Joint Identity-Aware Mixstyle and Graph-Enhanced Prototype for Clothes-Changing Person Re-Identification

Zhiwei Zhao, Bin Liu, *Member, IEEE*, Yan Lu, Qi Chu, *Member, IEEE*, Nenghai Yu, and Chang Wen Chen, *Fellow, IEEE*

***Abstract*—In recent years, considerable progress has been witnessed in the person re-identification (Re-ID). However, in a more realistic long-term scenario, the appearance shift arising from the clothes-changing inevitably deteriorates the conventional methods that heavily depend on the clothing color. Although the current clothes-changing person Re-ID methods introduce external human knowledge (i.e, contour, mask) and sophisticated feature decoupling strategy to alleviate the clothing shift, they still face the risk of overfitting to clothing due to the limited clothing diversity of training set. To more efficiently and effectively promote the clothes-irrelevant feature learning, we present a novel joint Identity-aware Mixstyle and Graph-enhanced Prototype method for clothes-changing person Re-ID. Specifically, by treating the cloth-changing as fine-grained domain/style shift, the identity-aware mixstyle (IMS) is proposed from the perspective of domain generalization, which mixes the instance-level feature statistics of samples within each identity to synthesize novel and diverse clothing styles, while retaining the correspondence between synthesized samples and latent label space. By incorporating the IMS module, the more diverse styles can be exploited to train a clothing-shift robust model. To further reduce the feature discrepancy caused by clothing variations, the graph-enhanced prototype constraint (GEP) module is proposed to explore the graph similarity structure of style-augmented samples across memory bank to build informative and robust prototypes, which serve as powerful exemplars for better clothing-irrelevant metric learning. The two modules are integrated into a joint learning framework and benefit each other. The extensive experiments conducted on clothes-changing person Re-ID datasets validate the superiority and effectiveness of our method. In addition, our method also shows good universality and corruption robustness on other Re-ID tasks.**

***Index Terms*—Clothes-changing person Re-ID, graph-enhanced prototype, identity-aware mixstyle, long-term, robustness.**

Manuscript received 19 December 2022; revised 9 May 2023; accepted 17 August 2023. Date of publication 1 September 2023; date of current version 14 February 2024. This work was supported by the National Natural Science Foundation of China under Grant 62272430. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Stephane Lathuiliere. *(Corresponding author: Bin Liu.)*

Zhiwei Zhao, Bin Liu, Yan Lu, Qi Chu, and Nenghai Yu are with the School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230052, China, and also with the Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Beijing 100864, China (e-mail: zwzhao98@mail.ustc.edu.cn; flowice@ustc.edu.cn; luyan17@mail.ustc.edu.cn; qchu@ustc.edu.cn; ynh@ustc.edu.cn).

Chang Wen Chen is with the Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: changwen.chen@polyu.edu.hk).

Digital Object Identifier 10.1109/TMM.2023.3311143

I. Introduction

PERSON re-identification (Re-ID) [1] aims at associating and matching pedestrian images across multiple cameras, which plays a vital role in intelligent video analysis and multimedia retrieval. Most previous works [2], [3], [4], [5], [6] mainly study the person Re-ID in a short-term scenario, where the underlying assumption is that the clothes of pedestrian are stable over a short span of time. By introducing feature learning [2], [5] and metric learning [4], [6] strategies, significant progress has been witnessed. However, in a more realistic long-term scenario, pedestrian often changes their clothes across days and weathers, and the appearance shift arising from clothes-changing significantly degrades the performance of short-term person Re-ID methods that heavily rely on the clothing information. To address this problem, in this article, we mainly focus on the more realistic yet challenging **clothes-changing** person Re-ID task.

The main idea of current works [7], [8], [9], [10], [11], [12], [13] in clothes-changing person Re-ID is to pursue the clothing-irrelevant biometric features and can be roughly divided into two categories. The former is based on the auxiliary external knowledge such as body contours [7], [9], keypoints [12], semantic mask [8] and 3D information [10] to assist biometric feature learning. The latter designs various feature disentanglement strategies [13], [14], [15] to decouple the feature into clothing-related and clothing-unrelated components. Despite some progress has been achieved, the existing approaches still have limitations. Due to the relatively limited diversity of pedestrian clothing in the training set, although these existing methods impose the sophisticated knowledge and feature regularizations, they still have the risk of overfitting to the seen clothes, and can not sufficiently learn the clothing-irrelevant cues of a person. Therefore, it is necessary to seek a simple yet effective method to better learn the clothes-irrelevant features.

To more efficiently and effectively promote the clothes-irrelevant feature learning, motivated by the fact that the clothes-changing of pedestrian can be treated as a fine-grained **domain/style** shift, we consider the potential solution mainly lies in two aspects, **(1)** Enriching the clothing style diversity of pedestrians during the training phase from the perspective of domain generalization. **(2)** Reducing the feature discrepancy for each pedestrian caused by the clothing shift from the perspective of metric learning. In light of the above discussions, in this article,

we present a novel joint Identity-aware Mixstyle and Graph-enhanced Prototype method for long-term clothes-changing person Re-ID task.

Specifically, we first propose the identity-aware mixstyle (**IMS**) module to enrich clothing style diversity of training set for shift-robust model training. The IMS module is inspired by the Mixstyle [16] method that performs style mixture between two arbitrary samples for domain generalization task. Different from Mixstyle, our IMS module randomly mixes instance-level feature statistics of samples **within each identity** to synthesize novel and diverse clothing styles for cloth-changing person Re-ID task. The proposed IMS module enjoys two merits. Firstly, since the Mixstyle blends the feature statistics of two random instances without considering identity label and feature statistics still encapsulate identity-related information, the Mixstyle is inevitably disturbed by other identities and corrupts inherent label space, resulting in sub-optimal performance. Compared with Mixstyle, the IMS module can effectively guarantee the correspondence between the synthesized samples and label space by highlighting the within-identity style mixture, leading to superior performance. Secondly, the novel and diverse styles are effectively synthesized on-the-fly in shallow network layers by IMS module during the training phase, which can be used to enhance the model's robustness against clothes-changing and to enforce model to learn clothing-irrelevant features.

To further decrease the feature discrepancy of each identity caused by clothing shift, the graph-enhanced prototype constraint (**GEP**) module is developed to capture more informative and robust prototypes with the graph structure knowledge, where the prototype are served as powerful exemplar for better clothing-irrelevant metric learning. Concretely, the GEP module initially builds a KNN graph to propagate the relation of style-augmented samples across a dynamic memory bank, hence the complementary messages are spread and a compact clustering structure of samples can be discovered. Then the GEP module estimates prototypes by the mean embedding of graph-enhanced features belonging to the same identity across the memory bank, which can well depict the intrinsic characteristics of identities. Finally, the compactness between style-augmented instances and prototypes is enforced to alleviate the feature discrepancy caused by clothing variation. The two modules are unified in a joint learning framework and collaborate each other. The IMS provides more diverse styles of samples to estimate more informative prototypes, while the GEP in turn serve prototypes as powerful exemplars to guide better metric learning for style-augmented instances.

Our main contributions can be summarized as follows:

1) Treating the cloth-changing as fine-grained domain/style shift, we propose the identity-aware mixstyle for clothes-changing person Re-ID from the perspective of domain generalization, which mixes instance-level feature statistics of samples within each identity to synthesize novel and diverse styles for shift-robust model training.
2) The graph-enhanced prototype constraint is developed to build informative prototypes by exploring the graph relation structure of samples across memory bank and toward better clothing-irrelevant metric learning.
3) The extensive experiments conducted on several *clothes-changing* person Re-ID datasets validate the effectiveness and superiority of our method against the state-of-the-arts. The further experiments on the *short-term* Re-ID and *corruption* Re-ID tasks highlight the universality and robustness of our method.

