Computer Vision and Image Understanding 240 (2024) 103905

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

IGMG: Instance-guided multi-granularity for domain generalizable person re-identification

Amran Bhuiyan [a,*], Jimmy Xiangji Huang [a], Aijun An [b]

[a] Information Retrieval and Knowledge Management Research Lab, York University, 4700 Keele Street Toronto, Ontario M3J 1P3, Canada
[b] Department of Electrical Engineering and Computer Science, York University, 4700 Keele Street Toronto, Ontario M3J 1P3, Canada

ARTICLE INFO

Communicated by William Robson Schwartz

Keywords:
Re-identification
Domain Generalization
Normalization
multi-granularity
Style-discrepancy

ABSTRACT

Learning generalized feature embedding is crucial for various computer vision tasks, including domain generalizable person re-identification (ReID). ReID aims to develop deep learning-based feature embeddings that can effectively recognize individuals in both trained (source) domains and unseen target domains. However, many state-of-the-art ReID methods suffer from overfitting as they train and test within the same source domain. To address this issue, we investigate the potential of multi-granularity approaches in mitigating domain shift challenges in person re-identification. Specifically, we propose a novel framework called instance-guided multi-granularity (IGMG), which leverages style-free features through non-parametric Instance Normalization (IN) at multiple granularity levels. While high-level abstract concepts are often not shared across different classes, low- and mid-level features can offer more shareable information to enhance the model's generalization capabilities. By incorporating this concept, our framework can dynamically eliminate style variations across various levels of abstraction. As a result, it enables the model to capture fine-grained details and high-level semantics, leading to enhanced robustness against changes in data distribution. To validate the effectiveness of our approach, we conduct extensive experiments on multiple benchmark ReID datasets. The results consistently demonstrate that our proposed framework exhibits strong generalization capabilities, performing consistently well on unseen target domains. The code is available at https://github.com/mdamranhossenbhuiyan/IGMG/.

1. Introduction

Person re-identification (ReID) is a task that involves identifying an individual across multiple camera views without any overlap. It has gained significant attention due to its wide range of applications in video surveillance. In this task, a person re-identification system aims to match a target person's given image (probe) captured by one camera with a gallery of images taken from different cameras. The task is highly challenging due to factors such as the flexible nature of the human body, variations in viewpoints, and varying illumination conditions. These factors can cause significant visual differences between images of the same person, while different individuals may exhibit similar appearances, further complicating the recognition process.

Supervised person re-identification (ReID) approaches have been extensively studied in the literature, with several notable methods proposed (Bhuiyan et al., 2014; Ahmed et al., 2015; Hermans et al., 2017; Ge et al., 2018; Sun et al., 2018; Luo et al., 2019; Bhuiyan et al., 2020; Hafner et al., 2022; Lin et al., 2023; Liu et al., 2023; Ji et al., 2023). These approaches involve training and testing models on different splits of the same dataset or domain. However, one common drawback is the difficulty in generalizing the learned model to unseen target datasets due to domain shifts. To illustrate this issue, Fig. 1(a)–(d) displays samples from different ReID datasets, highlighting the visible style discrepancies in terms of color saturation, contrasts, lighting, resolutions, clothing styles, seasons, and content variations. These differences in image data, captured under diverse conditions and by different cameras, contribute to the poor generalization capability of models trained on one dataset and tested on another. The t-SNE visualization in Fig. 2, which represents high-dimensional data in a lower-dimensional space, further demonstrates the distinct distributions and disparities between the Market1501 and DukeMTMC datasets. Existing approaches often tackle these challenges by overfitting the model to the training data, resulting in better recognition accuracy within the same dataset. However, when evaluated on previously unseen target datasets, these models suffer from a significant drop in performance, highlighting the limitations of their generalization capability. This problem is commonly referred to as style or content discrepancies in the literature, emphasizing the need to address the distribution differences in re-identification tasks.

The field of deep ReID has witnessed the emergence of various techniques to address its challenges. One particular approach is unsupervised domain adaptation (UDA) ReID (Wang et al., 2018c; Yang

[*] Corresponding author.
E-mail address: amran@yorku.ca (A. Bhuiyan).

https://doi.org/10.1016/j.cviu.2023.103905
Received 26 June 2023; Received in revised form 5 December 2023; Accepted 17 December 2023
Available online 29 December 2023

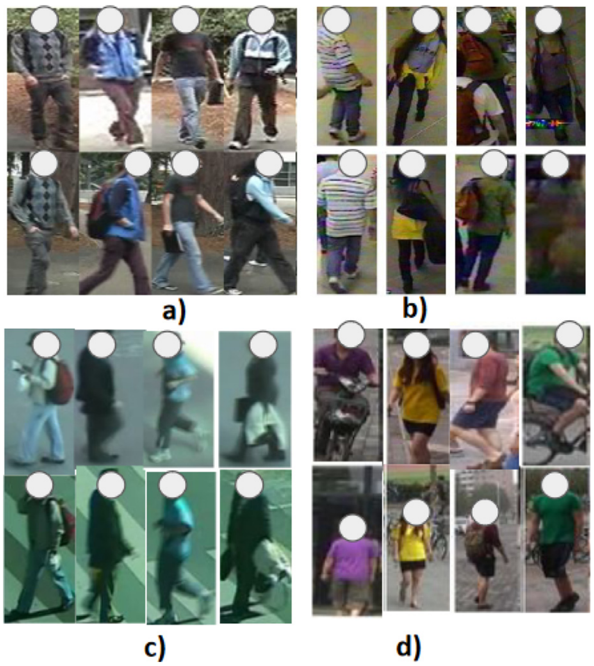

Fig. 1. Figure (a)–(d) showcases examples of domain shifts observed in re-identification across multiple datasets. Each column within each dataset represents bounding boxes of the same individual, while each row represents the same individual captured by different cameras. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

et al., 2019; Zhong et al., 2019; Mekhazni et al., 2020; Yu et al., 2019; Lin et al., 2020; Wang and Zhang, 2020; Li et al., 2023), which leverages unlabeled data from the target domain. However, UDA-ReID still relies on having access to target domain data, which can be costly or unattainable in some cases. Consequently, researchers have turned to Domain Generalization (DG) methods, which aim to train models that can generalize to unknown target domains using labeled data from one or more source domains instead of relying on actual target domain data. Therefore, in practical real-world applications, DG-ReID represents the most viable approach compared to other ReID techniques.

While Domain Generalization (DG) approaches have been successfully applied to various applications (Ulyanov et al., 2016; Muandet et al., 2013; Shankar et al., 2018), their direct application to ReID poses a challenge due to the mismatch in label spaces between source and target domains. Unlike other applications, ReID involves different individuals in different datasets, making it infeasible to assume a shared label space. Therefore, DG-ReID methods need to address significant domain discrepancies to achieve effective discrimination in unseen target domains with different label spaces. Several ReID approaches (Jia et al., 2019; Jin et al., 2020; Choi et al., 2021; Ni et al., 2022; Dai et al., 2021; Zhao et al., 2021; Luo et al., 2020) have been developed to tackle the domain generalization settings, categorized into meta-learning-based methods and invariant methods. Meta-learning approaches, such as those used in ReID methods like Choi et al. (2021), Ni et al. (2022), Dai et al. (2021) and Zhao et al. (2021), typically employ a split of the training set into meta-training and meta-test sets to simulate the generalization process for unknown domains. However, the optimization algorithms involved in meta-learning can be computationally expensive and time-consuming, particularly in large-scale ReID scenarios. On the contrary, invariants methods (Jia et al., 2019; Jin et al., 2020) directly extract domain-invariant features that usually rely on different normalization techniques to normalize the global feature representation by using statistics computed in mini-batches and regularize feature representations from heterogeneous domains, and thus the trained model is often capable of adapting to unseen domains. Relying on global feature representation, most of the DG-ReID techniques in this area integrated Instance Normalization (IN) to eliminate style differences between the domains. However, their IN module relies on the learnable affine parameters, which can potentially introduce biases toward specific domains during training

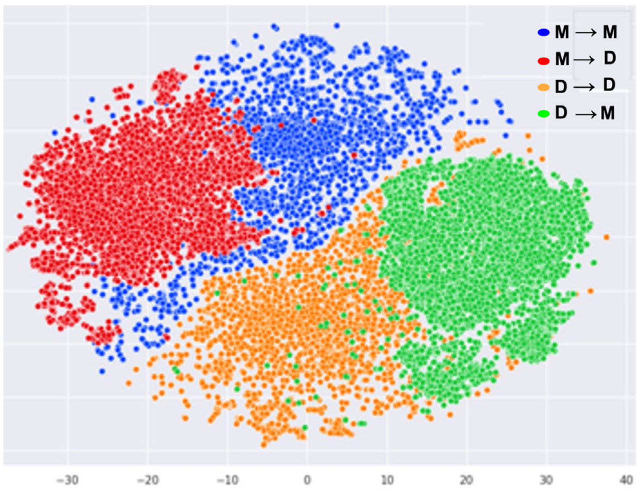

Fig. 2. The t-SNE visualization illustrates the domain shifts in re-identification between the Market1501 (M) and DukeMTMC (D) datasets. The feature points displayed here were extracted using the same experimental setup and baseline model (Bag-of-Trick (BOT)) (Luo et al., 2019) trained in a supervised manner using labels from the corresponding datasets. Red points represent DukeMTMC features extracted with a Market1501-trained model, while green points denote Market1501 features using a DukeMTMC-trained model.

