Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Leverage cross-domain variations for generalizable person ReID representation learning

Qilei Li[a,b], Shitong Sun[c,*], Weitong Cai[c], Shaogang Gong[c]

[a] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, 430079, Hubei, China
[b] National Engineering Research Center for Educational Big Data, Central China Normal University, Wuhan, 430079, Hubei, China
[c] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, UK

ARTICLE INFO

Keywords:
Generalizable model
Person re-identification
Cross-domain style variations
Feature augmentation

ABSTRACT

Generalizable person ReID has significant practical value in challenging the fragile i.i.d. assumption by learning a domain-generalizable person representation applicable to out-of-distribution test samples. Existing 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 work, we formulate Cross-Domain Variations Mining (CDVM) 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 modelled to represent per-identity cross-view appearance changes. This approach 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 state-of-the-art methods by significant margins.

1. Introduction

Person Re-identification (ReID) aims to retrieve a specific identity across disjoint camera views. Over the last decade, this field has garnered significant attention due to its wide-ranging practical applications [1]. Advances in Convolution Neural Networks (CNNs) have notably contributed to enhancing ReID performance, particularly when the training and test data are drawn from the same distribution [2]. 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 [3].

Generalizable ReID [4] aims to address the problem when test data distribution may be different from that of the training data. This scenario poses inherent challenges due to the absence of prior knowledge about the unseen test data from target domains, therefore prohibiting distribution alignment during training. 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. 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 [4]. Alternatively, one assumes that the domain gap is mainly caused by style (appearance) variations [5] which can be mitigated with batch or/and instance normalization [6]. 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.

[*] Corresponding author.
E-mail address: shitong.sun@qmul.ac.uk (S. Sun).

https://doi.org/10.1016/j.patcog.2025.112292
Received 30 October 2024; Received in revised form 26 July 2025; Accepted 11 August 2025
Available online 15 August 2025

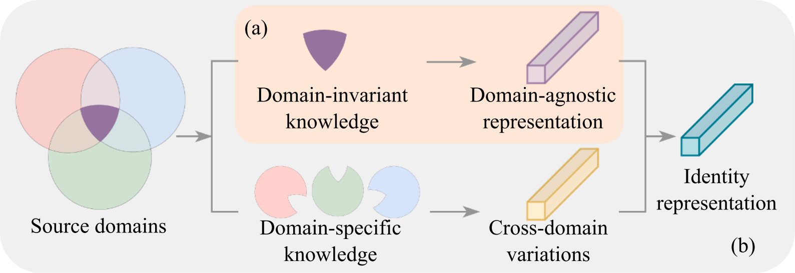

(a)
Domain-invariant knowledge
Domain-agnostic representation
Source domains
Domain-specific knowledge
Cross-domain variations
Identity representation
(b)

Fig. 1. Comparison of disentanglement learning-based models and the proposed CDVM model. (a). Disentangled representation learning-based methods solely utilize domain-invariant knowledge, making 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.

Firstly, they inevitably diminish contextual information and sacrifice the discrimination of the identity representation [7]. 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 [2]. 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 [8] or adversarial diversification [9]. 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 work, we introduce a new *Cross-Domain Variations Mining* (CDVM) model to overcome these limitations. 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 comparison of the CDVM model to conventional disentanglement learning-based models is illustrated in Fig. 1. 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 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:

- To our best 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.
- We formulate a principled mechanism CDVM to learn a context-aware generalizable ReID model sensitive to cross-camera person-wise variations, optimizing 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.

The rest of the paper is structured as follows: Section 2 presents the related work. Section 3 illustrates the proposed method in detail. Section 4 analyses the experimental results. The paper is concluded in Section 5.

2. Related works

Generalizable Person ReID. Despite the great progress made over recent years, most existing ReID methods [10] are built upon the fragile i.i.d assumption and their performance degrades significantly when deployed on a new test domain due to the covariant shift. To solve this problem, Generalizable ReID, has garnered increasing attention in recent years as a potential solution. Existing methods are generally categorized into three groups. The first line of these methods [3,4] 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 [6]. 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 mimics the training-testing discrepancy has been widely studied to enable the extracted features to be domain-agnostic [11]. On the other hand, ensemble learning-based techniques often aggregate descriptors derived from by multiple experts to assemble a more resilient representation [12]. Despite their efficacy, demonstrated effectiveness, these two strategies have limitations in effectively managing cross-domain conflicts and exploring cross-domain correlations. In this paper, 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.

