale datasets, *i.e* PRID, GRID, VIPeR, and iLIDs. However, due to the widely used DukeMTMC dataset was officially taken off due to privacy issues, recent works [156, 167] proposed a new protocol by removing DukeMTMC and using the remaining four datasets (M+C2+C3+CS) for training, called Protocol-1. Under this protocol, all the samples, regardless of the original training/testing splits, are used for training. We made a fair comparison with the SOTA competitors by performing 10-trial evaluations [64, 129] on the random split query/gallery sets, and reported the averaged results in Table 6.1. Compared to the other SOTA models trained with the same datasets, our model shows clear advantages and outperforms the latest SOTA model META [156] by 5.5% in mAP and 5.0% in Rank1 scores. Compared with the other SOTA methods trained including the DukeMTMC dataset, our method remains competitive.

Table 6.2: Performance comparisons of CDVM on under protocol-2.

Setting	Method	Reference	M+MS+CS→C3		M+CS+C3→MS		M+CS+C3→M		Average	
			mAP	R1	mAP	R1	mAP	R1	mAP	R1
Protocol-2: Training Sets	SNR [64]	CVPR20	8.9	8.9	6.8	19.9	34.6	62.7	16.8	30.5
	QAConv$_{50}$ [90]	ECCV20	25.4	24.8	16.4	45.3	63.1	83.7	35.0	51.3
	MetaBIN [17]	CVPR21	28.8	28.1	17.8	40.2	57.9	80.1	34.8	49.5
	M^{3}L [177]	CVPR21	34.2	34.4	16.7	37.5	61.5	82.3	37.5	51.4
	ACL [167]	ECCV22	41.2	41.8	20.4	45.9	74.3	89.3	45.3	59.0
	META [156]	ECCV22	36.3	35.1	**22.5**	**49.9**	67.5	86.1	42.1	57.0
	CDVM	Ours	**41.7**	**42.8**	20.7	46.4	**74.8**	**89.8**	**45.4**	**59.7**

Comparison under protocol-2 and protocol-3. The proposed CDVM model was further evaluated on four *large-scale* datasets with a leave-one-out strategy, *i.e* using three domains for training and the left one for testing. Note that due to all the identities in CUHK-SYSU are captured by the same camera, it was only used for training. For protocol-2, only the train splits of these datasets were leveraged for training. In contrast, for protocol-3, all the available labeled samples, regardless of the original splits, were used in training. We reported the comparison results in Table 6.2 and Table 6.3. It can be observed that the proposed CDVM model achieves superior performance when generalizing to CUHK03 and Market1501 and remains competitive when MSMT17 was

Table 6.3: Performance comparisons of CDVM on under protocol-3.

Setting	Method	Reference	M+MS+CS→C3		M+CS+C3→MS		M+CS+C3→M		Average	
			mAP	R1	mAP	R1	mAP	R1	mAP	R1
Protocol-3: Full Sets	SNR [64]	CVPR20	17.5	17.1	7.7	22.0	52.4	77.8	25.9	39.0
	QAConv$_{50}$ [90]	ECCV20	32.9	33.3	17.6	46.6	66.5	85.0	39.0	55.0
	MetaBIN [17]	CVPR21	43.0	43.1	18.8	41.2	67.2	84.5	43.0	56.3
	M^{3}L [177]	CVPR21	35.7	36.5	17.4	38.6	62.4	82.7	38.5	52.6
	ACL [167]	ECCV22	49.4	50.1	21.7	47.3	76.8	90.6	49.3	62.7
	META [156]	ECCV22	47.1	46.2	**24.4**	**52.1**	76.5	90.5	49.3	62.9
	CDVM	Ours	**50.9**	**50.7**	22.6	50.1	**77.6**	**90.8**	**50.4**	**63.9**

leveraged as the target domain. This illustrates that the CDVM model can benefit more when more identities are available to provide abundant style variations in training.

6.3.3 Ablation Study

Table 6.4: Components analysis. The proposed components were progressively incorporated into the baseline to study the individual contribution.