This paper proposes a novel approach called Instance-guided Multiple Granular (IGMG) model for Domain Generalization Person Re-identification (DG-ReID). Our model draws inspiration from the Multiple Granular Network (MGN) introduced by Wang et al. (2018a) and aims to generate domain-invariant features by integrating global and local information at various levels of granularity. The IGMG model leverages coarse to fine-grained features, encompassing low-level characteristics like color and texture, mid-level attributes such as body parts and poses, and high-level descriptors like clothing style and accessories. By considering features at multiple abstraction levels, our deep learning model becomes capable of capturing intricate details and semantic information, thereby enhancing its robustness to changes in data distribution. As a result, the proposed model exhibits strong performance even when tested on new target datasets with different distributions. To implement this idea, we design an instance-guided multiple granular network, comprising one global branch and two local branches, each with separate parameters. The network is integrated into the ResNet50 backbone architecture, starting from the 4th residual stage. In the local branches of our IGMG model, we divide the feature maps into multiple stripes that represent distinct image parts. This division allows us to independently learn more focused saliency preferences for local features, drawing inspiration from the works of Wang et al. (2018a) and Sun et al. (2018).

In addition, we incorporate a non-learnable Instance Normalization (IN) module in our proposed IGMG network to serve as a gating mechanism for the different branches. We have observed that using a non-learnable IN module, which relies on fixed statistics independent of the domain, is more effective in capturing the inherent characteristics of the data and facilitating generalization to new domains. Choosing the appropriate convolutional layers in the backbone CNN for integrating the IN module becomes a crucial consideration to achieve the desired gating effect. We argue that integrating the IN module at the shallow layers of the backbone CNN is advantageous. Shallow layers in deep CNN architectures typically extract low- and mid-level features, which can be viewed as more abstract concepts such as parts or intricate texture patterns associated with a particular domain's style. On the other hand, deeper layers tend to capture high-level features that are more content-centric rather than style-based. By integrating the IN module in the shallower layers, we enhance the extraction of style-free back-propagated features through the gradients of the loss function, which correspond to the style-free information. This encourages the lower and middle layers to learn filters that extract more domain-invariant patterns. Consequently, the network becomes more adept at capturing discriminative information while being less influenced by domain-specific stylistic variations.

The main contribution of this paper is threefold:

- We present a unique multi-granular network designed for person re-identification. Our approach distinctively integrates high-level semantics with fine-grained nuances. This results in a comprehensive and adaptable feature representation, offering robustness against data variations across domains.
- We pioneer a technique for deriving style-independent features. Central to our method is the innovative use of a non-parametric Instance Normalization (IN) module. As far as we are aware, this is the inaugural application of non-parametric IN as a control mechanism for branching within Domain Generalizable Person Re-identification.
- The inclusion of the non-parametric IN module serves a two-fold purpose: it proficiently distinguishes style from content and adeptly addresses the longstanding challenge of style disparity, a noted concern in previous research.

The structure of the remaining sections in this paper is as follows: In Section 2, we present background information on deep learning models for person ReID and Domain Generalization Person ReID. The details of our proposed Instance-guided Multiple Granular (IGMG) approach are provided in Section 3. Section 4 comprehensively describes the experimental methodology, including benchmark datasets, protocol, performance measures, and presents the comparative results. Finally, in Section 5, we draw a comprehensive conclusion and discuss future directions.

2. Related work

This section offers a concise summary of the current state-of-the-art deep ReID approaches, as well as deep domain generalized person ReID methods..

2.1. Deep person ReID.

Deep person ReID has experienced significant advancements in recent years, branching out from the initial Siamese network framework. One prominent approach is supervised ReID (Zheng et al., 2019; Chen et al., 2019a; Luo et al., 2019; Chen et al., 2019b; Zhang et al., 2020a; Bhuiyan et al., 2020; Sun et al., 2018; Masson et al., 2021; Bhuiyan and Huang, 2022; Hafner et al., 2022; Lin et al., 2023; Liu et al., 2023; Ji et al., 2023), which has demonstrated high recognition accuracy in addressing challenges related to misalignment within a dataset. These methods often employ techniques such as part-based approaches (Sun et al., 2018; Qian et al., 2018; Chen et al., 2019a; Bhuiyan et al., 2020) and attention-aware networks (Chen et al., 2019b,a; Luo et al., 2019; Zhang et al., 2020a; Lin et al., 2023; Liu et al., 2023) to improve performance. Supervised ReID models perform well on previously encountered datasets but struggle with unseen domains due to domain shifts and style differences, making them unsuitable for real-world applications where obtaining labeled data is challenging. To handle this issue, researchers put their efforts into designing unsupervised domain adaptation (UDA) based ReID approaches (Wang et al., 2018c; Yang et al., 2019; Mekhazni et al., 2020; Zhong et al., 2019; Yu et al., 2019; Wang and Zhang, 2020; Ge et al., 2020; Feng et al., 2021; Liu et al., 2022; Li et al., 2023; Chen et al., 2023) where there is instant target domain data without annotations. It requires the collection of target domain data and updating the domain accordingly. UDA ReID approaches can roughly be categorized into four kinds: clustering the target data to generate pseudo labels (Fu et al., 2019; Zhai et al., 2020; Ge et al., 2020; Xuan and Zhang, 2021; Feng et al., 2021; Liu et al., 2022; Chen et al., 2023), self-supervised training on the target domain (Yang et al., 2019; Zou et al., 2020), generating images similar to the target domain style and source domain label space for data augmentations (Wang et al., 2018c; Zhong et al., 2018, 2019; Zhang et al., 2020b) and finally aligning feature distribution between source and target domains (Wu et al., 2019; Mekhazni et al., 2020; Li et al., 2021). However, existing Unsupervised Domain Adaptation (UDA) ReID methods still necessitate accessing and updating with target data, a process that can be time-intensive for real-world usage. Recent studies, like those by Delussu et al. (2021, 2023), have explored the human-in-the-loop (HITL) approach to address domain shifts in ReID. This strategy, contrasting traditional UDA and Domain Generalization (DG) methods, leverages real-time human feedback on the target data during the ReID system's operation. While in situations where no target data is available for model training, it is worth noting that, unlike DG methods, HITL strategies utilize target data obtained during system use.

2.2. Domain generalization person re-identification.

The inception of domain generalization approaches stems from the feasible practical assumption that there is no target data (i.e., no matter whether labeled or unlabeled) is available while deploying the ReID model. Compared to domain generalization for other vision applications (Ulyanov et al., 2016; Muandet et al., 2013), DG-ReID has to deal with the additional challenges of having disjoint label spaces where the target domain typically has different identities from the source domain. Recently, there have been a number of DG-ReID approaches (Jia et al., 2019; Jin et al., 2020; Ni et al., 2022; Choi et al., 2021; Dai et al., 2021; Luo et al., 2020; Zhang et al., 2023; Wu et al., 2023) that uses different pipelines to enhance the generalization capability of the learned model. SuA-SpML (Zhang et al., 2023), Meta-GA (Wu et al., 2023) MetaBIN (Choi et al., 2021), RaMoE (Dai et al., 2021) and MDA (Ni et al., 2022) utilize the meta-learning pipelines where the training set is divided into meta-training sets and meta-test sets to simulate the generalization process of an unknown domain. As indicated in the meta-training pipeline, the meta-training process of meta-learning is complex and slow which could be considered as backlash for deploying it in the DG-ReID setting. On the other hand, Cross-Domain Mixup (Luo et al., 2020) and DomainMix. Wang et al. (2020) relied on data augmentation techniques by mixing data from different domains without considering the abrupt transition among the domains. Building on the same approach, DCCL (Gong et al., 2023) structures its training using a Curriculum Learning method. This method mirrors how humans learn throughout their lives, starting from simpler tasks and gradually moving to more complex ones to adapt and learn in unfamiliar domains. Nevertheless, the data augmentation in such a manner is just another way of expanding the existing category of data, and the model training using those data might lead to overfitting to the known domains. Finally, there are some DG-ReID approaches (Jia et al., 2019; Jin et al., 2020; Jiao et al., 2022) rely on developing invariant techniques that can optimize and extract domain invariant features using different deep neural network architectures. DualNorm (Jia et al., 2019) utilized the IN module in each bottleneck in the shallow layers of the deep network and obtained robust domain-invariant features over multiple unknown domains. Following the same pipeline, SNR (Jin et al., 2020) proposed style normalization and restitution that integrate IN as in Jia et al. (2019) followed by identity-relevant feature restitution for better disentanglement. Adopting a similar framework, DTIN-Net (Jiao et al., 2022) leverages unnormalized features to refine normalized ones for better domain and instance adaptation, while also using multi-task learning in multi-source scenarios to enhance