Disentangled Representation Learning. The objective of disentanglement learning [13] is to explore the distinct and explanatory components, decoupling a representation into domain-invariant and domain-specific parts. It is generally achieved through adversarial training [14] where the aim is to deceive a domain discriminator, to enable the learned features to be domain-agnostic. Alternatively, VAE [15] can be employed to model a normal distribution and create a shared space, thus aligning the learned features. Disentanglement has also been studied for the generalization of person ReID. For instance, EOM [3] designed a disentanglement module incorporating a cycle-consistency constraint, while Zhang [4] et al. constructed 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 [16]. In this paper, we design a disentanglement module constrained by maximizing the consensus of domain-shared knowledge so as to learn an identity prototype that is domain-agnostic.

Data Augmentation. Training a neural network with diverse data can improve its generalizability on new, unseen domains [17] thereby improving its robustness against spurious correlations. Data augmentation [18] serves as a cost-effective method to enrich data diversity. Traditional data augmentations were most commonly applied within the raw image space, often through geometric transformations or random erasing. The emergence GANs [19] has enmabled the generation of new, realistic augmented counterparts featuring different contents or styles. More recent studies [20] have explored semantic transformations by directly manipulating the feature descriptor. For instance, Li et al. [8] proved training with features perturbed by Gaussian noise can facilitate the creation of a robust decision boundary. In a separate study, Li et al. [20] manipulated feature distribution by modeling the uncertainty of samples within a minibatch. Huang et al. [21] proposed learning beneficial noise as a form of graph augmentation, where the noise is explicitly optimized to improve generalization. Zhang et al. [22] framed data augmentation as a process of estimating positive-incentive noise for contrastive learning. In this paper, 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.

Batch/Instance Normalization. Batch Normalization (BN) [23] has been widely adopted by contemporary deep models, normalizing intermediate features using statistics computed over the samples in the minibatch. Instance Normalization (IN) [24] is a variant aimed at reducing style variation by holistically shifting per-instance activation moments. IN [6] in conjunction with BN [7] have also found adoption in ReID to eliminate style information associated with identity. However, considering that a specific identity captured by disjoint cameras showcases distinct styles [12], a model trained solely with normalized features is limited in extracting a discriminative representation against such style variations. Additionally, IN might dilute essential complementary information that is crucial for general visual recognition. In this paper, 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.

3. Methodology

Problem Definition. In generalizable person ReID, the goal is to train a model on multiple labeled source domains such that it performs well on unseen target domains, without accessing any target data during training. Let the source domain set be $\mathcal{D} = \{D^k\}_{k=1}^{K}$, where each domain D^k contains N_k labeled pedestrian images: $D^k = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$, and $y_i^k \in \mathcal{Y}^k$ is the identity label. Each domain has a unique identity label space, i.e., $\mathcal{Y}^i \cap \mathcal{Y}^j = \emptyset$ for $i \neq j$. The task is to learn a feature extractor f_θ that maps an image x to a discriminative feature $f_\theta(x)$, enabling reliable retrieval of pedestrian images across disjoint domains and identities. This setting constitutes a heterogeneous zero-shot learning problem due to the complete disjointness between the training and test identities. The challenge lies in learning representations that are both discriminative and robust to domain shifts caused by variations in camera views, environments, and styles.

3.1. Overview

The objective of the proposed method is to improve the model's robustness in learning domain-agnostic representations by considering cross-domain style variations as prompts, aiming to model the style shift for an identity captured by different cameras. This objective is archived progressively from three aspects, namely: generalizability, robustness, and discrimination. Initially, we learn a domain-invariant prototype by factorizing the intermediate features with the global-local encoders. Subsequently, we enhance model's robustness by exploring expanded cross-domain variations to simulate the style variations of the same identity captured by disjoint cameras. Finally, we implicitly learn the importance of identity attributes to improve feature discrimination, highlighting the most dominant attributes in identifying a person while maintaining consistency across all pedestrians.