Components			CUHK03		MSMT17		Market1501	
$\mathcal{L}_{\text{dent}}$	f_{aug}	$\mathcal{L}_{\text{attr}}$	mAP	R1	mAP	R1	mAP	R1
✗	✗	✗	33.9	34.2	17.5	43.1	69.5	87.2
✓	✗	✗	36.6	37.1	18.1	44.3	71.6	88.2
✓	✓	✗	40.9	41.6	19.5	45.7	73.4	88.9
✓	✓	✓	41.7	42.8	20.7	46.6	74.8	89.8

Components analysis. We investigated the individual contribution of different components in the CDVM model to study its effectiveness. As shown in Table 6.4, the performance was progressively improved by incorporating the proposed constraints. Specifically, introducing the disentanglement loss $\mathcal{L}_{\text{dent}}$ can reduce the domain gap compared with the baseline model. Further performing cross-domain style augmentation improved the model's robustness against the potential style variations and so to make the representations more robust to cross-camera view variations in specificity. Finally, employing the attribute constraint further improved the discrimination capacity of the learned representation.

Global-Local Discripency Constraint. One premise for decoupling the domain-invariant and domain-specific knowledge from the learned representation is that the global and local branches are learning distinctive information. To achieve this goal, we designed the constraint $\mathcal{L}_{\text{ws}}$ to explicitly enlarge the discrepancy of the learnable parameters between the global and local encoders. We ablated its effectiveness in knowledge disentanglement. The comparison result is shown in Table 6.5. By employing the discrepancy constraint $\mathcal{L}_{\text{ws}}$, the performance is consistently improved on all the benchmarks. This shows the inadequacy of the conventional disentanglement design and the potential advantages of optimizing jointly both the generalizability and specificity criteria by this discrepancy constraint.

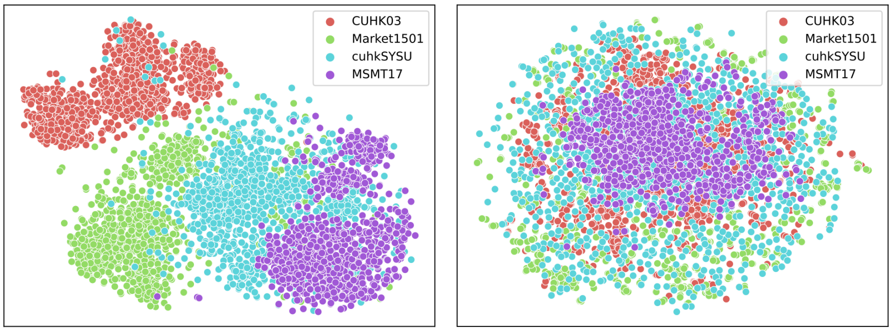
Table 6.5: Effects of the discrepancy constraint in feature disentanglement.

Components	CUHK03		MSMT17		Market1501	
	mAP	R1	mAP	R1	mAP	R1
w/o $\mathcal{L}_{\text{ws}}$	35.1	35.8	17.8	43.6	70.1	87.1
w $\mathcal{L}_{\text{ws}}$	36.6	37.1	18.1	44.3	71.6	88.2

Cross-Domain Variation Expansion. To encourage the model to be robust against style variation, we performed cross-domain multi-view feature augmentation by sampling the style factors over the cross-domain statistics. We considered the randomness and combinations of the cross-domain activations to achieve a more diverse augmentation. We validated the superiority of this augmentation strategy over the vanilla counterpart, *i.e* treat the statistics as determined factors without any expansion, and the results are reported in Table 6.6. We observed that (1) Considering the cross-domain variations improves the performance compared with the baselines, which verifies our assumption that enhance the diversity of identity is beneficial for learning a robust and generalization model. (2) Compared with taking the cross-domain variations as determined factors, exploiting the elastic expansion with random direction and intensity can yield better results. This validates the superiority of the proposed cross-domain expansion strategy.