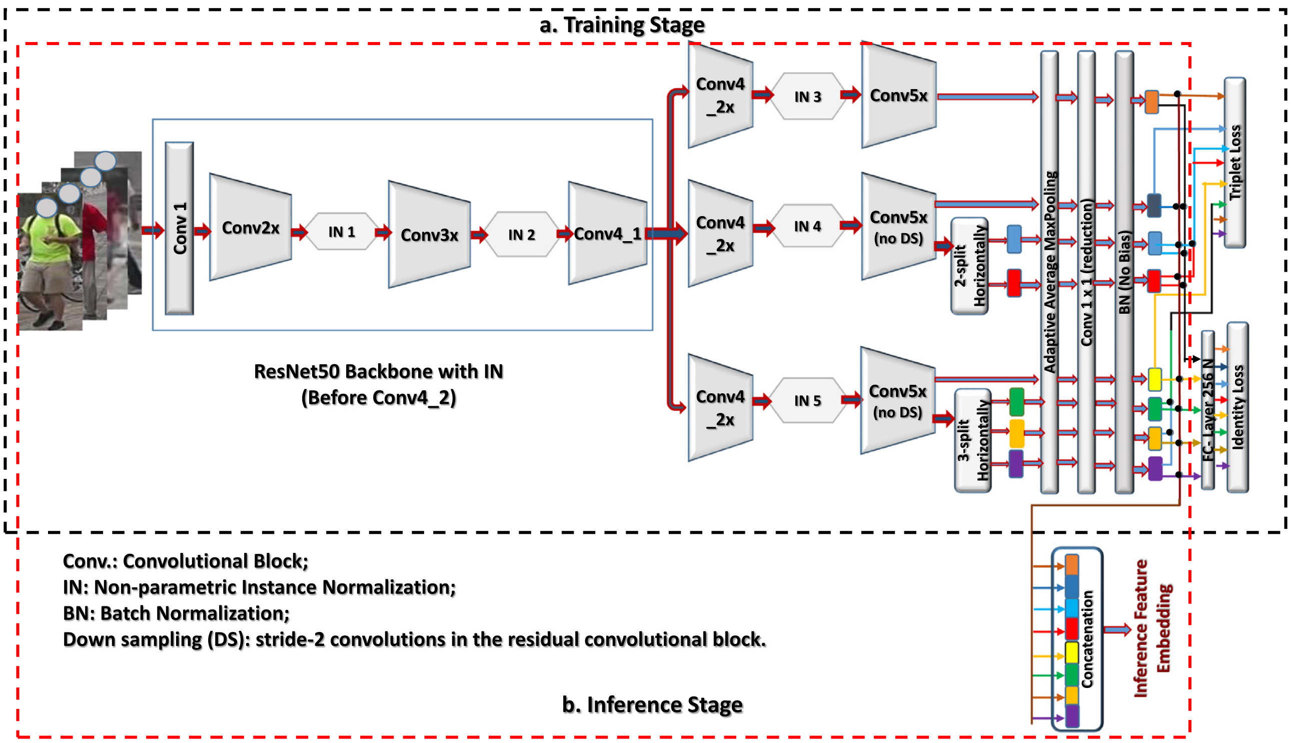

Fig. 3. Illustrations of the proposed Instance Guided Multi-granularity (IGMG) framework for DG-ReID. (a) Training architecture. Optimal feature multi-granularity is achieved by dividing the ResNet-50 backbone into global and local branches. Additionally, non-parametric Instance Normalization (IN) introduces the gating effect in each convolutional layer (except the final one) to filter out the style features related to dataset bias. Each route from *Conv4_2x* to its respective loss function denotes a unique supervisory signal, with no weight sharing across branches. (b) During testing, all the reduced features are concatenated together as the final feature representation of a pedestrian image.

domain generalization capabilities further. However, all these invariant approaches (Jia et al., 2019; Jin et al., 2020; Jiao et al., 2022) utilized the learnable parameters of IN, which hinders its normalization process for removing the style from each sample. Additionally, most of these DG-ReID approaches (Jia et al., 2019; Jin et al., 2020; Ni et al., 2022; Choi et al., 2021; Dai et al., 2021; Luo et al., 2020) relied on global feature representation ignoring the benefit of using both global and local representations to deal with the domain shift with a different distribution.

Unlike these approaches (Jia et al., 2019; Jin et al., 2020; Ni et al., 2022; Choi et al., 2021; Dai et al., 2021; Luo et al., 2020), in our IGMG framework, we exploit the advantages of using multi-granular feature representation that combines global and local information in different abstractions. It can help address domain shift in person re-identification by allowing the model to learn features at different levels of abstraction, making it more robust to changes in the data distribution. Different from other DG-ReID approaches, we integrate the IN module that does not consider learnable affine parameters while training and initiates the gating effect on different branches of IGMG architectures to filter out the style information for enhancing the generalization capability of the learned feature.

3. Our proposed framework IGMG

Our objective in this study is to develop a system that learns features that are consistent across different domains and perform well on new, unseen data. The overall structure of our proposed model, referred to as IGMG, is illustrated in Fig. 3. The IGMG model consists of backbone CNNs that incorporate multiple granular representations and IN modules. We train the IGMG model end-to-end in a single step. When presented with an input image, the features extracted from the convolutional layers are passed through the IN module. This module employs a gating mechanism to remove style-related statistics, and the resulting features are then segmented into a few distinct granular components. This division helps reduce the impact of domain-specific variations to some extent.

3.1. Multi-granular network

In ReID, deep neural networks naturally differentiate body parts based on their inherent meanings, but they often overlook specific patterns in body parts, focusing mainly on the central body. Interestingly, when networks are trained on local features, their responses cluster around notable semantic patterns, with this clustering affected by the size of the regions considered. This behavior suggests that the granularity of regions impacts the network's focus on particular patterns. By leveraging this insight, we introduce the IGMG framework that employs the Multiple Granularity Network (MGN) architecture (Wang et al., 2018a), integrating both broad global and detailed multi-granularity local features. to enhance recognition, emphasizing better domain generalization.

To ensure the fulfillment of our overall objectives, selecting an appropriate deep CNN architecture is a crucial step in our proposed IGMG network. While there exist various deep CNN architectures like VGGs (Simonyan and Zisserman, 2014), Inception (Szegedy et al., 2016), and OsNet (Zhou et al., 2019), we have opted for ResNet (He et al., 2016) architecture. The ResNet backbone is comprised of multiple residual convolutional blocks, namely: *Conv1*, followed by *Conv2x*, *Conv3x*, *Conv4x*, and *Conv5x*. As illustrated in Fig. 3, the ResNet-50 architecture has been adapted to better fit the requirements of this work. Instead of being utilized in its entirety, specific blocks from ResNet-50 have been selected. More specifically, following the MGN architecture (Wang et al., 2018a), we have made specific modifications to the network structure beyond the *Conv4_1* layer. These modifications involve the partitioning of this portion of the network into three separate branches. These changes were made to ensure that the network's architecture aligns with the specific objectives and design principles of our proposed IGMG framework. It is important to highlight that our design emerged as the most optimal after thorough evaluations of different branch counts and stripe setups.

In Fig. 3, the upper branch is known as the Global branch (Part-1), and its primary focus is on learning global feature representations without considering partition information. To achieve this, the feature representation from the convolutional blocks is reduced in dimensionality, going from 2048 dimensions to 512 dimensions. This dimensionality reduction is achieved through a series of operations, including average max-pooling and a 1 × 1 convolution layer with batch normalization. The resulting feature map serves as input for both the triplet loss and the fully connected (FC) layer, specifically for the Global branch during training.

The architecture of both the middle and lower branches in the network mirrors that of the Global Branch, with one exception: the absence of down-sampling operations in the *Conv5x* block. In each branch, the output feature maps are evenly divided into multiple horizontal stripes, which serve as the foundation for learning local feature representations. These stripes undergo the same subsequent operations as the Global Branch. These branches are referred to as Part-g Branches, where 'g' represents the number of partitions on the unreduced feature maps. For example, in Fig. 3, the middle branch can be denoted as the Part-2 Branch, while the lower branch can be designated as the Part-3 Branch.

During the testing phase, the reduced features are concatenated to obtain the final feature. This merging of global and local information aims to enhance the learned features' robustness and improve their ability to discriminate between different instances.

3.2. Instance normalization (IN)

Since IN (Ulyanov et al., 2016) can effectively remove the style content, it introduces the gating operation to filter out the style information propagated throughout the backbone architecture. Given a mini-batch of feature $\mathbf{F}^l_i \in \mathbb{R}^{b\times c_l\times h'\times w'}$ from the lth layer of the backbone architecture, the IN can perform instance normalization as follows:

$$\tilde{\mathbf{F}}^l_i = IN(\mathbf{F}^\mathbf{l}_\mathbf{i}) = \gamma(\frac{\mathbf{F}^\mathbf{l}_\mathbf{i} - \mu(\mathbf{F}^\mathbf{l}_\mathbf{i})}{\sigma(\mathbf{F}^\mathbf{l}_\mathbf{i})}) + \beta, \tag{1}$$

where $\mu(.)$ and $\sigma(.)$ denote the mean and standard deviation computed across spatial dimensions independently for each channel and each sample/instance, and $\gamma, \beta \in \mathbb{R}^c$ are affine parameters learned from data. In our IGMG framework, we propose that there is no necessity for the IN module to have learnable parameters during training. Therefore, we set the values of γ and β to 1 and 0, respectively. This design choice enhances the flexibility of the model, enabling it to adapt to varying input statistics. Consequently, the model exhibits improved generalization performance when applied to new domains or datasets. So the final equation for IN in our IGMG framework is:

$$\tilde{\mathbf{F}}^l_i = IN(\mathbf{F}^\mathbf{l}_\mathbf{i}) = \frac{\mathbf{F}^\mathbf{l}_\mathbf{i} - \mu(\mathbf{F}^\mathbf{l}_\mathbf{i})}{\sigma(\mathbf{F}^\mathbf{l}_\mathbf{i})} \tag{2}$$

Unlike Jin et al. (2020) and Jia et al. (2019), we integrate the non-parametric IN in between all the convolutional blocks of ResNet-50 (He et al., 2016), as shown in Fig. 3.

3.3. Network loss

To train the IGMG network, we employ the training techniques utilized in the Bag-of-Trick (BOT) approach (Luo et al., 2019). This approach combines triplet losses and identity (ID) losses, incorporating label smoothing at multiple levels of multi-granularity. For a given mini-batch of images, the total loss can be expressed as:

$$\mathcal{L} = \sum_{g=1}^{8} \alpha\mathcal{L}^g_{triplet} + \lambda\mathcal{L}_{ID}{}^g, \tag{3}$$

where $\mathcal{L}^g_{triplet}$ and $\mathcal{L}_{ID}{}^g$ denote the triplet loss and ID loss of the gth granular output, respectively. α and λ are the weight of different granular levels for the triplet and identity losses, respectively.