3.2. Domain-invariant knowledge disentanglement

We define *domain-invariant information* as identity-discriminative features that remain stable across all domains (e.g., body structure, clothing texture), and are unaffected by environmental conditions. In contrast, *domain-specific information* refers to the style-related variations-such as illumination, resolution, occlusion, and camera viewpoint-that are characteristic of a particular domain. These style factors are critical in generalizable ReID, as they can mislead a model into learning domain-biased representations. In our model, we disentangle these two types of features to retain identity relevance while adapting to domain-specific styles. One premise of a generalizable model is the capacity to extract a domain-agnostic representation for the pedestrian image captured under arbitrary conditions. We fulfil 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 Fig. 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

$$F_{\text{inv},s}^k = f_{(g,s)}\left(F_s^k\right), \quad F_{\text{spe},s}^k = f_{(l,s)}^k\left(F_s^k\right), \tag{1}$$

where $f_{(g,s)}$ and $f_{(l,s)}^k$ are the functionalities corresponding to the global and local branches at sth 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

$$\mathcal{L}_{\text{ws}} = \frac{1}{KS}\sum_{k=1}^{K}\sum_{s=1}^{S}\text{cosine}\left(\theta_{(g,s)}, \theta_{(l,s)}^k\right), \tag{2}$$

where $\theta_{(g,s)}$ and $\theta_{(l,s)}^k$ are the learnable parameters respectively for $f_{(g,s)}$ and $f_{(l,s)}^k$. (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 $F_{\text{inv, s}}^k$, while $f_{(\text{g,s})}$ aims to learn domain-invariant feature to fool the discriminator. To ensure training stability, we adopt the gradient reversal layer (GRL) [14], which integrates adversarial optimization into the forward-backward pass without requiring alternating updates. This, combined with a warm-up strategy and cosine learning rate schedule, helps avoid training divergence and facilitates smooth convergence. This is performed by solving the following min-max game:

$$\mathcal{L}_{\text{inv}} = -\frac{1}{S}\sum_{s=1}^{S}\min_{\theta_{D_{(c,s)}}}\max_{\theta_{(g,s)}} k\log D_{(c,s)}\left(f_{(g,s)}\left(F_{\text{inv},s}^k\right)\right), \tag{3}$$

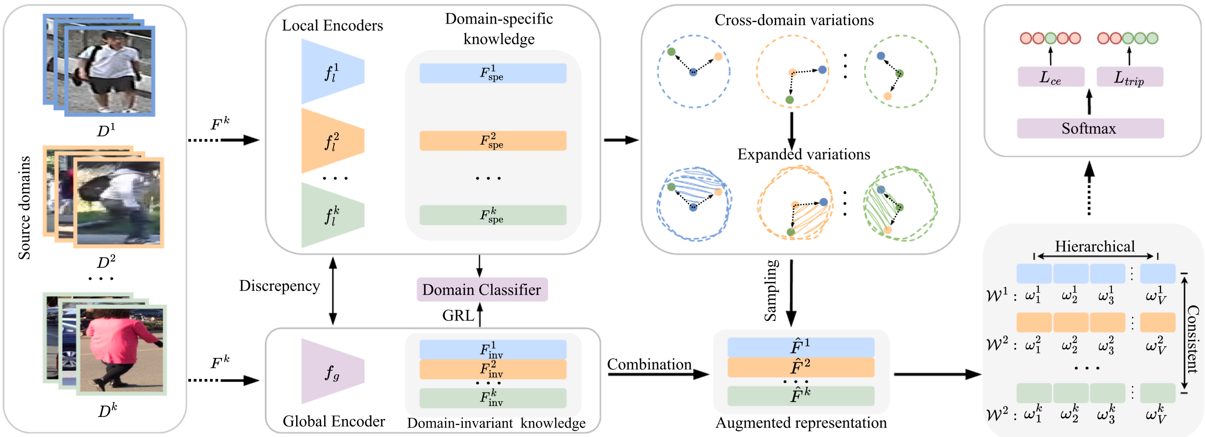


Fig. 2. Overview of the proposed *Cross-Domain Variations Mining* (CDVM) model. The overall objective is to enhance the model's robustness against domain shift when applied to an unseen environment. This is achieved by disentangling *domain-invariant features*, which encode identity-related characteristics consistent across domains, and *domain-specific features*, which primarily capture style-related variations such as illumination, resolution, and camera viewpoints. The model then applies cross-domain variations to simulate identity-level style discrepancy for robust multi-view feature augmentation.

where $D_{(c,s)}$ is the domain classifier parameterized by $\theta_{D_{(c,s)}}$. This is realized by applying a gradient reverse layer (GRL) [14] 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:

$$\mathcal{L}_{\text{spe}} = -\frac{1}{S}\sum_{s=1}^{S} k \log D_{(c,s)}\left(f^k_{(l,s)}\left(F^k_{\text{spe},s}\right)\right), \tag{4}$$