Table 6.6: Effects of different strategies in exploring cross-domain variations. “Determined” takes variations as fixed factors, “Elastic” considers combinations and randomness.

Method	CUHK03		MSMT17		Market1501	
	mAP	R1	mAP	R1	mAP	R1
Determined	37.3	38.1	19.0	45.1	72.0	88.3
Elastic	40.9	41.6	19.5	45.7	73.4	88.9



(a) Baseline

(b) Ours

Figure 6.3: T-SNE visualization of four ReID benchmarks. The target domain is set to MSMT17, and the remaining three are used for training. The baseline model learns domain-biased representations, while our model exploits cross-domain variations and extracts domain-invariant representations.

Visualization. To further validate the effectiveness of the proposed model, we conducted t-SNE visualization on the representations extracted by different models. The target domain was MSMT17 and the other three domains were leveraged for training. We sampled 1000 instances in each domain. Results are shown in Figure 6.3. From it, we observed that the baseline model is prone to learning domain-bias representations while the proposed CDVM model is more robust in extracting domain-invariant representation.

6.4 Summary

In this chapter, we presented a novel CDVM model to learn a generalizable ReID representation that simultaneously optimizes model generalizability and specificity. The motivation of CDVM model design is that the cross-domain variations can be used to perform multi-view augmentation on one identity, so as to simulate the style variations between the query and gallery views. To achieve this goal, we first explored cross-domain consensus to learn a domain-agnostic prototype which is then optimized with cross-domain variations for implicitly multi-view feature augmentation. Moreover, we further boosted the discrimination of the augmented representation by formulating an identity attribute constraint to reassemble the representation considering individual attribute significance. We validated the effectiveness of the proposed CDVM model extensively on 9 benchmark datasets. We show that the proposed new model outperforms existing SOTA methods by a notable margin.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

1. **In Chapter 3**, we introduced a Local-Global Associative Assembling (LOGA) method for video person Re-Identification (ReID), which selectively assembles video frames of diverse qualities to derive a more reliable and discriminative representation of a video tracklet. This is achieved by assessing the frames' quality based on the association of their local part alignments and global appearance correlation, to avoid integrating undesired visual information into the tracklet's representation, thus preventing identity mismatch. Unlike existing approaches that focus on either local or global information separately, our LOGA method constructs a locally assembled global appearance prototype of a tracklet to mitigate biased quality assessment resulting from identity-irrelevant misalignment or spatially insensitive appearance miscorrelation. Extensive experiments on four benchmark datasets demonstrate the performance advantages of LOGA over a wide range of State-of-the-Art (SOTA) video ReID methods. Furthermore, detailed ablation studies are conducted to provide in-depth discussions about the rationale and essence of different components in our model design.

2. **In Chapter 4**, we introduced a regularization technique called Primary-Auxiliary Objectives Association (PAOA) to facilitate the learning of a generalizable ReID model capable of extracting domain-unbiased representations that are more adaptable to unseen novel domains for person ReID. PAOA encourages the model to eliminate the influence of

domain-specific knowledge and focus on learning discriminative pedestrian information by associating the learning of an auxiliary pedestrian detection objective with a primary instance classification objective. To address the challenges posed by noisy auxiliary labels, we further developed a referenced-gradient calibration strategy to adjust the gradient of the auxiliary object when it conflicts with the primary object. The PAOA framework is task-agnostic, making it easily adaptable to other tasks by incorporating a related auxiliary task and a shared learning module.

3. **In Chapter 5**, we presented a Feature-Distribution Perturbation and Calibration (PECA) model to learn a generic yet discriminative representation across multiple source domains, with the aim of generalizing to arbitrary unseen target domains for a more accurate person Re-Identification (ReID) in unseen domains. PECA conducts simultaneous model regularization on local per-domain feature distributions and global cross-domain feature distributions to achieve a better domain-invariant feature space representation. Leveraging diverse features synthesized by local perturbation, PECA expands per-domain feature distribution to enhance robustness to domain shifts. Through global calibration, feature distributions of different domains are holistically represented and referenced in a shared feature space, thereby ignoring domain-specific data characteristics (*e.g* mean and variance of feature distributions) and resulting in higher model generalizability. Experiments on extensive ReID datasets demonstrate the performance advantages of the proposed PECA model over a wide range of SOTA competitors. Additionally, extensive ablation studies provide an in-depth analysis of the individual components in the PECA model.