Table 1
Different evaluation protocols for large-scale multi-source DG-ReID.

Setting	Training Data	Testing Data
Protocol-2	M+D+C3+MS	PRID, GRID, VIPeR
Protocol-3	Leave-one-out for M+D+C+MT	

Following the state-of-the-art ReID approaches (Luo et al., 2019; Bhuiyan and Huang, 2022; Jin et al., 2020; Jia et al., 2019), we utilize the widely used hard positive and negative mining technique for a mini-batch (Hermans et al., 2017). For each sample in a mini-batch, it selects the hardest positive and the hardest negative samples within the batch when forming the triplets for computing the loss:

$$\mathcal{L}^g_{triplet} = \frac{1}{N_s}\sum_{a=1}^{N_s}\Big[m + \max_{y_p=y_a} d\left(\tilde{\mathbf{f}}^g_a, \tilde{\mathbf{f}}^g_p\right) - \min_{y_p\neq y_a} d\left(\tilde{\mathbf{f}}^g_a, \tilde{\mathbf{f}}^g_n\right)\Big]_+ \tag{4}$$

where $[.]_+ = \max(.,0)$, m denotes a margin, N_s is the set of all hard triplets and d is the Cosine distance. $\tilde{\mathbf{f}}^g_a$, $\tilde{\mathbf{f}}^g_p$ and $\tilde{\mathbf{f}}^g_n$ denote the gth granular features of anchor, positive and negative samples with their labels y_a, y_p and y_n, respectively. Utilizing label smoothing (LS) (Wang et al., 2018b) into the ID loss has proven effective in preventing the ReID model from overfitting the training IDs. From this observation, we integrate LS with the ID loss. Given an image of an individual with true ID label y and p^g_i as the gth granular ID prediction logits of class i, the ID loss can be written as:

$$\mathcal{L}^g_{ID} = \sum_{i=1}^{N} -q^g_i \log(p^g_i), \tag{5}$$

where LS construction of q^g_i is:

$$q^g_i = \begin{cases} 1 - \frac{N-1}{N}\epsilon, & \text{if } i = y \\ \frac{\epsilon}{N}, & \text{otherwise,} \end{cases} \tag{6}$$

where ϵ is a constant to force the model to be less confident on the training set. In this work, ϵ is set to be 0.1. For small datasets with fewer IDs, LS has proven to be effective in improving the performance of the model.

4. Empirical evaluation

In this section, we provide information about the datasets used, implementation details, and performance metrics for validating our proposed IGMG framework. Subsequently, we conduct a comprehensive comparison with state-of-the-art methods, considering both single-source and multi-source scenarios, including a large-scale dataset for DG-ReID. Furthermore, we present an ablation study to evaluate the impact of various components incorporated in the proposed IGMG approach. Additionally, we perform qualitative result analysis by visualizing rank lists.

4.1. Experimental settings

4.1.1. Implementation details.

In our IGMG framework, we have the flexibility to incorporate various feature extractors. However, for this study, we choose the ResNet50 architecture (He et al., 2016) pretrained on Imagenet as our backbone due to its widespread usage in re-identification tasks. In the proposed architecture, the pretrained weights after the *Conv4_1* block of the original ResNet50 are shared among the three branches that follow the same block. To maintain consistency, all input images are resized to a resolution of 384 × 128. Following the best practice in ReID, our training process utilizes the PK triplet sampling technique. This involves randomly selecting P identities and then, for each one, picking K random instances, giving us a batch size of P*K. To put it simply, our batches are made up of 64 images: 4 unique identities with

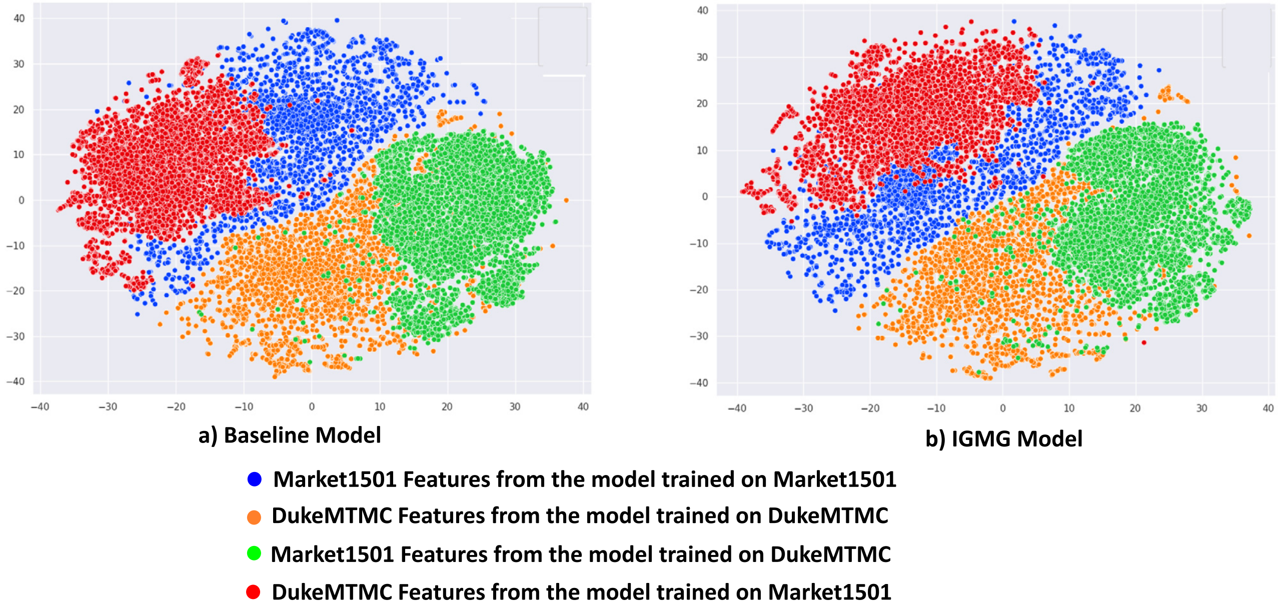

Fig. 4. Visualization of the features via t-SNE in the single-source setting.

16 images each. For multi-source experiments, different identities and cameras from various datasets are mixed together within a batch. Data augmentation techniques such as random cropping, flipping, and color jittering are applied. However, we choose not to utilize random erasing (REA) as it has shown negative effects in cross-domain ReID tasks, as mentioned in previous works (Jin et al., 2020; Dai et al., 2021). For optimization, we employ the Adam optimizer with a weight decay of 0.0005. The model is trained for a total of 60 epochs, with a warmup strategy implemented in the first 2000 iterations. The learning rate is initialized to 3.5×10^{-4} and is reduced by a factor of 0.1 after the 40th epoch. All models are implemented using PyTorch and trained on a single NVIDIA V100 GPU with 12 GB HBM2 memory. The duration of training fluctuates based on various experimental settings and the specific datasets used, particularly influenced by the number of unique identities present in each dataset. As an example, in the single-source experiment of Market1501 → DukeMTMC, our proposed IGMG model required approximately 2 h of training time. This was based on the presence of 750 distinct identities within the Market1501 dataset used for training purposes.

4.1.2. Datasets.

Following the state-of-the-art DG-ReID approaches (Jia et al., 2019; Jin et al., 2020; Dai et al., 2021; Ni et al., 2022), we conduct our experiments on the public ReID datasets, including Market1501 (Zheng et al., 2015), DukeMTMC-reID (Zheng et al., 2017), CUHK-03 (Li et al., 2014), MSMT17 (Wei et al., 2018), and three small datasets including datasets PRID2011 (Hirzer et al., 2011), GRID (Loy et al., 2010), and VIPeR (Gray and Tao, 2008). For simplicity, we denote Market1501 as M, DukeMTMC-reID as D, MSMT17 as MS, and CUHK03 as C3.

4.1.3. Evaluation settings

For single-source DG ReID, we choose either Market1501 (Zheng et al., 2015), and DukeMTMC-reID (Zheng et al., 2017) as the training set and directly test on the other dataset.

For multi-source DG-ReID, we adopted the protocols implemented by the state-of-the-art DG-ReID (Dai et al., 2021; Jin et al., 2020; Jia et al., 2019), as shown in Table 1. Under Protocol-2 (Dai et al., 2021), we utilize all images from M+D+C3+MS, which includes both training and testing sets. Our tests target four smaller ReID datasets: PRID, GRID, VIPeR, and iLIDs. The final results for these datasets are derived from the average performances across 10 random gallery and probe set splits. For Protocol-3 (Dai et al., 2021), we employ a "leave-one-out" setting with M+D+C3+MS. This means we pick one domain from M+D+C3+MS for testing (using only its test set) and train using all other domains, including both their training and testing sets.

Following the common trend of evaluation (Jin et al., 2020; Jia et al., 2019; Luo et al., 2019; Dai et al., 2021), we use two main metrics to evaluate our proposed IGMG framework and baseline methods. First, we look at the rank-01 accuracy, which we refer to as R-1 in our tables. In the context of ReID, rank-01 accuracy refers to the percentage of queries where the correct match is the top result in the ranking list. Second, we measure the mean average precision, or mAP which measures the average precision of retrieval results across all queries, providing a comprehensive evaluation of both the precision and recall of the model.