II. Related Work

A. Short-Term Person Re-ID

Benefiting from the development of deep neural networks, recent years have witnessed considerable success in person Re-ID. Existing short-term person Re-ID methods can be divided into feature-learning and metric-learning based methods. The former focuses on designing effective strategies to learn discriminative features and the representative methods leveraged the horizontal slicing [2], pose estimation [17] and semantic parsing [18] to capture local pedestrian features, respectively. The other feature-learning methods [5], [19] concentrate on utilizing auxiliary information such as view/camera label to learn camera-invariant features. The metric-learning based methods emphasize the design of metric loss functions to facilitate the intra-class compactness and inter-class separability, the typical loss functions include contrastive loss and hard-triplet loss [4]. However, most of the above works focus on the short-term scenarios and heavily rely on clothing color cues, which are ineffective when facing clothes-changing of pedestrian in a long-term scenario.

B. Clothes-Changing Person Re-ID

In recent years, several works [7], [8], [9], [10], [11], [12], [20], [21], [22] about clothes-changing person Re-ID have appeared, how to discover clothing-irrelevant biometric features lies in the heart of these works. Meanwhile, some clothes-changing person Re-ID datasets like PRCC [7], LTCC [12], VC-Clothes [23], COCAS [24], DeepChange [25] and LaST [26] have also emerged. In general, the existing works of clothes-changing Re-ID can be broadly grouped into two categories.

The former is based on the external human prior knowledge to assist the clothing-irrelevant feature mining, Yang et al. [7] designed a contour-based network to learn curve patterns for overcoming clothing change, Hong et al. [8] developed a shape-appearance mutual learning method to learn identity-guided human mas and enhance the shape features. Yu et al. [24] proposed a two-branch biometric-clothes network with the aid of clothes detector and human parser, and collected the COCAS datasets. Qian et al. [12] leveraged the human keypoints to encode shape embedding and distilling the identity-relevant shape feature, while Jia et al. [20] adopted the human semantic parser to randomly assembe the semantic cloth patches. Chen et al. [10] presented to learn texture-insensitive shape embedding by 3D human prior. However, despite the introduction of the complex and time-consuming knowledge regularizations, these methods still face the risk of overfitting to clothing due to the limited clothing diversity of training set. As a contrast, our method does not rely on any external knowledge and can efficiently enrich

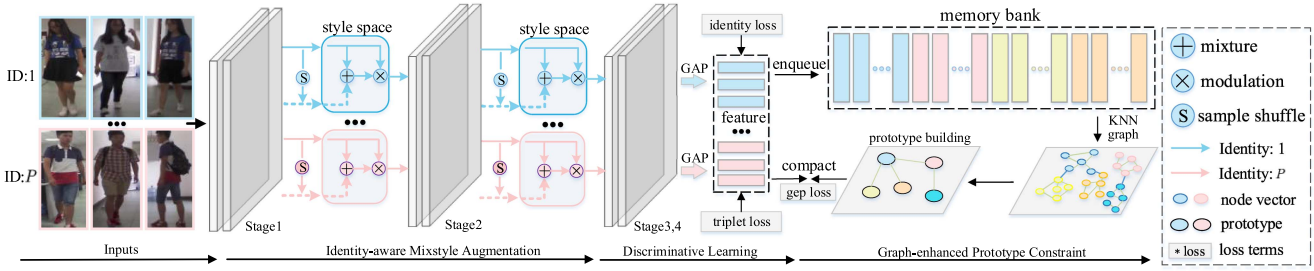

Fig. 1. Overview framework of our method. Given a batch of samples with cloth-changing shift, the identity-aware mixstyle (IMS) module mixes instance-level feature statistics of samples within each identity to simulate more novel and diverse styles for robust model training against the clothing shift. Furthermore, the graph-enhanced prototype constraint (GEP) is proposed to eliminate the feature discrepancy caused by clothing shift. It leverages the graph similarity structure of samples across a memory bank to build more informative prototype, and further utilizes the established prototype as exemplar to enhance instance-to-prototype compactness. The two modules are integrated into a joint learning framework to effectively learn clothing-irrelevant features.

the style diversity during training to enhance the robustness of model against the clothing style shift.

Another group of methods revolves around the feature decoupled learning. Xu et al. [14] leveraged an adversarial feature disentanglement strategy to decouple the identity-related and identity-unrelated (clothing) feature, while Gu et al. [21] presented an adversarial loss (CAL) that based on the clothes label to mine clothes-irrelevant features. However, the feature decoupling training [14] is sophisticated and not stable, and the clothes labels used in [21] are usually time-consuming to collect in practice, which hinders the scalability of such methods. Based on the CAL, Cui et al. [22] proposed to disentangle the clothing relevant and irrelevant features by the reconstruction of human component regions. However, this method relies heavily on both the human parser, contour and clothing labels. In contrast to the above methods, our approach does not involve complex decoupling strategies or clothes label annotation or external knowledge during training, which is more simple and efficient to learn clothing-irrelevant features for clothes-changing person Re-ID.

C. Domain Generalization

Domain Generalization (DG) [27] aims to learn a model that can generalize well to the unseen domain when given multiple source domains. To overcome the domain shift problem and achieve out-of-distribution generalization, the current DG methods are mainly based on data augmentation [16], [28] and domain-invariant learning [29], [30]. The data augmentation based methods focus on manipulating the inputs with transformations or simulating domain shift with label-preserving. Typical methods including the domain randomization [28], Mixstyle [16]. Inspired by the DG task, we consider that both the clothes-changing and the background shift that appear in clothes-changing person Re-ID can be treated as fine-grained domain shift, hence it is feasible to enhance the robustness of Re-ID model from the perspective of domain generalization. However, we find that directly applying the original Mixstyle [16] for the long-term cloth-changing Re-ID is sub-optimal. In view of this, we present a novel identity-aware mixstyle module to inherit the advantages and avoid the shortcomings of original Mixstyle for clothes-changing person Re-ID.

III. Proposed Method

In this section, we first describe the overview framework of our method. Then the details of the identity-aware mixstyle and the graph-enhanced prototype constraint modules are elaborated respectively. Finally, we show the overall optimization manner of the joint learning framework.

A. Overview Framework

The overview architecture of our method is illustrated in Fig. 1, which involves two key novel components, the identity-aware mixstyle (**IMS**) and the graph-enhanced prototype constraint (**GEP**). Given a batch of pedestrian images with the clothing variations, the IMS module probabilistically mixes instance-level feature statistics of samples within each identity to synthesize novel and diverse styles at multiple shallow layers, which efficiently promotes the shift-robust model training. Meanwhile, the synthesized samples are enqueued into a dynamic memory bank. The GEP module utilizes the graph similarity structure of samples across the memory bank to build more informative prototypes, and further adopts the prototypes as exemplars to strengthen the compactness between style-augmented instances and prototype. The two modules are unified in an end-to-end framework and mutually benefit to effectively learn the clothing-irrelevant features.

B. Identity-Aware Mixstyle Augmentation

The cloth-changing and background shift that occur in clothes-changing person Re-ID can be naturally considered as fine-grained domain/style shift, which influences the model performance. From the perspective of domain generalization, the identity-aware mixstyle is dedicated to synthesizing novel and diverse styles of pedestrians during training phase to enrich the clothing style diversity of training set. By incorporating more style-augmented instances into training phase, the robustness of the learned model against the cloth-changing and background shift can be remarkably enhanced, and the clothing-irrelevant features can be more fully mined.