4. **In Chapter 6**, we presented a novel Cross-Domain Variations Mining (CDVM) model aimed at learning a generalizable ReID representation that optimizes both model generalizability and specificity simultaneously. The motivation of CDVM model is that cross-domain variations can be leveraged to perform multi-view augmentation on a single identity, thereby simulating the style variations between query and gallery views. To achieve this objective, we first explored cross-domain consensus to learn a domain-agnostic prototype, which is then optimized with cross-domain variations for implicit multi-view feature augmentation. Furthermore, we enhanced the discrimination of the augmented representation by formulating an identity attribute constraint to reassemble the representation while considering the significance of individual attributes. We extensively validated the effec-

tiveness of the proposed CDVM model on nine benchmark datasets, demonstrating that our model outperforms existing SOTA methods by a notable margin.

7.2 Future Work

1. **Enhancing Discriminative Frame Selection with Optical Flow** In Chapter 3, we evaluated frame quality by jointly analyzing local region alignment and global appearance correlation. This approach filters out low-quality frames, thereby improving the discriminative power of aggregated tracklet-level representations. However, this strategy presupposes that the network is capable of establishing temporal correlations through the learning process by integrating information from individual frames. However, DNNs might struggle to fulfill this objective when the frame rate is relatively low, potentially failing to utilize contemporary information from adjacent frames. To address this, a promising direction for future research is to investigate optical flow [92] to identify pedestrian movement patterns. By leveraging optical flow as a guide to discern movement-induced occlusions, so as to enhance scene change detection and further refine the discriminative frame selection process.

2. **Promotive Multitask Learning by Mixture of Auxiliary Experts** In Chapter 4, we designed a primary-auxiliary learning framework termed as PAOA. In this framework, the auxiliary task is designed as a pedestrian saliency detection task for which labels are generated on the fly by a pretrain saliency detection model. However, such labels may not be reliable, particularly when pedestrians are obscured by large objects or when multiple pedestrians are present in the image simultaneously. To solve this issue, one possible solution is to leverage the advancements in large-scale foundational models, like Grounding DINO [94] for object localization and SAM [68] for semantic segmentation. By integrating additional auxiliary tasks into our model structure, we can create an auxiliary-expert joint learning framework to provide multifaceted deep image priors, and inherently refine inaccurate labels.

3. **Learnable Feature Distribution Local Perturbation** In Chapter 5, we developed the local feature perturbation module to expand the feature space within each local domain, which is complemented by a global calibration process that aligns the cross-domain representations into a unified feature space. The local perturbation module utilizes a Gaus-

sian distribution to introduce variations within the overall feature distribution. However, this Gaussian distribution may not accurately reflect the actual diversity within domains in real-world scenarios, which could potentially deteriorate feature representation. A potential remedy to mitigate this problem is to employ a trainable meta-network [110] that captures the diversity within domains, based on which to expand per-domain feature distribution, thereby preserving the discriminative while enhancing the diversity.

4. **Open-vocabulary ReID with Pretrained Vision-Language Models** In Chapter 6, we introduced the CDVM model to improve its robustness against domain shift. by CDVM exploring cross-domain variations to mimic the style variations of an identity captured by different cameras. However, CDVM is a closed-vocabulary approach where the model is trained with a pre-defined set of identities known with limited attribuduring training. This restricts the model's ability to adapt to new or unseen identities encountered in real-world scenarios and makes it less effective in handling open-vocabulary situations where the identities are not predefined or known beforehand. To address this limitation, a promising solution is to leverage large-scale pretrained vision-language models [93, 77] trained on vast amounts of data across different domains and modalities. These models encode rich semantic information from images and textual descriptions, enabling them to infer relationships between identities and adapt to new or unseen identities encountered during inference. Therefore, it can enhance the model's ability to handle open-vocabulary scenarios by providing more comprehensive and discriminative representations.