4.2. Comparison with state-of-the-art methods

4.2.1. Single-source DG-ReID.

The purpose of this experiment is to evaluate our proposed IGMG framework using a single-source dataset. We trained our model using the Market (and alternatively, Duke) dataset, then tested on the Duke (or Market) as the target. Additionally, after training on the MSMT17 dataset, we tested our model on both the Market and Duke datasets. To demonstrate the effectiveness of our IGMG approach, we compared it with two recent state-of-the-art approaches: DG-ReID and multigranularity-based MGN approach.

Table 2 presents a comparison of the performance of our IGMG approach with the state-of-the-art approaches. Our IGMG framework consistently outperformed the best-performing state-of-the-art DG-ReID methods, MDA (Ni et al., 2022) and DTIN-Net (Jiao et al., 2022), in all evaluation measures. For example, our IGMG approach achieved a 7.2% improvement in rank-01 accuracy and a 5.5% improvement in mAP over DTIN-Net for the $M \rightarrow D$ setting, and a 6% improvement in rank-01 accuracy and a 4.7% improvement in mAP for the $D \rightarrow M$ setting. It is important to note that some state-of-the-art approaches, such as SNR (Jin et al., 2020) and MDA (Ni et al., 2022), utilized both the train and test splits of a dataset as the source domain. In our experiment, we only used the train split as the source domain to ensure a fair comparison with other DG-ReID approaches. Despite this limitation, our IGMG framework still significantly outperformed SNR and MDA approaches. To highlight the importance of multigranularity,

Table 2
Comparisons between ours and the state-of-the-art in a **single-source DG-ReID setting**.

Methods	Reference	M → D		D → M		MS → M		MS → D	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
ECN baseline (Zhong et al., 2019)	CVPR	28.9	14.8	43.1	17.7	43.2	20.7	47.4	27.4
PN-GAN (Qian et al., 2018)	ECCV	29.9	15.8	–	–	–	–	–	–
QA$Conv_{50}$ (Liao and Shao, 2020)	ECCV	48.8	28.7	58.6	27.2	72.6	43.1	69.4	52.6
DSU (Li et al., 2022)	ICLR	52.0	32.0	63.7	32.4	–	–	–	–
SNR (Jin et al., 2020)	CVPR	55.1	33.6	66.7	33.9	70.1	41.4	–	–
MetaBIN (Choi et al., 2021)	CVPR	55.2	33.1	69.2	35.9	–	–	–	–
SuA-SpML (Zhang et al., 2023)	TIP	55.5	34.8	65.8	36.3	–	–	–	–
Meta-GA (Wu et al., 2023)	MTA	55.1	34.9	69.3	36.6	–	–	–	–
DualNorm (Jia et al., 2019)	BMVC	56.2	35.2	72.2	39.9	78.7	51.5	75.1	59.6
MDA (Ni et al., 2022)	CVPR	56.7	34.4	70.3	38.0	–	–	–	–
DTIN-Net (Jiao et al., 2022)	ECCV	57.0	36.1	69.8	37.4	–	–	–	–
MGN (Wang et al., 2018a)	ACM MM	61.9	41.1	71.1	38.6	78.6	49.8	75.7	58.1
IGMG	**Ours**	**64.2**	**41.6**	**75.8**	**42.1**	**80.9**	**53.5**	**77.7**	**61.6**

Table 3
Comparisons between ours and the state-of-the-art in multi-source DG ReID on **Protocol-2 setting**. The best results are highlighted in bold.

Methods	Reference	Target: PRID		Target: GRID		Target: VIPeR	
		R-1	mAP	R-1	mAP	R-1	mAP
SNR (Jin et al., 2020)	ECCV	49.0	60.0	30.4	41.3	55.1	65.0
DMG-Net (Bai et al., 2021)	CVPR	59.9	69.7	37.3	47.2	62.3	70.9
RaMoE (Dai et al., 2021)	CVPR	56.9	66.8	43.4	53.9	63.4	72.2
DualNorm (Jia et al., 2019)	BMVC	58.6	69.8	40.9	49.8	61.6	70.6
DTIN-Net (Jiao et al., 2022)	ECCV	**67.4**	**77.4**	**49.4**	**58.4**	64.0	71.9
IGMG	**Ours**	63.8	73.0	45.2	53.4	**65.2**	**72.9**

Table 4
Comparisons between ours and the state-of-the-art in multi-source DG ReID on **Protocol-3 setting**. The best results are highlighted in bold.

Methods	Reference	D+C+MS→ M		M+MS+C → D		M+D+MS → C		M+D+C → MS	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
QA$Conv_{50}$ (Liao and Shao, 2020)	ECCV	65.7	35.6	66.1	35.6	23.5	21.3	24.3	7.5
M^3L (Zhao et al., 2021)	CVPR	75.9	50.2	69.2	51.1	33.1	32.1	33.0	12.9
SNR (Jin et al., 2020)	ECCV	76.5	49.3	66.2	47.4	28.5	28.6	35.7	14.1
DEX (Ang et al., 2021)	BMVC	81.5	55.5	73.7	55.0	36.7	33.8	43.5	18.7
DSU (Li et al., 2022)	ICLR	77.8	52.6	71.7	53.9	34.6	32.5	46.9	19.0
QAConv-GS (Liao and Shao, 2022)	CVPR	79.1	53.8	72.4	54.5	35.9	34.2	46.5	17.1
MixNorm (Qi et al., 2022)	TMM	78.9	51.4	70.8	49.9	29.6	29.0	**47.2**	19.4
RaMoE (Dai et al., 2021)	CVPR	82.0	56.5	73.6	56.9	34.6	33.5	34.1	13.5
DualNorm (Jia et al., 2019)	BMVC	82.3	56.7	74.5	56.4	36.5	35.0	45.4	19.6
SuA-SpML (Zhang et al., 2023)	TIP	81.0	56.6	70.8	51.2	–	–	–	–
DCCL (Gong et al., 2023)	TCSVT	82.7	**60.4**	72.5	55.0	36.9	36.0	45.3	18.2
IGMG	**Ours**	**83.1**	57.1	**76.2**	**57.7**	**37.3**	**36.3**	41.6	**19.7**

we assessed the MGN approach (Wang et al., 2018c) In our evaluation, it ranked as the second-best model among other DG-ReID models. This highlights the effectiveness of introducing feature multi-granularity and non-parametric instance normalization, which improves the generalization capability of the model across unseen domains, regardless of the size of the source domain in the single-source setting.

Visualization of the feature distribution in Single-Source Setting. Fig. 4 depicts the visualization of feature distributions for the source and target domains in a single-source setting. Upon examining the figure, we can observe that the baseline method demonstrates fewer overlapping regions between the features of different domains. In contrast, our approach yields larger overlapping regions, indicating that our method effectively learns domain-invariant features. This capability enables our model to capture generalizable patterns that can be applied to unseen domains. Hence, the results of our experiment validate the effectiveness of our approach in learning domain-invariant features and achieving better generalization across domains

4.2.2. Large-scale DG-ReID (Protocol-2)

This experiment aims to analyze the effect of our proposed IGMG approach on large-scale multi-source ReID benchmarks in the protocol-0 setting. Following protocols designed by the state-of-the-art DG-ReID approaches (Ni et al., 2022; Dai et al., 2021; Jin et al., 2020) and for drawing comparisons, all the small-scale target datasets in protocol-2 tested in a single-shot setting using specific numbers of query and gallery images. Specifically, for the VIPeR dataset, we use 316 query images and 316 gallery images. We use 125 query images and 900 gallery images for the GRID dataset. Following (Ni et al., 2022; Dai et al., 2021), all results are the average of 10 random splits on the target datasets for protocol-2.

Table 3 reports the comparative performance of our approach with the available state-of-the-art DG-ReID approaches on protocol-2 settings. As shown in Table 3, the IGMG framework consistently works well on all the considered datasets by outperforming the considered baseline DualNorm. In particular, the margin of improvements for the PRID dataset is 6.9% in rank-01 and 6.2% over the RaMoE (Dai et al., 2021). The proposed IGMG framework trails DTIN-Net (Jiao et al., 2022) in performance on the PRID and GRID datasets but shows marginal improvement on the VIPeR dataset. This is largely because DTIN-Net's effectiveness stems from its multi-task training, which is particularly suited for multi-source scenarios. However, when applied to single-source DG-ReID, its performance diminishes. In such single-source situations, our IGMG framework considerably outperforms DTIN-Net, as detailed in our Single-source DG-ReID experimental analysis in Table 2.

Table 5
Effect of different components of Multi-Granularity of our IGMG approach. G1 — Global Branch of Part-01; P2(:) — All the components of Part-02; P3(:) — All the components of Part-03, where G'x' refer to the global feature and l'x' refer to the local features of the corresponding branches, respectively.

Methods	M → D		D → M	
	R-1	mAP	R-1	mAP
G1	54.4	31.2	61.1	28.6
G1+ P2(G2)	58.6	35.5	64.7	32.6
G1+P2(G2+l1)	59.5	36.1	66.8	33.5
G1+P2(G2+l1+l2)	61.3	38.8	68.0	34.4
G1+P2(G2+l1+l2)+P3(G3)	61.8	38.8	69.6	36.3
G1+P2(G2+l1+l2)+P3 (G3+l1)	62.0	39.1	71.5	39.4
G1+P2(G2+l1+l2)+P3 (G3+l1+l2)	63.3	41.0	73.8	41.1
G1+P2(G2+l1+l2)+P3 (G3+l1+l2+l3)	**64.2**	**41.4**	**75.8**	**42.1**
G1+P2(G2+l1+l2)+P3 (G3+l1+l2+l3)+ P4(G4)	63.1	40.8	73.4	39.8
G1+P2(G2+l1+l2)+P3 (G3+l1+l2+l3)+P4 (G4+l1)	62.5	39.2	73.1	38.9

4.2.3. Large-scale DG-ReID (Protocol-3)

The target datasets presented in protocol-2 are small and thus may not reflect the model generalizations correctly. That is why protocol-3 was designed by Dai et al. (2021), Liao and Shao (2020) and Zhao et al. (2021) to evaluate the proposed models on large-scale datasets as shown in Table 1.