The final disentanglement objective is formulated as

$$\mathcal{L}_{\text{dise}} = \mathcal{L}_{\text{ws}} + \mathcal{L}_{\text{inv}} + \mathcal{L}_{\text{spe}}. \tag{5}$$

3.3. Cross-domain multi-view augmentation

The global encoder is specifically designed to extract a disentangled domain-agnostic prototype $F^k_{\text{inv},s}$, as defined by Eq. (1). However, considering the potential for significant style shifts in testing samples, due to occlusion, scale variations, and illumination changes, 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 [6] over the current minibatch as:

$$\begin{aligned}\mu_s^k &= \frac{1}{B_k HW}\sum_{b=1}^{B_k}\sum_{h=1}^{H}\sum_{w=1}^{W} F_s^k(b,h,w),\\ (\sigma_s^k)^2 &= \frac{1}{B_k HW}\sum_{b=1}^{B_k}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(F_s^k(b,h,w)-\mu_s^k\right)^2,\end{aligned} \tag{6}$$

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 in 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

$$\begin{aligned}\hat{\mu}_s^k &= \frac{1}{K-1}\sum_{i=1,i\neq k}^{K} \mu_s^k + \epsilon_\mu \delta\left(\mu_s^k\right),\\ \hat{\sigma}_s^k &= \frac{1}{K-1}\sum_{i=1,i\neq k}^{K} \sigma_s^k + \epsilon_\sigma \delta\left(\sigma_s^k\right).\end{aligned} \tag{7}$$

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

$$F^k_{\text{sty,s}} = \hat{\sigma}_s^k \frac{F^k_{\text{spe, s}} - \mu_s^k}{\sigma_s^k} + \hat{\mu}_s^k. \tag{8}$$

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

$$\hat{F}_s^k = \text{FC}\left(\text{cat}\left(F^k_{\text{inv},s}, F^k_{\text{sty,s}}\right)\right), \tag{9}$$

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.

3.4. Subspace hierarchical and consistent constrain

Feature maps extracted at each block with different kernels focusing on various aspects of the input image. 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 hypernetwork 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

$$\begin{aligned} \mathcal{L}_{\mathrm{v}} &= \frac{1}{V(V-1)} \sum_{m=1}^{V} \sum_{\substack{n=1 \\ n \neq m}}^{V} \left[m_1 - \| w_m - w_n \|_2 \right]_+^2, \\ \mathcal{L}_{\mathrm{c}} &= \frac{1}{B(B-1)} \sum_{i=1}^{B} \sum_{\substack{j=1 \\ j \neq i}}^{B} \left[\| \mathcal{W}_i - \mathcal{W}_j \|_2 - m_2 \right]_+^2, \end{aligned} \tag{10}$$

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

$$\mathcal{L}_{\mathrm{attr}} = \mathcal{L}_{\mathrm{c}} + \mathcal{L}_{\mathrm{v}}. \tag{11}$$

3.5. *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 $\mathcal{L}_{\mathrm{attr}}$.

$$\mathcal{L} = \mathcal{L}_{\mathrm{ce}} + \mathcal{L}_{\mathrm{tri}} + \mathcal{L}_{\mathrm{cent}} + \alpha \mathcal{L}_{\mathrm{dise}} + \beta \mathcal{L}_{\mathrm{attr}}, \tag{12}$$

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. The overall training procedure is depicted in Algorithm 1.

Algorithm 1 Learning of the proposed CDVM model.

```
Input: Source domains $D_k = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$; Maximum iterations $max\_iter$;
Number of training block $S$.
Output: Trained $f_\theta$.
for $i = 1$ to $max\_iter$ do
  Randomly sample a mini-batch $\{(x_i, y_i^{(p)})\}$.
  for $s = 1$ to $S$ do
    Intermediate feature $F_{(i,s)}^k$ Extraction.
    Feature disentanglement by Eq. (5).
    Cross-domain augmentation by Eq. (8).
    Attribute significance modeling by Eq. (11).
  Global pooling to get identity representation.
  Compute the meta losses (Eq. (12)).
  Update parameters by gradient descent.
end for
```

Table 1
Statistics of ReID datasets used in the paper.