Specifically, we denote a batch of input pedestrian images as $\mathcal{X} = \{\boldsymbol{x}_1^1, \boldsymbol{x}_i^1, \boldsymbol{x}_K^1, \ldots, \boldsymbol{x}_1^p, \boldsymbol{x}_i^p, \boldsymbol{x}_K^p, \ldots, \boldsymbol{x}_1^P, \boldsymbol{x}_i^P, \boldsymbol{x}_K^P\}$,

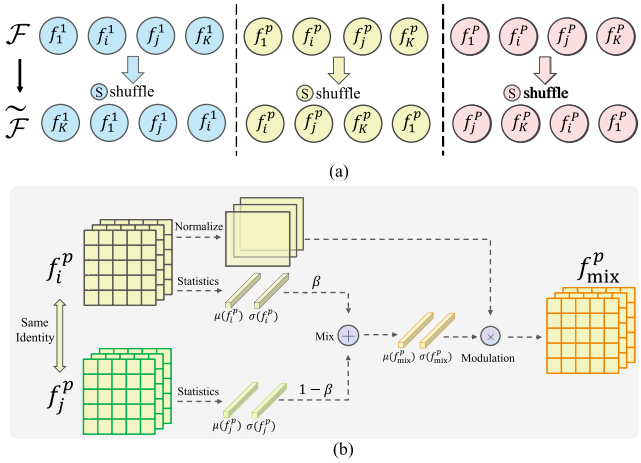

Fig. 2. Illustration of the identity-aware mixstyle module. (a) The description of how a reference batch $\widetilde{\mathcal{F}}$ involved in IMS module is generated, where the identity label is denoted by color. (b) The flowchart of IMS module, which probabilistically selects two samples of the same identity to conduct the style mixing at shallow CNN layers.

consisting of P identities and K instances for each identity, and $\boldsymbol{x}_i^p$ denotes the i-th instance of the p-th identity. Due to the random batch sampling of the clothes-changing training set, each identity in the $\mathcal{X}$ usually has multiple different clothes. As shown in the Fig. 1, the input image set $\mathcal{X}$ is then fed into the widely used ResNet50 [31] backbone to extract the feature map at shallow layers, such as the end layer of the stage1 and stage2 of the ResNet. The extracted feature map set is denoted as $\mathcal{F} \in \mathbb{R}^{N\times C\times H\times W}$, where N, C, H and W denote the batch-size, channel number, height and width of feature map, respectively. Specifically, $\mathcal{F} = \{\boldsymbol{f}_1^1, \boldsymbol{f}_i^1, \boldsymbol{f}_K^1, \ldots, \boldsymbol{f}_1^p, \boldsymbol{f}_i^p, \boldsymbol{f}_K^p, \ldots, \boldsymbol{f}_1^P, \boldsymbol{f}_i^P, \boldsymbol{f}_K^P\}$, where the $\boldsymbol{f}_i^p \in \mathbb{R}^{C\times H\times W}$ indicates the feature map of $\boldsymbol{x}_i^p$ and here we omit indicator of network layer for convenience.

After extracting the feature map set $\mathcal{F}$ at the shallow layers, the identity-aware mixstyle (IMS) augmentation then efficiently synthesizes more samples with novel and diverse styles in the style space on-the-fly. To seamlessly integrate the IMS module into the standard mini-batch training, as shown in the Fig. 2(a), the IMS module first generates a reference batch $\widetilde{\mathcal{F}} \in \mathbb{R}^{N\times C\times H\times W}$ by randomly shuffling the K instances within each identity of the $\mathcal{F}$, which means that the $\widetilde{\mathcal{F}} = \{\mathbf{S}[\boldsymbol{f}_1^1, \boldsymbol{f}_i^1, \boldsymbol{f}_K^1], \ldots, \mathbf{S}[\boldsymbol{f}_1^p, \boldsymbol{f}_i^p, \boldsymbol{f}_K^p], \ldots, \mathbf{S}[\boldsymbol{f}_1^P, \boldsymbol{f}_i^P, \boldsymbol{f}_K^P]\}$. As depicted in Fig. 1, the $\mathbf{S}$ denotes the shuffle operation of samples within each identity and it is random and independent for each identity in the mini-batch.

Next, different from the Mixstyle method, the IMS module randomly mixes the instance-level feature statistics of two samples belonging to the same identity. As shown in Fig. 2(b), given the $\boldsymbol{f}_i^p$ and $\boldsymbol{f}_j^p$ of the same p-th identity, which are indexed by the same position within $\mathcal{F}$ and $\widetilde{\mathcal{F}}$, respectively, we first compute the channel-wise feature statistics of them. As shown in (1) and (2), $\mu(\boldsymbol{f}) \in \mathbb{R}^C$ and $\sigma(\boldsymbol{f}) \in \mathbb{R}^C$ denote the channel-wise mean and standard deviation across the spatial dimension of $\boldsymbol{f}$, respectively.

$$\mu(\boldsymbol{f}) = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\boldsymbol{f}_{h,w}, \tag{1}$$

$$\sigma(\boldsymbol{f}) = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(\boldsymbol{f}_{h,w} - \mu(\boldsymbol{f}))^2}. \tag{2}$$

As suggested by Mixstyle [16] and AdaIN [32], style information of an image is mainly preserved in the above instance-level feature statistics $\mu(\boldsymbol{f})$ and $\sigma(\boldsymbol{f})$ of the shallow layers. Then the IMS module conducts the identity-aware style mixture to the feature statistics of $\boldsymbol{f}_i^p$ and $\boldsymbol{f}_j^p$ by

$$\mu_{\text{mix}}^p = \beta * \mu(\boldsymbol{f}_i^p) + (1-\beta) * \mu\left(\boldsymbol{f}_j^p\right), \tag{3}$$

$$\sigma_{\text{mix}}^p = \beta * \sigma(\boldsymbol{f}_i^p) + (1-\beta) * \sigma\left(\boldsymbol{f}_j^p\right), \tag{4}$$

where μ_{mix}^p and σ_{mix}^p indicates the mixed mean and standard deviation of $\boldsymbol{f}_i^p$ and $\boldsymbol{f}_j^p$ within the p-th identity, the mixing factor β is randomly sampled from priori beta distribution $\text{Beta}(\alpha, \alpha)$ and α controls the shape of Beta distribution. The β is independent and random for each instance and each layer. By (3)–(4), more novel and unknown styles can be synthesized during the training phase.

Finally, we leverage the statistics μ_{mix}^p and σ_{mix}^p to modulate the style-normalized $\boldsymbol{f}_i^p$ and obtain the style-mixed instance $\boldsymbol{f}_{\text{mix}}^p$ by the following equation.

$$\boldsymbol{f}_{\text{mix}}^p = \sigma_{\text{mix}}^p \frac{\boldsymbol{f}_i^p - \mu(\boldsymbol{f}_i^p)}{\sigma(\boldsymbol{f}_i^p)} + \mu_{\text{mix}}^p. \tag{5}$$

The IMS modules are hierarchically inserted at the end of stage1 and stage2 of ResNet and activated with probability of 0.5 following original Mixstyle [16]. The (1)–(5) applied on the $< \boldsymbol{f}_i^p, \boldsymbol{f}_j^p >$ can be naturally extended to $< \mathcal{F}, \widetilde{\mathcal{F}} >$ for standard mini-batch training. Due to the randomness and independence in the reference batch generation and style mixture stage, when $\boldsymbol{f}_i^p$ and $\boldsymbol{f}_j^p$ are with the different clothes, the probabilistic convex mixture between the feature statistics of them can simulate more novel clothing style/domain to enrich the clothing diversity. When $\boldsymbol{f}_i^p$ and $\boldsymbol{f}_j^p$ are with the same clothes, the IMS can simulate rich background shift. Note that both of the above cases fully emerge along with multiple training iterations. By involving the online style-augmented samples for training, the model robustness against the clothing shift and background variation can be remarkably boosted, and the overfitting problem can be alleviated.