Table 4 reports the comparative performance of our IGMG approach with the relevant state-of-the-art approaches under protocol-3. IGMG outperforms all the state-of-the-art approaches by a good margin. Especially our IGMG approach improves the second best RaMoE (Dai et al., 2021) by 8.6% rank-01 and 6.2% while tested on the MSMT17 dataset. It is important to note that while our approach does not achieve the highest rank-01 accuracy on MSMT17, this discrepancy may be attributed to the face blurring effects present in the MSMT17_V2 dataset, whereas most of the previously mentioned state-of-the-art DG-ReID approaches were assessed using MSMT17_V1, which features clear faces. Nevertheless, the mAP performance for MSMT17 displays a marginal improvement over other DG-ReID counterparts, owing to its comprehensive utilization of the whole query sets. Collectively, these results affirm that our IGMG model exhibits robust generalizability across varying target dataset sizes.

4.3. Ablation study

4.3.1. Assessing the impact of multi-granularity components in our IGMG framework

To evaluate how each specific component (referred to as "Part-based analysis") impacts our IGMG framework, we carried out experiments where we progressively added components to the baseline model. The performance improvement was compared under the single-source DG-ReID setting for both $M \rightarrow D$ and $D \rightarrow M$ scenarios. The results of this analysis are presented in Table 5. As expected, among the alternative components, the performance of the global branch of Part-01 was found to be inferior. This aligns with our expectations since the global branch lacks the ability to capture fine-grained details necessary for effective domain adaptation in DG-ReID. On the other hand, the proposed IGMG network demonstrated improved performance with the addition of each granularity component. The gradual improvement observed in the results supports our claim that integrating multi-granularity can effectively address the domain shift problem in DG-ReID. Furthermore, we extended our evaluation beyond the indicated number of partitions and found that the optimal performance was achieved with the given granular architecture, which aligns with the results reported in MGN (Wang et al., 2018a). Overall, this experimental analysis confirms that the integration of multiple granularity components in our IGMG framework leads to significant performance improvements in DG-ReID, effectively mitigating the challenges posed by domain shift.

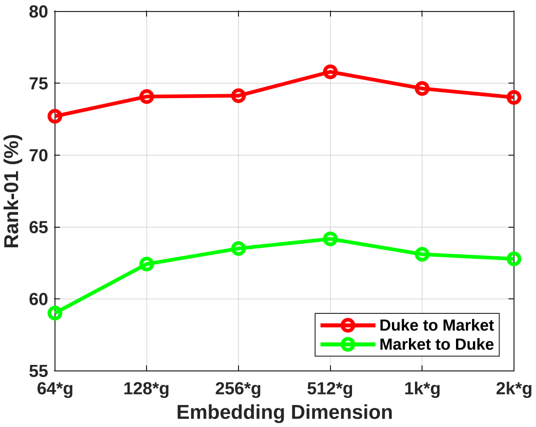

Fig. 5. Results for the optimization of the embedding size on Rank-01.

Table 6
Effectiveness of various normalization techniques on the proposed IGMG approach.

Methods	Gated Norm	Backbone Norm.	M → D		D → M	
			R-1	mAP	R-1	mAP
Baseline	–	BN	41.7	24.6	53.9	25.4
Baseline	IN (learnable)	BN	54.8	33.8	65.4	33.6
Baseline	IN (w/o learnable)	BN	56.8	36.1	69.8	37.1
IGMG	–	BN	61.9	41.1	71.1	38.6
IGMG	IN (learnable)	BN	63.3	41.0	72.3	40.4
IGMG	w/o learnable IN	FrozenBN	59.6	37.5	71.3	39.6
IGMG	w/o learnable IN	IN	62.6	40.0	75.8	42.1
IGMG	w/o learnable IN	w/o learnable IN	60.6	38.9	70.4	37.2
IGMG	w/o learnable IN	BN	**64.2**	**41.6**	**75.8**	**42.1**

4.3.2. Evaluating the role of IN within our IGMG approach

In this part of the study, we aimed to validate the effectiveness of non-learnable instance normalization (IN) compared to the learnable variant. We conducted experiments in the single-source DG-ReID setting for both $M \rightarrow D$ and $D \rightarrow M$ scenarios. Here, the "gated norm" pertains to the modules used between convolutional layers, while the "backbone norm" pertains to those within the convolutional block's residual connections. "FrozenBN" denotes the freezing of Batch norm learnable parameters to the default ImageNet values. For the purpose of comparison, we considered the Bag-of-trick (BOT) approach (Luo et al., 2019) as our baseline.

Table 6 presents the results of this experimental analysis. It demonstrates that when using IN without learnable affine parameters, both in the baseline and IGMG framework, consistently better performance was achieved compared to using IN with learnable affine parameters. Furthermore, the margin of improvement of *Baseline + IN (w/o learnable)* over the *Baseline* was higher than that of *IGMG (w/o learnable IN)* over *IGMG (without IN)*. These results clearly highlight the advantages of our proposed contribution. Firstly, the optimization based on multi-granularity proves to be effective in improving the model's generalization capabilities. Secondly, by using non-learnable parameters in the IN module, the model focuses on capturing the inherent structure of the data rather than relying on the specific statistics of each domain. Combining these two factors gives rise to our proposed *Instance guided Multi-Granularity (IGMG)* framework, which exhibits significant performance improvements over related alternatives.

4.3.3. Location of IN within the IGMG framework

In our IGMG framework, the application of Non-parametric Instance Normalization (IN) within the IGMG framework is carefully considered to address domain-specific biases in the extracted feature representations effectively. We investigate the optimal locations for applying IN and validate our findings in the Single-source DG-ReID setting;

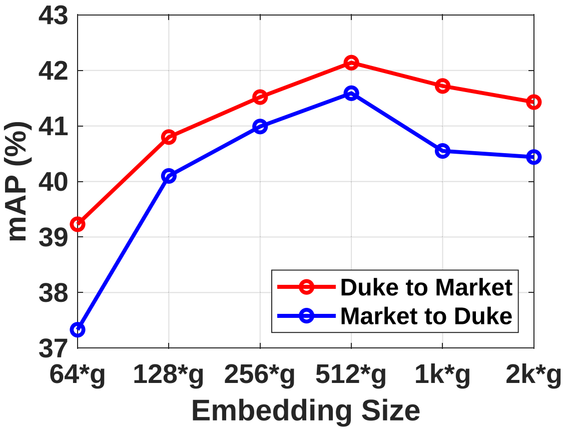

Fig. 6. Results for the optimization of the embedding size on mAP. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
Location of IN gated among convolutional blocks.

Groups	M → D		D → M	
	R-1	mAP	R-1	mAP
IN0,IN1,IN2, IN_{3-5}, IN_{6-8}	58.3	35.7	65.1	31.9
IN0,IN1,IN2, IN_{3-5}	61.0	38.7	70.6	37.1
IN0,IN1,IN2	59.8	37.3	70.1	36.5
IN1,IN2, IN_{3-5}	**64.2**	41.6	**75.8**	**42.1**
IN2, IN_{3-5}	64.0	**41.9**	73.7	41.2
IN1,IN2	61.0	39.3	73.2	41.3

specifically the $M \rightarrow D$ and $D \rightarrow M$ scenarios. IN0 is applied to the feature representation extracted from the Conv1 block. IN3, IN4, and IN5 are collectively referred to as IN_{3-5}. IN6, IN7, and IN8 are collectively referred to as IN_{6-8}, and they are applied to the final feature representation of the respective branch.

The experimental results are presented in Table 7. Consistent with the observations in SNR (Jin et al., 2020) and DualNorm (Jia et al., 2019), we find that applying non-parametric IN from the initial to mid-level layers yields better performance compared to other alternatives. These findings further support the effectiveness and relevance of our approach. The schematic illustration in Fig. 3 is based on the optimal performance we achieved in this study.

4.3.4. Ablation study of the embedding size optimization

To optimize our IGMG framework, an important hyperparameter to consider is the embedding size. We investigated the ideal embedding size based on the single-source DG-ReID setting for both $M \rightarrow D$ and $D \rightarrow M$ scenarios. The results of this analysis are depicted in Fig. 5.

In Fig. 6, it is important to note that g represents the total number of granular levels considered in our IGMG framework, which in this case is $g = 8$. Therefore, the actual embedding size is obtained by multiplying the embedding size of the individual granular component by 8. Based on the analysis, it was observed that the optimal performance was achieved when the embedding size of the individual granular component was set to 512. This finding holds true for both the $M \rightarrow D$ and $D \rightarrow M$ scenarios. In summary, the experimental analysis determined that an embedding size of 512 for the individual granular component leads to the optimal performance of our IGMG framework in the single-source DG-ReID setting

4.3.5. Performance on different backbones

To further validate our proposed IGMG approach, an ablation study is conducted considering different CNN backbones architecture. In this regard, we consider two performance evaluation scenarios: performance evaluation on another backbone and performance on different variants of ResNet architectures.

Table 8
Performance of our IGMG on OSNet (Zhou et al., 2019).