Bibliography

[1] Jon Almazan, Bojana Gajic, Naila Murray, and Diane Larlus. Re-id done right: towards good practices for person re-identification. *arXiv preprint arXiv:1801.05339*, 2018.

[2] Eugene PW Ang, Lin Shan, and Alex C Kot. Dex: Domain embedding expansion for generalized person re-identification. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021.

[3] Haoyue Bai, Rui Sun, Lanqing Hong, Fengwei Zhou, Nanyang Ye, Han-Jia Ye, S-H Gary Chan, and Zhenguo Li. Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[4] Yan Bai, Jile Jiao, Wang Ce, Jun Liu, Yihang Lou, Xuetao Feng, and Ling-Yu Duan. Person30k: A dual-meta generalization network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] Liqiang Bao, Bingpeng Ma, Hong Chang, and Xilin Chen. Preserving structural relationships for person re-identification. In *International Conference on Multimedia & Expo Workshops (ICMEW)*, 2019.

[6] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3286–3295, 2019.

[7] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media*, 5:117–150, 2019.

[8] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[11] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Feng Chen, Nian Wang, Jun Tang, Dong Liang, and Hao Feng. Self-supervised data augmentation for person re-identification. *Neurocomputing*, 415:48–59, 2020.

[13] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Spatial-temporal attention-aware learning for video-based person re-identification. *IEEE Transactions on Image Processing*, 28(9):4192–4205, 2019.

[14] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Zengqun Chen, Zhiheng Zhou, Junchu Huang, Pengyu Zhang, and Bo Li. Frame-guided region-aligned representation for video person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[16] Yeong-Jun Cho and Kuk-Jin Yoon. Pamm: Pose-aware multi-shot matching for improving person re-identification. *IEEE Transactions on Image Processing*, 27(8):3739–3752, 2018.

[17] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[18] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[20] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Towards unified human parsing and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[21] Cian Eastwood, Ian Mason, Christopher KI Williams, and Bernhard Schölkopf. Source-free adaptation to measurement shift via bottom-up feature restoration. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[22] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[23] Sarah Erfani, Mahsa Baktashmotlagh, Masud Moshtaghi, Xuan Nguyen, Christopher Leckie, James Bailey, and Rao Kotagiri. Robust domain generalisation by enforcing distribution invariance. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[24] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[26] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[27] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[28] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[29] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1414–1430, 2016.

[30] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press Cambridge, 2016.

[31] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[33] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proceedings of IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, 2007.

[34] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008.

[35] Hongyang Gu, Jianmin Li, Guangyuan Fu, Chifong Wong, Xinghao Chen, and Jun Zhu. Autoloss-gms: Searching generalized margin-based softmax loss function for person reidentification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[36] Jianyang Gu, Kai Wang, Hao Luo, Chen Chen, Wei Jiang, Yuqiang Fang, Shanghang Zhang, Yang You, and Jian Zhao. Msinet: Twins contrastive search of multi-scale interaction for object reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[37] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[38] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[39] Ke Han, Chenyang Si, Yan Huang, Liang Wang, and Tieniu Tan. Generalizable person re-identification via self-supervised batch norm test-time adaption. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[41] Lingxiao He and Wu Liu. Guided saliency feature learning for person re-identification in crowded scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[42] Tianyu He, Xin Jin, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Dense interaction learning for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[43] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[44] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[45] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017.

[46] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102, 2011.

[47] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis Scandinavian Conference*, pages 91–102, 2011.

[48] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

[49] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[50] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[51] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[52] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Feature completion for occluded person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4894–4912, 2021.