Our proposed IMS module is inspired by Mixstyle [16], but avoids the shortcomings of Mixstyle by designing the within-identity style mixture. As suggested by [33], the feature statistics not only encapsulate considerable style information but also contain some identity-related cues. However, the original Mixstyle [16] blends the feature statistics of two instances across the batch with an identity-unaware manner, which unavoidably introduces interference from other identities and corrupts the correspondence between synthetic samples and inherent label space, thus leading to sub-optimal performance. Compared with Mixstyle, the IMS emphasizes the importance of

mixing style statistics of two samples within each identity for clothes-changing person Re-ID, and adopts the intra-identity sample shuffle and intra-identity style mixture, which effectively guarantees the consistency between synthesized samples and labels, thus well maintaining discrimination of learned features and resulting in superior performance.

We compare our IMS module with the identity-aware mixup (IMU), which interpolates the raw image space of paired images by mixup [34], Certainly, both the IMS and IMU belong to data augmentation techniques, which can be employed to enhance the model robustness. However, they differ significantly in that IMS manipulates more powerful style statistics of feature maps in multiple shallow layers, while IMU only blends images at the input pixel level. Compared to IMU, IMS can synthesize more diverse styles more effectively and interpretably, which is useful for improving the style robustness of models against the clothing changes of pedestrian. We also quantitatively validated the advantages of IMS over IMU through comparative experiments in Section IV.

C. Graph-Enhanced Prototype Constraint

The IMS module efficiently reinforces the robustness of model against cloth-changing by incorporating the online style-augmented samples for training. Meanwhile, how to exploit these samples and enforce model to mine clothes-irrelevant features needs to be explored. In view of this, the graph-enhanced prototype constraint (GEP) module is proposed to build more informative and robust prototype with the graph similarity structure of style-diversified samples across memory bank. These prototypes serve as powerful exemplars to guide better clothing-irrelevant metric learning and help to mitigate the feature discrepancy caused by clothes-changing.

As shown in Fig. 1, after the IMS augmentation at shallow layers, the intermediate feature maps are fed into the deep layers (stage3-4) and followed by global average pooling (GAP) to obtain the features. We denote the extracted features of $\mathcal{X}$ as $\mathcal{V} = \{\boldsymbol{v}_i\}_{i=1}^{N}$, where $\boldsymbol{v}_i \in \mathbb{R}^{2048}$ indicates the feature vector. Then the $\mathcal{V}$ is enqueued into a memory bank $\mathcal{M} \in \mathbb{R}^{B\times 2048}$, the B denotes the size of memory bank and is set to 8192. The memory bank $\mathcal{M}$ is constructed as a dynamic queue with the current batch enqueued and the oldest batch dequeued following [35]. Specifically, at each training iteration, we enqueue the feature vectors of the current batch $\mathcal{X}$, and dequeue the oldest batch. The memory bank is initialized by filling the features of a set of training images extracted from the initial model. During training, the memory bank dynamically records a large quantity of style-augmented samples in past and current iterations, which efficiently enlarges the scale and diversity of available samples for the graph construction and the prototype estimation.

In clothes-changing Re-ID, due to the considerable clothes-changes and batch-sampling uncertainty, the typical center loss [36] using naive class centroid of mini-batch is insufficient for effective metric learning. To fully exploit style-diversified samples and discover more informative prototype for desirable metric guidance, the GEP module first builds the KNN graph to propagate the relation of style-diversified samples across memory bank $\mathcal{M}$, which can exploit superior global clustering structure based on graph similarity. Considering the KNN graph $\mathcal{G} = (\mathcal{M}, \mathcal{E})$, each node in $\mathcal{G}$ represents a feature vector $\boldsymbol{m}_i$ in the memory bank $\mathcal{M}$ and it is connected to $\mathbb{K}$ nearest neighbors, yielding $\mathbb{K}$ edges belonging to $\mathcal{E}$. The edge affinity e_{ij} is determined by the cosine similarity between the feature vector $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$. To generate the attention coefficient ω_{ij} for each node in $\mathcal{M}$, we first obtain the top-K neighbor index $\mathcal{K}_i$ from the memory bank $\mathcal{M}$ for each node by the sorted function $\mathcal{T}(\cdot, \mathbb{K}, \mathcal{M})$, as shown in (6). Then the attention coefficient ω_{ij} is obtained by normalizing edge affinity e_{ij} among the top-K neighbors using a softmax function as described in (7).

$$\mathcal{K}_i = \mathcal{T}(\{e_{ij}\}_{j=1}^{B}, \mathbb{K}, \mathcal{M}), \tag{6}$$

$$\omega_{ij} = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{K}_i}\exp(e_{ik})}. \tag{7}$$

The attention coefficient ω_{ij} reflects the relation intensity between each node and it's top-$\mathbb{K}$ style-augmented neighbors in the $\mathcal{M}$. Then each node attentively aggregates the features of top-$\mathbb{K}$ neighbors by linear combination, and the aggregated feature vector of node $\boldsymbol{m}_i$ can be calculated by

$$\vec{\boldsymbol{m}}_i = \sum_{k\in\mathcal{K}_i, m_k\in\mathcal{M}} \omega_{ik}\boldsymbol{m}_k. \tag{8}$$

To stabilize the training process, the $\vec{\boldsymbol{m}}_i$ and $\boldsymbol{m}_i$ are further aggregated together to obtain final feature vector $\overline{\boldsymbol{m}}_i$ by

$$\overline{\boldsymbol{m}}_i = \frac{1}{2}(\vec{\boldsymbol{m}}_i + \boldsymbol{m}_i). \tag{9}$$

By exploring the underlying graph similarity structure of style-diversified samples across memory bank $\mathcal{M}$, each node could absorb complementary messages from neighbors and a compact clustering structure can be discovered. This superior structure is beneficial to subsequently build more informative and reliable prototypes rather than only based on local and isolated batch. We further estimates the identity-specific prototype $\boldsymbol{c}_p \in \mathbb{R}^{2048}$ by the mean embedding of graph-enhanced nodes belonging to the p-th identity across entire memory bank $\mathcal{M}$ by the (10), where N_p denotes the sample number of p-th identity in memory bank, $|\mathcal{M}|$ is size of memory bank, y_i denotes the identity label of $\overline{\boldsymbol{m}}_i$, $\mathbb{1}(y_i = p)$ is the indicator function, which equals to 1 if $y_i = p$ otherwise 0.

$$\boldsymbol{c}_p = \frac{1}{N_p}\sum_{i=1}^{|\mathcal{M}|}\mathbb{1}(y_i = p)\cdot\overline{\boldsymbol{m}}_i. \tag{10}$$

The dynamically estimated prototype set $\mathcal{C}$ could efficiently capture more intrinsic characteristics of identities from the view of global relation, therefore shows better robustness to clothing shift and offers superior metric properties. Unlike the graph-based conventional person Re-ID method MNE [37] that directly optimizes the graph-enhanced embeddings of mini-batch with cross-entropy loss, the GEP module uses the estimated prototypes $\mathcal{C}$ reinforced by the graph structure knowledge of style-diversified samples across the $\mathcal{M}$ as powerful exemplars

to promote better clothing-irrelevant feature learning for cloth-changing Re-ID, which can effectively eliminate intra-class feature discrepancy for style-augmented instances $\mathcal{V}$ of current batch. We achieve this constraint by minimizing the distance between the instances of $\mathcal{V}$ and the corresponding prototype with (11). More importantly, different from the MNE [37] that needs auxiliary image set to construct the graph structure during testing, our GEP constraint in training process seamlessly transferred graph-enhanced structure knowledge to feature set $\mathcal{V}$ that no graph enhancement applied, which avoids graph construction and post-processing during testing.