Methods	Backbones	M → D		D → M	
		R-1	mAP	R-1	mAP
Baseline	OSNet	44.9	26.4	58.8	29.1
SNR Jin et al. (2020)	OSNet	54.8	33.8	65.4	33.6
IGMG (Ours)	OSNet	**58.8**	**35.9**	**69.9**	**36.2**

Table 9
Performance of our proposed IGMG approach on different variants of ResNet.

Methods	Backbones	M → D		D → M	
		R-1	mAP	R-1	mAP
Baseline	ResNet34	39.7	22.6	49.8	23.4
IGMG (Ours)	ResNet34	**61.5**	**37.4**	**71.3**	**39.5**
Baseline	ResNet50	41.7	24.6	53.9	25.4
IGMG (Ours)	ResNet50	**64.2**	**41.6**	**75.8**	**42.1**
Baseline	ResNet101	46.8	28.4	55.9	26.4
IGMG (Ours)	ResNet101	**65.6**	**42.4**	**76.1**	**42.6**

Performance on Backbone other than ResNet: In this experiment, we aimed to evaluate the effectiveness of our proposed approach on a backbone architecture other than ResNet. We chose the Omni-Scale Network (OSNet) (Zhou et al., 2019) as the alternative backbone, which is a lightweight network commonly used in ReID tasks. Adapting our IGMG architecture for OSNet involved making some modifications compared to the ResNet-based framework.

Most of the procedures followed by the ResNet architecture remained the same for OSNet, except for the downsampling (DS) operation and the number of Instance Normalization (IN) blocks. Unlike ResNet, OSNet does not have a downsampling operation, so we did not consider it in the adaptation of our IGMG framework. However, we integrated six IN modules between the OSNet blocks of the backbone and the multi-granular architectures. In contrast, the ResNet-based IGMG framework utilized five IN modules. The experimental results for our IGMG approach on OSNet architecture, using the single-source setting of $M \rightarrow D$ and $D \rightarrow M$ scenarios, are presented in Table 8. It is evident that our proposed IGMG approach outperformed the *Baseline with OSNet* architecture significantly in all evaluation metrics for both $M \rightarrow D$ and $D \rightarrow M$ scenarios. Moreover, IGMG with OSNet achieved superior performance compared to the state-of-the-art OSNet-SNR (Jin et al., 2020), surpassing it by 4.0% and 4.5% in rank-01 for $M \rightarrow D$ and $D \rightarrow M$ scenarios, respectively.

Performance on Different Variants of ResNet: To further demonstrate the generalization capability of our proposed IGMG framework, we conducted experiments using different variants of the ResNet architecture with varying complexities. In addition to the results reported in our main manuscript using ResNet-50, we implemented the same experiments with ResNet-34 (a lighter variant) and ResNet-101 (a deeper variant).

We evaluated all models in the single-source setting of $M \rightarrow D$ and $D \rightarrow M$ scenarios. The results, as presented in Table 9, consistently show that our IGMG approach achieves significant performance gains across all three ResNet variants compared to the *Baseline (Bag of Tricks)*. While the IGMG framework with the deeper ResNet-101 exhibits superior performance among the ResNet variants, the performance gap compared to ResNet-50 is marginal. Therefore, considering the complexity and maintaining a fair comparison with state-of-the-art approaches, we chose ResNet-50 as the backbone for our extensive experimental evaluation in the main manuscript. Nonetheless, it is worth noting that our IGMG framework consistently performs well and outperforms relevant baselines and state-of-the-art approaches, regardless of the complexities of the backbone architectures used.

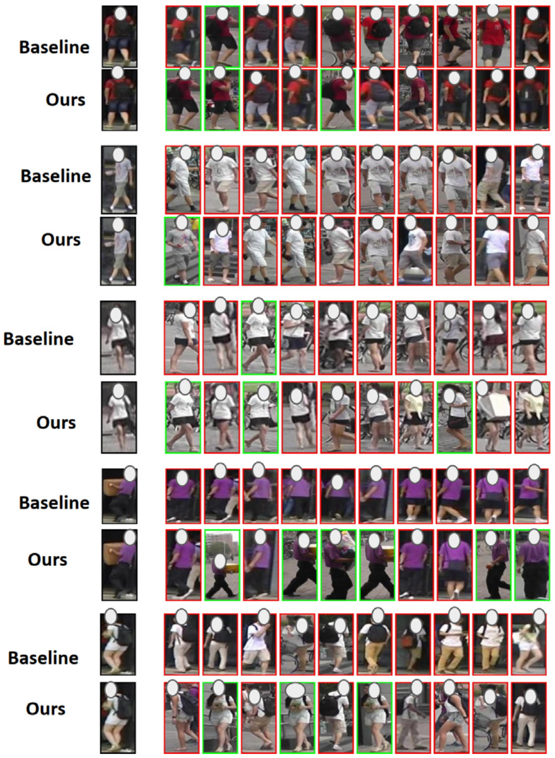

Fig. 7. Visual comparison of query images to top 10 matching images from the gallery set for five random persons in the Market1501 dataset under the single-source setting of $D \to M$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

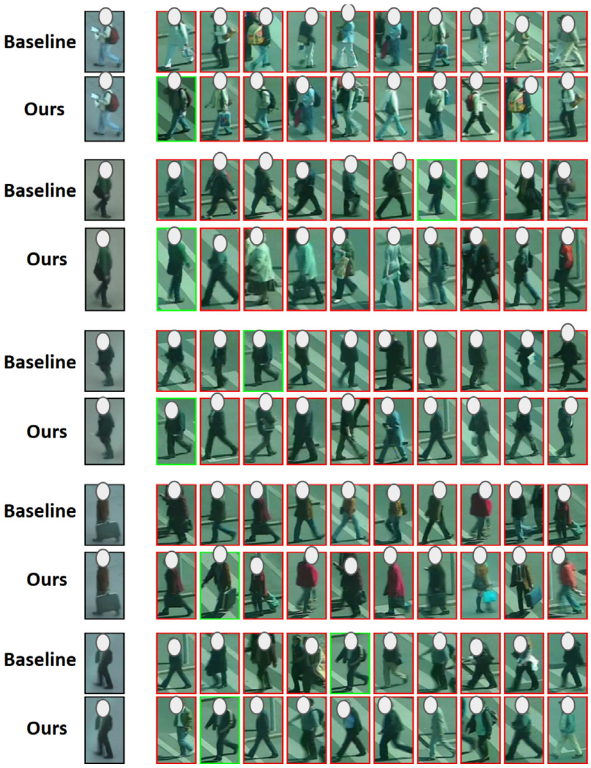

Fig. 8. Visual comparison of query images to top 10 matching images from the gallery set for five random persons in the PRID dataset under the Protocol-2 setting.

4.3.6. Qualitative matching results

The objective of this experiment is to visualize the ranking list of the gallery set corresponding to the matching results with the query. We conducted this experiment on a target dataset using the single-source setting and Protocol-2. In the single-source environment, we trained the model using the $D \to M$ scenario. In Protocol-2, we used the PRID dataset, which adopts a single-shot setting for evaluation, resulting in only one positive match available in the gallery setting, as illustrated in Fig. 8.

In Figs. 7 to 8, the first and second rows represent the ranking results produced by the baseline approach and our proposed IGMG approach, respectively. Images surrounded by green boxes indicate a match between the query and the gallery set. As observed in Figs. 7 to 8, our proposed IGMG approach consistently identifies the true match in rank-01, regardless of the capture conditions, whereas the baseline approach finds it in later ranks. We also showcase a case where our proposed IGMG approach does not find the true match in rank-01 due to a few style variations. However, these matches are eventually recognized within the first few ranks, predominantly in rank-02. These visualizations demonstrate the consistent performance of our proposed IGMG approach, regardless of the dataset size.

5. Conclusions and future work

We present a novel framework designed to address the challenges of domain-generalized person re-identification (DG-ReID). Our proposed Instance-guided Multi-Granular (IGMG) framework leverages the concept of multi-granularity, which enables the utilization of information at multiple levels of granularity, leading to more robust and flexible feature representations. By combining these different levels of information, our IGMG model captures a comprehensive representation of the underlying data, enhancing its resilience to domain shift and improving generalization to unseen domains. Furthermore, we exploit the advantages of Instance Normalization (IN) without learnable affine parameters, which introduces a gating effect into the backbone convolutional neural network (CNN). This gating effect filters out domain-specific style statistics, thereby further enhancing the model's generalization ability. Experimental results on multiple benchmark datasets demonstrate that our proposed framework consistently outperforms state-of-the-art DG-ReID methods, regardless of the dataset size. Moreover, our proposed architecture is versatile and applicable to various vision applications. Therefore, future experiments can explore the potential of our IGMG approach in different applications, using diverse backbone architectures. This flexibility and scalability make our framework a promising solution for a wide range of computer vision tasks.

CRediT authorship contribution statement

Amran Bhuiyan: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jimmy Xiangji Huang:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, Validation, Writing – review & editing. **Aijun An:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Md Amran Hossen Bhuiyan reports was provided by York University. Md Amran Hossen Bhuiyan reports a relationship with York University that includes: employment.

Data availability

Data will be made available on request.

Acknowledgments

This research is supported by research grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada (RGPIN-2020-07157, RGPIN-2019-06799, and ALLRP 546302-19), York Research Chairs (YRC) program, and Dapasoft Inc. The authors gratefully appreciate the anonymous reviewers and associate editor for their valuable comments and constructive suggestions that greatly helped to improve the quality of the paper. We also acknowledge the Digital Research Alliance of Canada. for providing us with the computing resources to conduct experiments.

References

Ahmed, E., Jones, M., Marks, T.K., 2015. An improved deep learning architecture for person re-identification. In: CVPR.