Datasets	Abbr.		ID	Img
Market1501 [25]	M		1501	29,419
MSMT17 [26]	MS		4101	126,441
CUHK02 [27]	C2		1816	7264
CUHK03 [28]	C3		1467	14,097
CUHK-SYSU [35]	CS		11,934	34,574
Datasets	Probe		Gallery	
	ID	Img	ID	Img
PRID [30]	100	100	649	649
GRID [31]	125	125	900	900
VIPeR [32]	316	316	316	316
iLIDS [33]	60	60	60	60

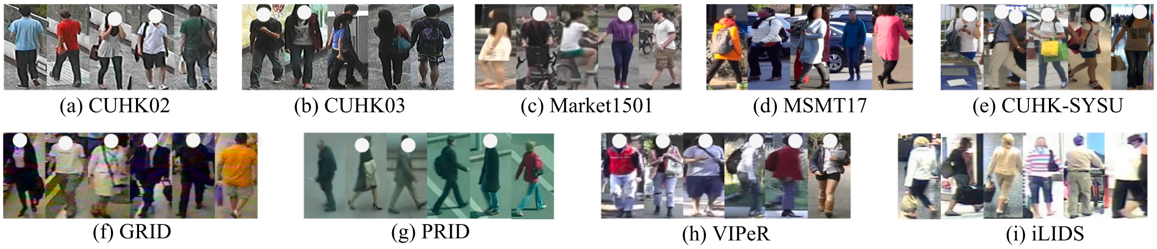
4. Experiments

Datasets and protocols. We conducted multisource domain generalization on a wide range of 9 benchmarks, including five large-scale datasets: Market1501 (M) [25], MSMT17 (MT) [26], CUHK02 (C2) [27], CUHK03 (C3) [28], CUHK-SYSU (CS) [29], and four small-scale datasets: PRID [30], GRID [31], VIPer [32], and iLIDs [33]. For CUHK03, we used the "labeled" subset to keep a fair comparison with the SOTA competitors [12,34]. The statistics of these datasets are shown in Table 1, and a few samples are visualized in Fig. 3, which exhibits significant domain shift caused by variations on illumination, viewpoint, resolution, and scene context.

Evaluation Metrics. We adopt two widely used evaluation metrics for person ReID:
(1) *Cumulative Matching Characteristic (CMC) Rank-1 Accuracy*, which measures the probability that the top-ranked gallery image corresponds to the correct identity. Let $\mathbb{1}(r_i = 1)$ be an indicator function that equals 1 if the ith query's correct match appears in the top-1 retrieval result; then the Rank-1 accuracy is:

$$\text{Rank-1} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(r_i = 1), \tag{13}$$

where N is the number of queries.
(2) *Mean Average Precision (mAP)*, which reflects both precision and recall across the ranking list. For each query q_i, the average precision (AP)



(a) CUHK02 (b) CUHK03 (c) Market1501 (d) MSMT17 (e) CUHK-SYSU
(f) GRID (g) PRID (h) VIPeR (i) iLIDS

Fig. 3. Example identity samples from different domains. Significant domain gaps are caused by the variation on nationality, illumination, viewpoints, resolution, scenario, etc.

Table 2
Comparisons with the SOTA methods 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 [36]	52.0	39.2	41.1	29.3	60.1	51.2	78.4	70.2	57.9	47.5
	SNR [6]	66.5	52.1	47.7	40.2	61.3	52.9	89.9	84.1	66.4	57.3
	RaMoE [37]	67.3	57.7	54.2	46.8	64.6	56.6	90.2	85.0	69.1	61.5
	DMG-Net [11]	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$_{50}$ [38]	62.2	52.3	57.4	48.6	66.3	57.0	81.9	75.0	67.0	58.2
	M^{3}L [34]	65.3	55.0	50.2	40.0	68.2	60.8	74.3	65.0	64.5	55.2
	MetaBIN [7]	70.8	61.2	57.9	50.2	64.3	55.9	82.7	74.7	68.9	60.5
	META [12]	71.7	61.9	60.1	52.4	68.4	61.5	83.5	79.2	70.9	63.8
	CDVM (Ours)	**74.1**	**64.8**	**66.1**	**56.0**	**69.6**	**63.6**	**87.7**	**83.1**	**74.4**	**66.9**

is:

$$\mathrm{AP}_i = \frac{1}{m_i}\sum_{k=1}^{n} P_i(k)\cdot \mathrm{rel}_i(k), \tag{14}$$

where $P_i(k)$ is the precision at the kth rank, $\mathrm{rel}_i(k)$ is a binary indicator of relevance at rank k, m_i is the total number of relevant images for q_i, and n is the list length.