[53] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[54] Panwen Hu, Jiazhen Liu, and Rui Huang. Concentrated multi-grained multi-attention network for video based person re-identification. *arXiv preprint arXiv:2009.13019*, 2020.

[55] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, pages 292–302. PMLR, 2020.

[56] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[57] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[58] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*. PMLR, 2020.

[59] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[60] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2012.

[61] Jieru Jia, Qiuqi Ruan, and Timothy M Hospedales. Frustratingly easy person re-identification: Generalizing person re-id in practice. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019.

[62] Bingliang Jiao, Lingqiao Liu, Liying Gao, Guosheng Lin, Lu Yang, Shizhou Zhang, Peng Wang, and Yanning Zhang. Dynamically transformed instance normalization network for generalizable person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[63] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Feature alignment and restoration for domain generalization and adaptation. *arXiv*, 2020.

[64] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[65] Nikolaos Karianakis, Zicheng Liu, Yinpeng Chen, and Stefano Soatto. Person depth reid: Robust person re-identification with commodity depth sensors. *arXiv preprint arXiv:1705.09882*, 2017.

[66] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[67] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[68] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[69] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4422–4431, 2019.

[70] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[71] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[72] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[73] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[74] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[75] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[76] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[77] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on machine learning (ICML)*, 2023.

[78] Mengliu Li, Han Xu, Jinjun Wang, Wenpeng Li, and Yongli Sun. Temporal aggregation with clip-level attention for video-based person re-identification. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, 2020.

[79] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[80] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1770–1782, 2019.

[81] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[82] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[83] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[84] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[85] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2013.

[86] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[87] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[88] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[89] Xingze Li, Wengang Zhou, Yun Zhou, and Houqiang Li. Relation-guided spatial attention and temporal refinement for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[90] Shengcai Liao and Ling Shao. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[91] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. *arXiv preprint arXiv:1908.01683*, 2019.

[92] Daizong Liu, Xiang Fang, Wei Hu, and Pan Zhou. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. *IEEE Transactions on Multimedia*, 25:8539–8553, 2023.

[93] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[94] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[95] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[96] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[97] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[98] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[99] Chen Change Loy, Tao Xiang, and Shaogang Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90:106–129, 2010.

[100] Chuanchen Luo, Chunfeng Song, and Zhaoxiang Zhang. Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[101] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[102] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019.

[103] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *Proceedings of the International Conference on machine learning (ICML)*, 2021.

[104] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. Aggregating deep pyramidal representations for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 0–0, 2019.

[105] Neeraj Matiyali and Gaurav Sharma. Video person re-identification using learned clip similarity aggregation. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, 2020.

[106] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[107] Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of disentanglement in generalisation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[108] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017.

[109] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the International Conference on machine learning (ICML)*, 2013.

[110] Ivona Najdenkoska, Xiantong Zhen, and Marcel Worring. Meta learning to bridge vision and language models for multimodal few-shot learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[111] Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[112] Hao Ni, Jingkuan Song, Xiaopeng Luo, Feng Zheng, Wen Li, and Heng Tao Shen. Meta distribution alignment for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[113] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[114] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210, 2010.

[115] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[116] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Proceedings of the Conference on Neural Information Processing Systems Workshop(NeurIPS-W)*, 2017.

[117] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[118] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[119] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

[120] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[121] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[122] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 31, 2018.

[123] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[124] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 486–504, 2018.

[125] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[126] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[127] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[128] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[129] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[130] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[131] Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3), 2019.

[132] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the International Conference on machine learning (ICML)*, 2020.

[133] Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li. Mhsa-net: Multihead self-attention network for occluded person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[134] Lei Tan, Pingyang Dai, Rongrong Ji, and Yongjian Wu. Dynamic prototype mask for occluded person re-identification. In *Proceedings of the 30th ACM international conference on multimedia*, 2022.

[135] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[136] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv*, 2016.

[137] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, 2017.