$$\mathcal{L}_{\text{gep}}(\mathcal{V},\mathcal{C}) = \frac{1}{PK}\sum_{p=1}^{P}\sum_{k=1}^{K}||\boldsymbol{v}_k^p - \boldsymbol{c}_p||_2. \tag{11}$$

D. Overall Optimization

As shown in Fig. 1, a batch of pedestrian images $\mathcal{X}$ are first fed into the network $\phi(\cdot)$ embedded with the identity-aware mixstyle augmentation to extract the feature vector set $\mathcal{V}$. Then, we append a classifier and impose the identity loss $\mathcal{L}_{\text{identity}}$ on the extracted feature set $\mathcal{V}$ by

$$\mathcal{L}_{\text{identity}}(\mathcal{V}) = \mathbb{E}_i[-\log(p(y_i|\boldsymbol{v}_i))], \tag{12}$$

where y_i denotes the ground truth label of $\boldsymbol{v}_i$ and $p(y_i|\boldsymbol{v}_i)$ indicates the predicted probability of $\boldsymbol{v}_i$ belonging to the y_i. In addition, we also apply the hard triplet loss $\mathcal{L}_{\text{triplet}}$ [8] to decrease the intra-identity distance and enlarge the inter-identity separability on the feature set $\mathcal{V}$ by

$$\mathcal{L}_{\text{triplet}}(\mathcal{V}) = \mathbb{E}_{\boldsymbol{v}\in\mathcal{V}}[d(\boldsymbol{v}_a,\boldsymbol{v}_p) + m - d(\boldsymbol{v}_a,\boldsymbol{v}_n)]_+, \tag{13}$$

where $\boldsymbol{v}_a$ is the anchor sample, and the $\boldsymbol{v}_p$ and $\boldsymbol{v}_n$ denote the hardest positive and negative sample of $\boldsymbol{v}_a$ in $\mathcal{V}$, respectively. $d(\cdot,\cdot)$ represents the distance operation and $[\cdot]_+ = \max(\cdot, 0)$. Furthermore, the GEP module effectively boosts the clothing-irrelevant metric learning with $\mathcal{L}_{\text{gep}}(\mathcal{V},\mathcal{C})$.

The entire model can be optimized by minimizing the overall loss function $\mathcal{L}_{\text{overall}}$ with (14), where the hyper-parameter λ is the loss weight of $\mathcal{L}_{\text{gep}}(\mathcal{V},\mathcal{C})$ constraint.

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{identity}}(\mathcal{V}) + \mathcal{L}_{\text{triplet}}(\mathcal{V}) + \lambda\mathcal{L}_{\text{gep}}(\mathcal{V},\mathcal{C}). \tag{14}$$

In the training phase, the IMS modules are only embedded in the shallow style space of the network front-end (i.e, stage-1,2) to synthesize samples with diverse styles, while the deep layers of the network (i.e., stage3 and stage4 of resnet50) do not have IMS module embedded. Under the optimization objective of minimizing the joint loss consisting of the identity loss, triplet loss, and GEP loss, the deep layers of the network are enforced to learn clothes-irrelevant features effectively. Simultaneously, the robustness of the model against the clothes changing can be remarkably enhanced.

Once the model training is finished, in the testing/inference stage, both the identity-aware mixstyle augmentation (IMS) and graph-enhanced prototype (GEP) modules are not used.

IV. Experiments

A. Datasets and Protocols

We conduct extensive experiments on **five** clothes-changing person Re-ID datasets and two short-term Re-ID datasets. The experiments are carried out on a wide range of tasks including cloth-changing, short-term and corruption person Re-ID.

1) The Clothes-Changing Person Re-ID Datasets:

1) The **PRCC** [7] dataset is collected from 3 cameras and each identity has clothes changing. For Camera C, the clothing of each person is different from the clothing they wore in Cameras A and B. There are 33,698 images from 221 identities in PRCC dataset.
2) The **VC-Clothes** [23] is a synthetic dataset by the game engine which contains 512 virtual identities of 19,060 images in 4 different scenes with remarkable clothes changes. There are 256 identities for training and the other unseen 256 identities for testing in VC-Clothes.
3) The **LTCC** [12] dataset contains 17,138 images of 152 identities, capturing 478 different outfits from 12 camera views. On average, there are 5 different outfits for each cloth-changing identity in LTCC.
4) The **DeepChange** [25] is a large, realistic long-term person Re-ID dataset with diverse clothes change and weather conditions (sunny, cloudy, windy, rainy, etc.), which consists of 178 K images of 1.1 K person identities from 17 cameras, collected over a course of 12 months.
5) The **LaST** [26] is a large-scale long-term person Re-ID dataset with larger spatial and temporal ranges, which exhibits rich variations in clothing, light, weather, and locations. It contains 5,000 identities and 71,248 images for training, 5,806 identities and 135,529 images for testing.

2) The Short-Term Object Re-ID Datasets: We also evaluated method on two short-term object Re-ID datasets include **Market1501** [38] and **VeRi-776** [39]. The Market1501 is a commonly-used person Re-ID datasets without the cloth-changing, Veri-776 is a widely used vehicle Re-ID dataset.

3) Evaluation Protocols: We adopt the Rank-1 accuracy and mean average precision (mAP) as the main evaluation metrics. For long-term clothes-changing datasets PRCC, VC-Clothes, LTCC, DeepChange, we follow their original evaluation protocols and conduct evaluation in cloth-changing evaluation setting. We also report the result under cloth-unchanging setting for PRCC and VC-Clothes. For PRCC, we use the single-shot matching by randomly choosing one image for each identity to form the gallery set and repeating 10 times to compute the average performance following [8], [10], [20]. The cloth-changing setting in PRCC means that each identity exists the clothes variation between query and gallery, while the images are all cloth-consistent for each identity in the cloth-unchanging setting. For VC-Clothes, following most works [8], [10], [21], we report the results on the camera 3&4 in cloth-changing setting and results on the camera 2&3 in the cloth-unchanging setting. For LTCC, following [8], [21], [22], the accuracy is calculated only using clothes-changing ground truth samples in cloth-changing setting. For DeepChange, we use true matches from different time and trajectory following [21], [25]. For LaST

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON LONG-TERM CLOTH-CHANGING RE-ID DATASETS

Methods	external knowledge	clothes label	PRCC				VC-Clothes				LTCC	
			Cloth-changing		Cloth-unchanging		Cloth-changing		Cloth-unchanging		Cloth-changing	
			Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
SPT [7]	contour	no	34.3	-	64.2	-	-	-	-	-	-	
ISP [18]	no	no	36.6	-	92.8	-	72.0	72.1	94.5	94.7	27.8	11.9
PCB [2]	no	no	41.8	38.7	99.8	97.0	62.0	62.2	94.7	94.3	23.5	10.0
IANet [42]	no	no	46.3	45.9	99.4	98.3	-	-	-	-	25.0	12.6
RCSANet [11]	extra data	no	50.2	48.6	100.0	97.2	-	-	-	-	-	-
3dPL [10]	pose+contour	yes	51.3	-	-	-	79.9	81.2	-	-	31.2	14.8
BOT+SPA [20]	semantic mask	yes	54.2	63.9	-	-	-	-	-	-	38.7	18.2
FASM [8]	pose+contour	no	54.5	-	98.8	-	78.6	78.9	94.7	94.8	38.5	16.2
CAL [21]	no	yes	55.2	55.8	100.0	99.8	81.4	81.7	95.1	95.3	40.1	18.0
DCR-ReID [22]	mask+contour	yes	57.2	57.4	100.0	99.7	-	-	-	-	41.1	20.4
baseline	no	no	47.4	55.5	99.1	99.2	75.9	77.4	94.5	94.6	35.2	15.1
Ours	**no**	**no**	**57.3**	**65.8**	99.7	**99.8**	**81.8**	**81.7**	94.7	94.9	**43.4**	**18.2**

"Cloth-changing" and "Cloth-unchanging" denote two evaluation protocols illustrated in section IV-A3. "-" Denotes that original paper not reported.

datasets, we follow the original evaluation protocols [21], [26]. For short-term Re-ID datasets Market1501 and Veri-776, we inherit their standard evaluation protocols [38], [39]. For corruption Re-ID, we apply 21 image corruptions [40] including noise, blur, weather and digital to test set of Market1501, which is used to verify the robustness of model against natural corruptions attacks.