The overall mAP is the average of APs across all queries:

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{AP}_i. \tag{15}$$

Implementation details. Following the conventional settings [39], we used ResNet50 [40] with IBN [41] 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 colorjitter. The proposed CDVM was trained for 120 epochs with an SGD optimizer [42], 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. (12) were both set to 0.5. The margins m_1 and m_2 in Eq. (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 framework with four A100 GPUs.

4.1. Comparisons to the state-of-the-art

Comparison under Protocol-1. One established evaluation protocol [6,36,37], is to train on five large-scale datasets, *i.e.,* DukeMTMC [43], 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 [12,39] 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 [6,36] on the random split query/gallery sets, and reported the averaged results in Table 2. Compared to the other SOTA models trained with the same datasets, our model shows clear advantages and outperforms the latest SOTA model META [12] 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.

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 3. It can be observed that the proposed CDVM model achieve superior performance when generalizing to CUHK03 and Market1501 and remains competitive when MSMT17 was 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. We note that the combination “M + CS + C3→MS” does not yield the best performance. This is mainly due to the significant domain gap: MSMT17 contains more diverse scenes and lighting conditions than the relatively

Table 3
Comparisons with the SOTA generalizable person ReID models on large-scale datasets (protocol-2 and protocol-3).

Setting	Method	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 [6]	8.9	8.9	6.8	19.9	34.6	62.7	16.8	30.5
	QAConv$_{50}$ [38]	25.4	24.8	16.4	45.3	63.1	83.7	35.0	51.3
	MetaBIN [7]	28.8	28.1	17.8	40.2	57.9	80.1	34.8	49.5
	M^{3}L [34]	34.2	34.4	16.7	37.5	61.5	82.3	37.5	51.4
	ACL [39]	41.2	41.8	20.4	45.9	74.3	89.3	45.3	59.0
	META [12]	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**
Protocol-3: Full Sets	SNR [6]	17.5	17.1	7.7	22.0	52.4	77.8	25.9	39.0
	QAConv$_{50}$ [38]	32.9	33.3	17.6	46.6	66.5	85.0	39.0	55.0
	MetaBIN [7]	43.0	43.1	18.8	41.2	67.2	84.5	43.0	56.3
	M^{3}L [34]	35.7	36.5	17.4	38.6	62.4	82.7	38.5	52.6
	ACL [39]	49.4	50.1	21.7	47.3	76.8	90.6	49.3	62.7
	META [12]	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**

Table 4
Components analysis. The proposed components were progressively incorporated into the baseline to study the individual contribution. The best results are in **bold**.

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
✗	✓	✗	39.1	39.3	18.9	45.1	73.1	88.5
✗	✗	✓	37.6	37.1	18.2	44.2	71.5	88.1
✓	✓	✓	**41.7**	**42.8**	**20.7**	**46.6**	**74.8**	**89.8**

homogeneous settings in Market1501, CUHK03, and CUHK-SYSU. As a result, the training domains offer limited style variation to effectively simulate the complexity of MSMT17.

4.2. Ablation study

We conducted comprehensive ablations studies to provide in-depth analyses and better understanding of each designed components of the CDVM model. All the variants were evaluated under protocol-2.

Components analysis. We investigated the individual contribution of different components in the CDVM model to study its effectiveness. As shown in Table 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 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 5
Effectiveness of the proposed discrepancy constraint $\mathcal{L}_{\text{ws}}$ 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

Table 6
Effects of different strategies in exploring cross-domain variations. "Determined" takes variations as fixed factors, "Elastic" considers combinations and randomness.

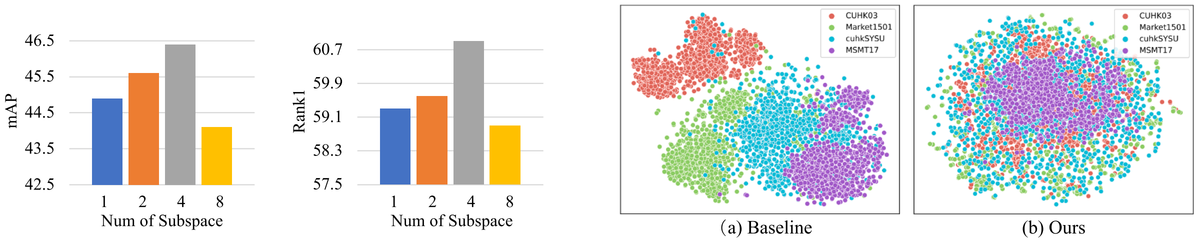
Method	CUHK03		MSMT17		Market1501	
	mAP	R1	mAP	R1	mAP	R1
Determined	37.3	38.1	19.0	45.1	72.0	88.3
Elastic	40.9	41.6	19.5	45.7	73.4	88.9

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, which is termed as "elastic" expansion. 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. 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.