[138] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

[139] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A Olshausen, and Trevor Darrell. Fully test-time adaptation by entropy minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[140] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[141] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[142] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[143] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[144] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[145] Zhikang Wang, Lihuo He, Xiaoguang Tu, Jian Zhao, Xinbo Gao, Shengmei Shen, and Jiashi Feng. Robust video-based person re-identification by hierarchical mining. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[146] Mirza M Xu B Warde-Farley and D Ozair S Courville A Bengio. Y goodfellow ij, pouget-abadie j. generative adversarial nets. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[147] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[148] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[149] Guile Wu and Shaogang Gong. Decentralised learning from independent multi-domain labels for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[150] Guile Wu and Shaogang Gong. Generalising without forgetting for lifelong person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[151] Guile Wu, Xiatian Zhu, and Shaogang Gong. Spatio-temporal associative representation for video person re-identification. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019.

[152] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[153] Wangmeng Xiang, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Part-aware attention network for person re-identification. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020.

[154] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv*, 2016.

[155] Boqiang Xu, Lingxiao He, Jian Liang, and Zhenan Sun. Learning feature recovery transformer for occluded person re-identification. *IEEE Transactions on Image Processing*, 31:4651–4662, 2022.

[156] Boqiang Xu, Jian Liang, Lingxiao He, and Zhenan Sun. Mimic embedding via adaptive aggregation: Learning generalizable person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[157] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[158] Cheng Yan, Guansong Pang, Xiao Bai, Changhong Liu, Xin Ning, Lin Gu, and Jun Zhou. Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. *IEEE Transactions on Multimedia*, 24:1665–1677, 2021.

[159] Fu-En Yang, Yuan-Chia Cheng, Zu-Yun Shiau, and Yu-Chiang Frank Wang. Adversarial teacher-student representation learning for domain generalization. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[160] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 2019.

[161] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, Jun Chen, and Jun Liu. Specific person retrieval via incomplete text description. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ACMICMR)*, 2015.

[162] Shijie Yu, Feng Zhu, Dapeng Chen, Rui Zhao, Haobin Chen, Shixiang Tang, Jinguo Zhu, and Yu Qiao. Multiple domain experts collaborative learning: Multi-source domain generalization for person re-identification. *arXiv*, 2021.

[163] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[164] Jianfu Zhang, Naiyan Wang, and Liqing Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[165] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. When unseen domain generalization is unnecessary? rethinking data augmentation. *arXiv preprint arXiv:1906.03347*, 2019.

[166] Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[167] Pengyi Zhang, Huanzhang Dou, Yunlong Yu, and Xi Li. Adaptive cross-domain learning for generalizable person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[168] Ruimao Zhang, Jingyu Li, Hongbin Sun, Yuying Ge, Ping Luo, Xiaogang Wang, and Liang Lin. Scan: Self-and-collaborative attention network for video person re-identification. *IEEE Transactions on Image Processing*, 28(10):4870–4882, 2019.

[169] Tianyu Zhang, Longhui Wei, Lingxi Xie, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Spatiotemporal transformer for video-based person re-identification. *arXiv preprint arXiv:2103.16469*, 2021.

[170] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv*, 2017.

[171] Yi-Fan Zhang, Zhang Zhang, Da Li, Zhen Jia, Liang Wang, and Tieniu Tan. Learning domain invariant representations for generalizable person re-identification. *IEEE Transactions on Image Processing*, 32:509–523, 2022.

[172] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

[173] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[174] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[175] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[176] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[177] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[178] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[179] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

[180] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2009.

[181] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[182] Zhedong Zheng and Yi Yang. Person re-identification in the 3d space. *arXiv preprint arXiv:2006.04569*, 2020.

[183] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017.

[184] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust partial matching for person search in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[185] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[186] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

[187] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5056–5069, 2021.

[188] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[189] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[190] Zijie Zhuang, Longhui Wei, Lingxi Xie, Haizhou Ai, and Qi Tian. Camera-based batch normalization: An effective distribution alignment method for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):374–387, 2021.

[191] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[192] Yang Zou, Xiaodong Yang, Zhiding Yu, BVK Vijaya Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.