B. Implementation Details

We adopt the ResNet50 initialized by ImageNet pre-trained weights as the backbone, and the downsampling of stage4 is removed following [3]. The IMS module is only applied after the end of stage1 and stage2, and the α in the beta distribution $\mathrm{Beta}(\alpha, \alpha)$ is set to 0.2 empirically. Following [12], the resolution of input pedestrian images are resized to 384×192 and the horizontal flipping and random erasing [41] are adopted for data augmentation. The batchsize is set to 64 and each batch consists of 8 identities and 8 instances for each identity for training. The model is trained by the Adam optimizer for 150 epochs on PRCC dataset, for 120 epochs on other datasets. The learning rate linearly increased from 3.5e-5 to 3.5e-4 in the first 10 epochs by warmup strategy and was further decayed 10 times after the 40 and 90 epoch subsequently. The top-$\mathbb{K}$ neighbors in the GEP module is set to 8, the margin parameter m in triplet loss is set to 0.3 and the size of memory bank B is set to 8192 empirically. The λ is set to 0.5 for PRCC and LaST, set to 2.0 for VC-Clothes, LTCC and DeepChange. The IMS module and GEP module are not used during inference stage and we adopt the l_2-normalized feature for distance calculation. For the Veri-776 dataset, the vehicle images are uniformly resized to 256×256.

C. Comparison With State-of-The-Art Methods

Comparison on clothes-changing Re-ID dataset: We first compare our proposed method with state-of-the-arts on five long-term clothes-changing datasets. As shown in the Table I, under the cloth-changing setting, our method consistently outperforms the state-of-the-art methods including CAL [21], FASM [8], 3dPL [10] on the PRCC, LTCC and VC-Clothes. Specifically, the Rank1 accuracy of our method is 57.3% and 43.4% on the PRCC and LTCC datasets, respectively. Our method surpass the CAL method [21] by **2.1**% and **3.3**% Rank1 on PRCC and LTCC, respectively. Compared to the recent DCR-ReID [22] method, our approach achieves similar Rank1 performance on PRCC and outperforms it by 2.3% Rank1 accuracy on LTCC. On the virtual dataset VC-Clothes, our method also outperforms the SOTA method CAL [21]. Furthermore, on the two large-scale long-term datasets DeepChange and LaST that with highly diverse clothes and weather variations, as depicted in Tables II and III, our approach still exceeds the CAL [21] with 1.1% Rank1 improvement on DeepChange, and 1.0% mAP improvement on LaST dataset. Note that our method is simpler and more effective than the compared methods. The CAL [21] relies heavily on the clothes labels. However, the clothes labels are usually time-consuming to collect in practice. The SPT [7] adopts the auxiliary contour information, RCSANet [11] uses additional Re-ID data. The 3dPL [10] and FASM [8] need external human pose+contour knowledge and sophisticated network

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE LARGE-SCALE LONG-TERM CLOTH-CHANGING RE-ID DATASET DEEPCHANGE

Methods	Input modality	**DeepChange** (clothes-changing)			
		Rank1	Rank5	Rank10	mAP
BOT ResNet50 [25]	RGB	47.4	59.4	71.1	12.9
BOT ResNet50 [25]	Grayscale	40.0	54.0	60.1	9.4
BOT ResNet152 [25]	RGB	50.2	62.2	67.8	14.5
BOT DenseNet121 [25]	RGB	47.8	60.4	65.8	13.4
Re-IDCaps [25]	RGB	44.2	56.4	62.0	13.2
ViT B16 [25]	RGB	49.7	61.8	67.3	14.9
CAL [21]	RGB	54.0	-	-	19.0
baseline	RGB	52.3	63.1	67.7	16.9
Ours	RGB	**55.1**	**64.9**	**69.6**	**18.3**

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE LARGEST LONG-TERM PERSON RE-ID DATASET LAST

Methods	**LaST Dataset**			
	mAP	Rank1	Rank5	Rank10
PCB [2]	15.2	50.6	68.0	73.9
MGN [43]	17.6	41.0	63.0	76.0
SFT [44]	19.3	61.2	75.3	79.9
OSNet [45]	21.0	64.3	78.9	82.6
HOReID [46]	25.5	68.3	82.3	86.2
CtF [47]	26.5	70.0	83.3	86.7
mAPLoss [26]	28.0	71.0	84.3	87.7
CAL [21]	28.8	73.7	-	-
baseline	25.7	69.3	83.4	86.7
Ours	**29.8**	**73.2**	**86.2**	**89.6**

TABLE IV
COMPARISON WITH THE REPRESENTATIVE METHODS ON THE SHORT-TERM PERSON RE-ID AND VEHICLE RE-ID DATASETS

Method	Person Re-ID		Method	Vehicle Re-ID	
	Market1501			Veri-776	
	Rank1	mAP		Rank1	mAP
BOT [3]	94.5	85.9	PRN [48]	92.2	70.2
CAL-FGVC [49]	94.5	87.0	PAMTRI [50]	92.9	71.8
FASM [8]	94.6	85.6	DFLNet [51]	93.2	73.3
OSNet [45]	94.8	84.9	ResNet50 [49]	94.5	72.0
DGNet [52]	94.8	86.0	CAL-FGVC [49]	95.4	74.3
ISP [18]	95.3	88.6	UMTS [53]	95.8	75.9
baseline	94.2	85.8	baseline	95.2	77.0
Ours	95.1	87.1	**Ours**	95.8	77.7

design. The recent DCR-ReID [22] method requires both the mask+countor knowledge dependencies and clothes label. In contrast to above methods, it is important to emphasize that our approach does not rely on any external knowledge dependencies or clothes-labels, and the scalability of our approach is better, achieving superior performance. These discussions highlight the advantages of our approach to cope with the clothes-changing person Re-ID. In addition, as shown in the Table I, compared with these methods, our method also shows competitive performance under cloth-unchanging setting, which illustrates the effectiveness of our method in the cloth-unchanging scenario.

Comparison on short-term Re-ID datasets: We next evaluated the universality of our approach on the short-term person Re-ID dataset Market1501 and vehicle Re-ID dataset VeRi-776. As shown in the Table IV, our method achieves clear performance gains on the baseline, and yields comparable performance compared with short-term Re-ID methods DGNet [52], ISP [18] and clothes-changing Re-ID method FASM [8]. It is worth mentioning that the methods 3dPL [10], FASM [8] and BOT+SPA [20] rely heavily on priori human structure and cannot be applicable for vehicle Re-ID. In constrast to these clothes-changing methods, as shown in Table IV, our method can also be adapted to vehicle Re-ID and achieves promising performance compared with current vehicle Re-ID methods, which validates the universality of our approach.