Effects to Number of Subspaces. We studied the effects of the number of subspaces on the attribute constraints (Eq. (11)). Fig. 4a shows that grouping features into more subspaces provides a positive impact on model generalization. However, further increasing the number of subspaces can bring an optimization issue due to laking of explicit supervision signal. Based on the analysis, we set the number of subspaces to 4 as the default in our model design.

4.3. 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 Fig. 4b. From it, we observed that the baseline model is prone to learn domain-bias representations while the proposed CDVM model is more robust in extracting domain-invariant representation. Additionally, we visualized the attention maps generated by both the ResNet50 baseline and the proposed CDVM model. The results in Fig. 5 demonstrate that CDVM is more attentative to the target object.



(a) Baseline (b) Ours

(a) Effect of subspace number. (b) t-SNE visualization across domains.

Fig. 4. (a) Increasing subspace granularity improves generalization but may lead to optimization challenges. (b) t-SNE shows that our model better extracts domain-invariant features compared to the baseline.

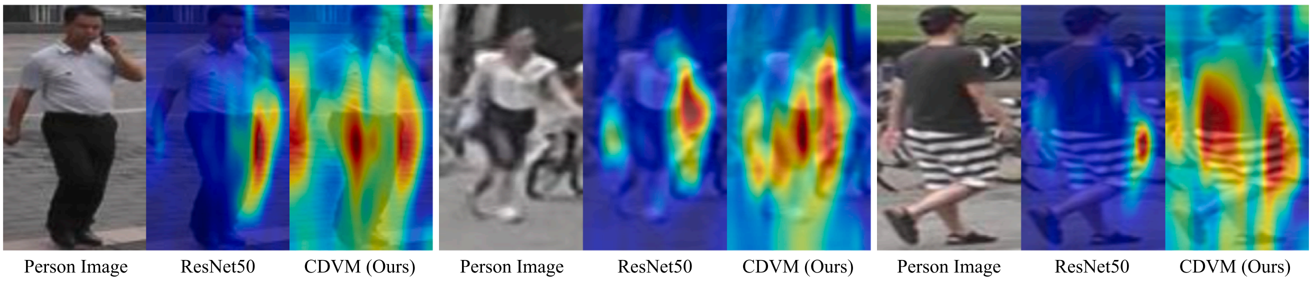


Fig. 5. Visualization of person image and attention map extracted by ResNet50 (baseline) and the proposed CDVM model.

5. Conclusions

In this work, we presented a novel *Cross-Domain Variations Mining* (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.

CRediT authorship contribution statement

Qilei Li: Methodology, Conceptualization; **Shitong Sun:** Conceptualization; **Weitong Cai:** Investigation; **Shaogang Gong:** Supervision, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The preliminary work for this study was conducted while Qilei Li was affiliated with Queen Mary University of London.