Ang, E.P., Shan, L., Kot, A.C., 2021. Dex: Domain embedding expansion for generalized person re-identification. arXiv preprint arXiv:2110.11391.

Bai, Y., Jiao, J., Ce, W., Liu, J., Lou, Y., Feng, X., Duan, L.Y., 2021. Person30k: A dual-meta generalization network for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2123–2132.

Bhuiyan, A., Huang, J.X., 2022. STCA: Utilizing a spatio-temporal cross-attention network for enhancing video person re-identification. Image Vis. Comput. 123, 104474.

Bhuiyan, A., Liu, Y., Siva, P., Javan, M., Ayed, I.B., Granger, E., 2020. Pose guided gated fusion for person re-identification. In: WACV. pp. 2675–2684.

Bhuiyan, A., Perina, A., Murino, V., 2014. Person re-identification by discriminatively selecting parts and features. In: European Conference on Computer Vision. Springer, pp. 147–161.

Chen, B., Deng, W., Hu, J., 2019a. Mixed high-order attention network for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 371–381.

Chen, G., Lin, C., Ren, L., Lu, J., Zhou, J., 2019b. Self-critical attention learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9637–9646.

Chen, F., Wang, N., Tang, J., Yan, P., Yu, J., 2023. Unsupervised person re-identification via multi-domain joint learning. Pattern Recognit. 138, 109369.

Choi, S., Kim, T., Jeong, M., Park, H., Kim, C., 2021. Meta batch-instance normalization for generalizable person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3425–3435.

Dai, Y., Li, X., Liu, J., Tong, Z., Duan, L.Y., 2021. Generalizable person re-identification with relevance-aware mixture of experts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16145–16154.

Delussu, R., Putzu, L., Fumera, G., 2023. Human-in-the-loop cross-domain person re-identification. Expert Syst. Appl. 226, 120216.

Delussu, R., Putzu, L., Fumera, G., Roli, F., 2021. Online domain adaptation for person re-identification with a human in the loop. In: 2020 25th International Conference on Pattern Recognition. (ICPR), IEEE, pp. 3829–3836.

Feng, H., Chen, M., Hu, J., Shen, D., Liu, H., Cai, D., 2021. Complementary pseudo labels for unsupervised domain adaptation on person re-identification. IEEE Trans. Image Process. 30, 2898–2907.

Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S., 2019. 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. pp. 6112–6121.

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

Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., et al., 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In: Advances in Neural Information Processing Systems, vol. 31.

Gong, T., Chen, K., Zhang, L., Wang, J., 2023. Debiased contrastive curriculum learning for progressive generalizable person re-identification. IEEE Trans. Circuits Syst. Video Technol..

Gray, D., Tao, H., 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European Conference on Computer Vision. Springer, pp. 262–275.

Hafner, F.M., Bhuiyan, A., Kooij, J.F., Granger, E., 2022. Cross-modal distillation for RGB-depth person re-identification. Comput. Vis. Image Underst. 216, 103352.

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

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

Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H., 2011. Person re-identification by descriptive and discriminative classification. In: Scandinavian Conference on Image Analysis. Springer, pp. 91–102.

Ji, R., Li, J., Zhang, L., 2023. Siamese self-supervised learning for fine-grained visual classification. Comput. Vis. Image Underst. 229, 103658.

Jia, J., Ruan, Q., Hospedales, T.M., 2019. Frustratingly easy person re-identification: Generalizing person re-id in practice. In: BMVC.

Jiao, B., Liu, L., Gao, L., Lin, G., Yang, L., Zhang, S., Wang, P., Zhang, Y., 2022. Dynamically transformed instance normalization network for generalizable person re-identification. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV. Springer, pp. 285–301.

Jin, X., Lan, C., Zeng, W., Chen, Z., Zhang, L., 2020. Style normalization and restitution for generalizable person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3143–3152.

Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., Duan, L.Y., 2022. Uncertainty modeling for out-of-distribution generalization. arXiv preprint arXiv:2202.03958.

Li, H., Dong, N., Yu, Z., Tao, D., Qi, G., 2021. Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification. IEEE Trans. Circuits Syst. Video Technol. 32 (5), 2814–2830.

Li, X., Li, Q., Liang, F., Wang, W., 2023. Multi-granularity pseudo-label collaboration for unsupervised person re-identification. Comput. Vis. Image Underst. 227, 103616.

Li, W., Zhao, R., Xiao, T., Wang, X., 2014. Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 152–159.

Liao, S., Shao, L., 2020. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In: European Conference on Computer Vision. Springer, pp. 456–474.

Liao, S., Shao, L., 2022. Graph sampling based deep metric learning for generalizable person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7359–7368.

Lin, Y., Xie, L., Wu, Y., Yan, C., Tian, Q., 2020. Unsupervised person re-identification via softened similarity learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3390–3399.

Lin, X., Zhu, L., Yang, S., Wang, Y., 2023. Diff attention: A novel attention scheme for person re-identification. Comput. Vis. Image Underst. 103623.

Liu, Y., Ge, H., Sun, L., Hou, Y., 2022. Complementary attention-driven contrastive learning with hard-sample exploring for unsupervised domain adaptive person re-id. IEEE Trans. Circuits Syst. Video Technol. 33 (1), 326–341.

Liu, Z., Mu, X., Lu, Y., Zhang, T., Tian, Y., 2023. Learning transformer-based attention region with multiple scales for occluded person re-identification. Comput. Vis. Image Underst. 229, 103652.

Loy, C.C., Xiang, T., Gong, S., 2010. Time-delayed correlation analysis for multi-camera activity understanding. Int. J. Comput. Vis. 90 (1), 106–129.

Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W., 2019. Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

Luo, C., Song, C., Zhang, Z., 2020. Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In: European Conference on Computer Vision. Springer, pp. 224–241.

Masson, H., Bhuiyan, A., Nguyen-Meidine, L.T., Javan, M., Siva, P., Ayed, I.B., Granger, E., 2021. Exploiting prunability for person re-identification. EURASIP J. Image Video Process. 2021 (1), 1–31.

Mekhazni, D., Bhuiyan, A., Ekladious, G., Granger, E., 2020. Unsupervised domain adaptation in the dissimilarity space for person re-identification. In: European Conference on Computer Vision. Springer, pp. 159–174.

Muandet, K., Balduzzi, D., Schölkopf, B., 2013. Domain generalization via invariant feature representation. In: International Conference on Machine Learning. PMLR, pp. 10–18.

Ni, H., Song, J., Luo, X., Zheng, F., Li, W., Shen, H.T., 2022. Meta distribution alignment for generalizable person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2487–2496.

Qi, L., Wang, L., Shi, Y., Geng, X., 2022. A novel mix-normalization method for generalizable multi-source person re-identification. IEEE Trans. Multimed..

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

Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., Sarawagi, S., 2018. Generalizing across domains via cross-gradient training. In: ICLR.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S., 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision. (ECCV), pp. 480–496.

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

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2016. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022.

Wang, W., Liao, S., Zhao, F., Kang, C., Shao, L., 2020. Domainmix: Learning generalizable person re-identification without human annotations. arXiv preprint arXiv:2011.11953.

Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X., 2018a. Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia. pp. 274–282.

Wang, D., Zhang, S., 2020. Unsupervised person re-identification via multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10981–10990.

Wang, C., Zhang, Q., Huang, C., Liu, W., Wang, X., 2018b. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: ECCV.

Wang, J., Zhu, X., Gong, S., Li, W., 2018c. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR.

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

Wu, D., Zhang, B., Lu, X., Li, Y., Gu, Y., Li, J., Ren, G., 2023. A domain generalization pedestrian re-identification algorithm based on meta-graph aware. Multimedia Tools Appl. 1–21.

Wu, A., Zheng, W.S., Lai, J.H., 2019. Unsupervised person re-identification by camera-aware similarity consistency learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6922–6931.

Xuan, S., Zhang, S., 2021. Intra-inter camera similarity for unsupervised person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11926–11935.

Yang, Q., Yu, H.X., Wu, A., Zheng, W.S., 2019. Patch-based discriminative feature learning for unsupervised person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3633–3642.

Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H., 2019. Unsupervised person re-identification by soft multilabel learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2148–2157.

Zhai, Y., Lu, S., Ye, Q., Shan, X., Chen, J., Ji, R., Tian, Y., 2020. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9021–9030.

Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z., 2020a. Relation-aware global attention for person re-identification. In: Proceedings of the Ieee/Cvf Conference on Computer Vision and Pattern Recognition. pp. 3186–3195.

Zhang, L., Liu, Z., Zhang, W., Zhang, D., 2023. Style uncertainty based self-paced meta learning for generalizable person re-identification. IEEE Trans. Image Process. 32, 2107–2119.

Zhang, W., Wei, Z., Huang, L., Xie, K., Qin, Q., 2020b. Adaptive attention-aware network for unsupervised person re-identification. Neurocomputing 411, 20–31.

Zhao, Y., Zhong, Z., Yang, F., Luo, Z., Lin, Y., Li, S., Sebe, N., 2021. 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. pp. 6277–6286.

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

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J., 2019. Joint discriminative and generative learning for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2138–2147.

Zheng, Z., Zheng, L., Yang, Y., 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3754–3762.

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

Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y., 2018. Camera style adaptation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5157–5166.

Zhou, K., Yang, Y., Cavallaro, A., Xiang, T., 2019. Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3702–3712.

Zou, Y., Yang, X., Yu, Z., Kumar, B., Kautz, J., 2020. Joint disentangling and adaptation for cross-domain person re-identification. In: European Conference on Computer Vision. Springer, pp. 87–104.