D. Ablation Studies

We adopt the widely used strong-baseline [3] in person Re-ID as our baseline, which employs the ResNet50 as the backbone and utilizes both the identity loss and triplet loss as the loss functions. The extensive ablation and comparison experiments are conducted on top of this baseline to demonstrate the effectiveness of our proposed method. As shown in the last 2 lines of the Tables I and II, our holistic method brings significant performance improvements to baseline across various clothes-changing ReID datasets and settings. For example, under cloth-changing setting, our method significantly surpasses the baseline model by **9.9%** and **8.2%** Rank1 on PRCC and LTCC, respectively, which suggests that our method can effectively handle clothes-changing person Re-ID.

The effectiveness of IMS module: The IMS module is designed to synthesize more novel and diverse styles during training, we first verify its effectiveness for robust model training against the clothing shift. As shown in the Table V, despite simple, the IMS module brings significant improvement of **7.2%** Rank1 to baseline for PRCC under cloth-changing setting, which highlights the effectiveness of our IMS module for overcoming the clothing shift challenge.

TABLE V
ABLATION STUDY ABOUT THE EFFECTIVENESS OF THE PROPOSED IMS AND GEP MODULES ON THE PRCC DATASET

Methods	PRCC			
	Cloth-changing		Cloth-unchanging	
	Rank1	mAP	Rank1	mAP
baseline	47.4	55.5	99.1	99.2
baseline + IMS	54.6	63.2	99.5	99.6
baseline + GEP	53.4	62.3	99.4	99.3
baseline + IMS + GEP	**57.3**	**65.8**	**99.7**	**99.8**

TABLE VI
COMPARISON EXPERIMENTS OF THE IDENTITY-AWARE MIXSTYLE "IMS" MODULE. THE "IMU" DENOTES THE IDENTITY-AWARE MIXUP

Methods	PRCC			
	Cloth-changing		Cloth-unchanging	
	Rank1	mAP	Rank1	mAP
baseline	47.4	55.5	99.1	99.2
baseline + IMU	49.5	58.3	99.2	99.3
baseline + Mixstyle	51.8	60.1	99.2	99.4
baseline + IMS	**54.6**	**63.2**	**99.5**	**99.6**

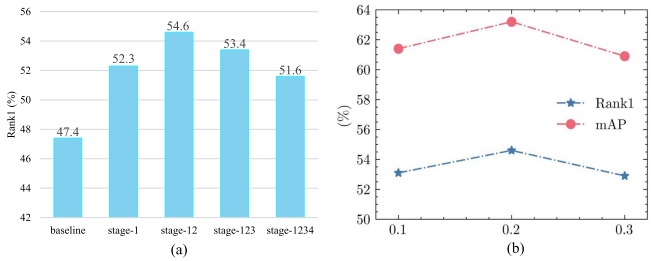

Fig. 3. (a) Analysis about where to insert the IMS module in the ResNet backbone. (b) The hyper-parameters analysis about the α which controls the Beta distribution in IMS module. (Under cloth-changing setting of PRCC) .

Furthermore, we compare the proposed IMS module with several counterpart methods on the baseline including (1) *original Mixstyle* (2) *identity-aware mixup* (IMU). The IMU method interpolates raw image space of paired images by mixup [34]. As shown in the Table VI, although the IMU and original Mixstyle [16] bring gains, the performance improvement of IMS module significantly surpasses them. Compared with original Mixstyle, the IMS module mixes instance-level feature statistics of two samples belonging to the same identity, which avoids the interference from other identities during mixture and efficiently retains the correspondence between synthesized samples and latent label space, hence achieves superior performance to original mixstyle. Compared with IMU, the IMS directly manipulates more powerful style statistics in shallow layers hierarchically, which is more interpretable and effective to synthesize diverse styles for training.

Where to insert the IMS module: Given a standard ResNet50 with 4 stages, we also analyzed where the IMS module should be inserted. The analyses are conducted on PRCC dataset under cloth-changing setting and we report the Rank1 accuracy. Some observations can be drawn from the Fig. 3(a). (1) Inserting the IMS module into the multiple stages can consistently bring significant performance gains to the baseline. (2) When applying the IMS module at the shallow layers of network (stage-12), the

TABLE VII
ABLATION EXPERIMENTS ABOUT THE "GEP" MODULE ON THE CLOTH-CHANGING SETTING OF PRCC AND LTCC DATASETS

prototype	batch	memory	graph	PRCC		LTCC	
				Rank1	mAP	Rank1	mAP
case-0	–	–	–	47.4	55.5	35.2	15.1
case-1	√	–	–	48.7	58.2	36.0	15.8
case-2	–	√	–	51.5	60.2	37.3	16.0
case-3	–	√	√	**53.4**	**62.3**	**40.5**	**17.4**

performance is the best. This is because the style information of pedestrian image is mainly stored in the feature statistics of feature maps in shallow layers.

The effectiveness of GEP module: As shown in the Table V, the GEP module brings remarkable performance gain by **+6.0%** Rank1 to the baseline on PRCC under cloth-changing setting, which indicates that GEP could effectively eliminate the feature discrepancy caused by clothing-shift.

Next, we conduct ablations to validate the necessity and significance of the GEP module on the PRCC and LTCC datasets. As shown in the Table VII, the case-1 (center loss) uses mini-batch to calculate the class centroid. The case-2 adopts a larger memory bank which records the past and current samples to estimate the identity prototype. The case-3 (i.e., GEP module) explores the underlying graph similarity relation of samples across memory bank to estimate prototype as constraint. As we can see, with the progressive incorporation of memory bank and KNN-graph modeling, the performance has been gradually improved. For clothes-changing person Re-ID, the center loss using naive class centroid of mini-batch is insufficient for effective metric learning (case-1).

In contrast, the GEP (case-3) module benefits from the larger scale of samples and the KNN-graph enhancement between style-diversified samples within the dynamic memory bank. The prototype constructed by the GEP module can capture more intrinsic characteristics for each identity, making it more robust to clothing shifts and serving as an exemplar for desirable metric learning. Although there may be a small number of negative samples among the K neighbors, we observed that most of samples in the top-K neighbors of the anchor are positive samples. In addition, the attention-based feature aggregation and the selection of proper $\mathbb{K}$ value also mitigate the effect of hard negative samples. By comparing the case-2 and case-3 in the Table VII, we experimentally verify the necessity of KNN-graph enhancement for GEP module, as it resulted in performance improvements of **+1.9%** and **+3.2%** Rank1 on PRCC and LTCC datasets, respectively.

The complementary of the IMS and GEP module: As shown in Table V, compared to using a single module alone, the combination of the identity-aware mixstyle augmentation and graph-enhanced prototype constraint brings higher performance, which suggests that the two modules are complementary and collaborate to learn clothes-irrelevant features.

E. The Further Corruption Robustness Analysis

The robustness analysis against corruptions: In addition to clothing shift, the person Re-ID in open-world faces various natural corruptions such as noise, blur and weather. The experiments are further conducted on the corruption-invariant Re-ID benchmark to verify the robustness of our method against corruptions. Following [40], we applied 21 types of corruptions to the testset of Market1501 and there are three evaluation settings: (1) corrupted query, (2) corrupted gallery, and (3) corrupted query and gallery. Following [40], the random erasing (RE) is removed during training for both baseline and our method in the corruption analysis experiments, since it hurts model corruption robustness. We can draw some observations from the Fig. 4. Firstly, the image corruptions significantly deteriorate the performance of the well-trained model. Secondly, our method significantly improves model robustness against image corruptions. For example, although our method only exceeds the baseline by 0.5% Rank-1 under Clean Eval. mode, the **6%** Rank1 gain is obtained by our method against baseline on the most challenging Corrupted Eval. mode. Thirdly, our model consistently outperforms the baseline under various corruption modes. These suggests that our method can effectively promote the model robustness against the covariate shift arising from image corruptions.