References

[1] G. Wu, X. Zhu, S. Gong, Learning hybrid ranking representation for person re-identification, Pattern Recognit. 121 (2022) 108239.
[2] X. Lan, X. Zhu, S. Gong, Unsupervised cross-domain person re-identification by instance and distribution alignment, Pattern Recognit. 124 (2022) 108514.
[3] C. Eom, B. Ham, Learning disentangled representation for robust person re-identification, NeurIPS (2019).
[4] Y.-F. Zhang, Z. Zhang, D. Li, Z. Jia, L. Wang, T. Tan, Learning domain invariant representations for generalizable person re-identification, IEEE Trans. Image Process. 32 (2022), pp. 509–523.
[5] Y. Wu, K. He, Group normalization, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
[6] X. Jin, C. Lan, W. Zeng, Z. Chen, L. Zhang, Style normalization and restitution for generalizable person re-identification, in: CVPR, 2020.
[7] S. Choi, T. Kim, M. Jeong, H. Park, C. Kim, Meta batch-instance normalization for generalizable person re-identification, in: CVPR, 2021.
[8] P. Li, D. Li, W. Li, S. Gong, Y. Fu, T.M. Hospedales, A simple feature augmentation for domain generalization, in: ICCV, 2021.
[9] R. Volpi, H. Namkoong, O. Sener, J.C. Duchi, V. Murino, S. Savarese, Generalizing to unseen domains via adversarial data augmentation, Adv. Neural Inf. Process. Syst. 31 (2018).
[10] C. Liu, S. Gong, C.C. Loy, On-the-fly feature importance mining for person re-identification, Pattern Recognit. 47 (4) (2014) 1602–1615.
[11] Y. Bai, J. Jiao, W. Ce, J. Liu, Y. Lou, X. Feng, L.-Y. Duan, Person30k: a dual-meta generalization network for person re-identification, in: CVPR, 2021.
[12] B. Xu, J. Liang, L. He, Z. Sun, Mimic embedding via adaptive aggregation: learning generalizable person re-identification, in: ECCV, 2022.
[13] W. Lin, J. Chu, L. Leng, J. Miao, L. Wang, Feature disentanglement in one-stage object detection, Pattern Recognit. 145 (2024) 109878.
[14] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, JMLR 17 (1) (2016) 2096–2030.
[15] M. Ilse, J.M. Tomczak, C. Louizos, M. Welling, DIVA: Domain invariant variational autoencoders, in: Medical Imaging with Deep Learning, PMLR, 2020, pp. 322–348.
[16] M.-H. Bui, T. Tran, A. Tran, D. Phung, Exploiting domain-specific features to enhance domain generalization, NeurIPS (2021).
[17] H. Bai, R. Sun, L. Hong, F. Zhou, N. Ye, H.-J. Ye, S.-H.G. Chan, Z. Li, DecAug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation, in: AAAI, 2021.
[18] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 1–48.
[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[20] X. Li, Y. Dai, Y. Ge, J. Liu, Y. Shan, L.-Y. Duan, Uncertainty modeling for out-of-distribution generalization, ICLR (2022).
[21] Huang, Siqi and Xu, Yanchen and Zhang, Hongyuan and Li, Xuelong, Learn beneficial noise as graph augmentation, in: Proceedings of International Conference on Machine Learning (ICML), 2025.
[22] H. Zhang, Y. Xu, S. Huang, X. Li, Data augmentation of contrastive learning is estimating positive-incentive noise, 2024. arXiv:2408.09929.
[23] S. Santurkar, D. Tsipras, A. Ilyas, A. Madry, How does batch normalization help optimization? NeruIPS (2018).
[24] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: ICCV, 2017.
[25] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: ICCV, 2015.
[26] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: CVPR, 2018.
[27] W. Li, X. Wang, Locally aligned feature transforms across views, in: CVPR, 2013.
[28] W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: deep filter pairing neural network for person re-identification, in: CVPR, 2014.
[29] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3415–3424.
[30] M. Hirzer, C. Beleznai, P.M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Scandinavian Conference on Image Analysis, 2011.
[31] C.C. Loy, T. Xiang, S. Gong, Time-delayed correlation analysis for multi-camera activity understanding, IJCV (2010).
[32] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: ECCV, 2008.
[33] W.-S. Zheng, S. Gong, T. Xiang, Associating Groups of People, in: BMVC, 2009.
[34] Y. Zhao, Z. Zhong, F. Yang, Z. Luo, Y. Lin, S. Li, N. Sebe, Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification, in: CVPR, 2021.
[35] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, End-to-end deep learning for person search, arXiv (2016).
[36] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, T.M. Hospedales, Generalizable person re-identification by domain-invariant mapping network, in: CVPR, 2019.
[37] Y. Dai, X. Li, J. Liu, Z. Tong, L.-Y. Duan, Generalizable person re-identification with relevance-aware mixture of experts, in: CVPR, 2021.

[38] S. Liao, L. Shao, Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting, in: ECCV, 2020.

[39] P. Zhang, H. Dou, Y. Yu, X. Li, Adaptive cross-domain learning for generalizable person re-identification, in: ECCV, 2022.

[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016.

[41] X. Pan, P. Luo, J. Shi, X. Tang, Two at once: enhancing learning and generalization capacities via IBN-Net, in: ECCV, 2018.

[42] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: ICLR, 2015.

[43] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in: ICCV, 2017.