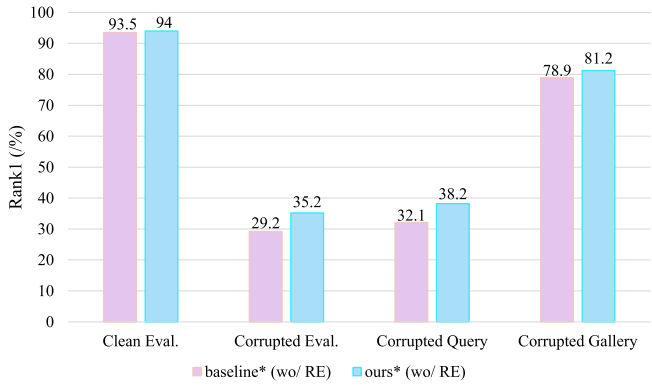

Fig. 4. Corruption robustness of our method compared to baseline with different modes on Market1501. Corrupted Eval. means that both query and gallery images are corrupted, while Clean Eval. denotes no corruption.

F. Hyper-Parameters Analysis

We analyzed the influences of some key hyper-parameters on PRCC dataset, including α, $\mathbb{K}$ and λ. Specifically, α determines the shape of Beta distribution and influences the random convex weight β in IMS module. We evaluated the α on PRCC and the α is set from {0.1,0.2,0.3}. As shown in the Fig. 3(b), the performance reaches peak when the $\alpha = 0.2$ and the IMS module consistently improves the performance when the α changes from 0.1 to 0.3. The $\mathbb{K}$ determines the number of nearest neighbors involved in the GEP module. As shown in the Fig. 5(a), the performance rises as the $\mathbb{K}$ initially increases and then slowly decreases, and the performance arrives peak when $\mathbb{K} = 8$ on PRCC. When the $\mathbb{K}$ is small, the relation knowledge of helpful neighbors is not sufficient, while the large $\mathbb{K}$ may introduce some noisy neighbors in the early training epochs. The λ controls the weight of GEP module driven by $\mathcal{L}_{\text{gep}}(\mathcal{V}, \mathcal{C})$. As depicted in Fig. 5(b), the performance achieves best when $\lambda = 0.5$ and different λ in a reasonable range consistently boosts performance.

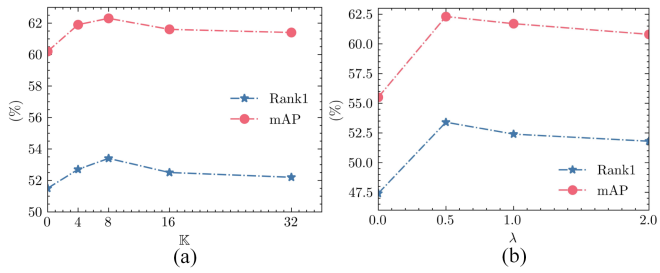

Fig. 5. Key hyper-parameters analysis of GEP module on the PRCC, including. (a) The analysis of the $\mathbb{K}$ which determines the number of neighbors, (b) The analysis of the λ which controls the weight of $\mathcal{L}_{\text{gep}}$ loss.

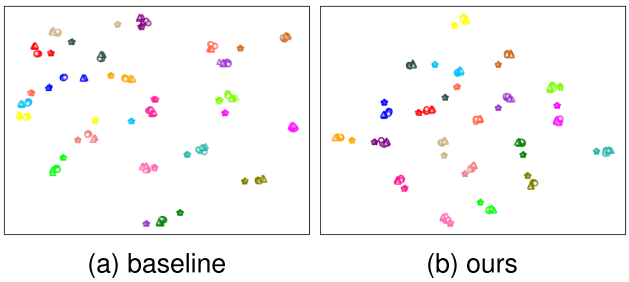

Fig. 6. t-SNE visualization of feature embedding from 20 random identities on PRCC testset. In camera A (marked by circle) and camera B (marked by triangle), the clothes of same identity are consistent, In camera C (marked by pentagram), the clothes are different from that of camera A/B for the same identity. Different identities are indicated by different colors.

G. Visualization Analysis

Feature distributions: We first visualize the feature embedding distributions of baseline and our method on PRCC by t-SNE [54]. In Fig. 6(a) for baseline, we can observe that for each identity, although the same-clothes samples exhibit the intra-class compactness, the cross-clothes samples of some identities still present considerable intra-class dispersion. For comparison, as shown in Fig. 6(b), our model displays good intra-class compactness, whether cross-clothes or the same clothes. These suggest that our method can effectively decrease the cross-clothes feature discrepancy.

Feature activation map: We next compare the activation of feature map between baseline and our method. As depicted in the Fig. 7, given some pedestrian images, the activation maps of them by the baseline model concentrate on several narrow local discriminative regions, such as face, shoes, and shoulders. In contrast, the activation maps obtained by our model are more broad, diverse and clothing-irrelevant, and the attention region includes global body shape and the limbs, which reduces the risk of overfitting to local regions and hence the model can be more robust to cloth-changing.

Retrieval results: Finally, we visualize some ranking lists by baseline and our method under cloth-changing setting. It can be seen from Fig. 8 that when facing large clothing variations and similar pedestrians, the baseline model underperformed. The retrieval results of baseline tend to overfit to the clothing color and texture (see the 2nd line). For comparison, our model effectively retrieves the correct cross-clothes ground truth, even suffering significant clothing changes and distraction from similar pedestrians (see the 3rd and 4th lines).

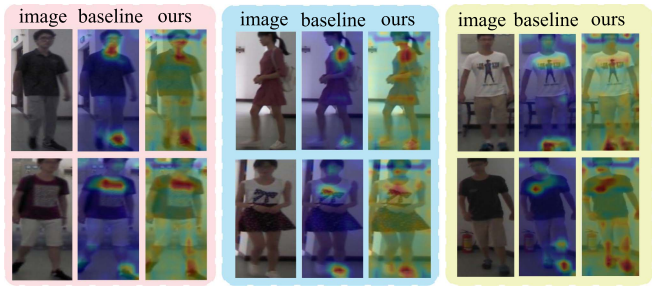

Fig. 7. Visualization of feature activation maps for some images on PRCC. In each triplet, the first column indicates the original image, the second column corresponds to the baseline model, and the third column corresponds to activation map of our model. In each grid (color), the original images share the same identity but with different clothes.

Fig. 8. Comparison of ranking list between baseline and our method under cloth-changing setting. The query images are selected from realistic PRCC and virtual VC-Clothes testset. Green box means the correct match, while the red box means the wrong match. Note that the evaluation of PRCC is single-shot, which indicates that there is only one ground truth in the gallery.

V. Conclusion

This article mainly studies the more realistic long-term cloth-changing person Re-ID task. To efficiently and effectively strengthen the clothes-irrelevant feature mining, we present a novel joint Identity-aware Mixstyle and Graph-enhanced Prototype framework for clothes-changing person Re-ID. By considering the cloth-changing as fine-grained domain/style shift, from the perspective of domain generalization, the identity-aware mixstyle is proposed to mix the instance-level feature statistics of samples within each identity to enrich the clothing style diversity for shift-robust model training. The graph-enhanced prototype constraint further explores the underlying graph similarity structure of samples across memory to build informative and robust prototypes, which serve as exemplars for better

clothing-irrelevant metric learning. The extensive experiments conducted on many clothes-changing person Re-ID datasets validated the effectiveness and advantage of our method. Further analyses show the good universality and corruption robustness of our method.

REFERENCES

[1] M. Ye et al., "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[2] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.

[3] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1487–1495.

[4] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *CoRR*, pp. 1–17, 2017, *arXiv:1703.07737*.

[5] F. Liu and L. Zhang, "View confusion feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6639–6648.

[6] C. Zhao et al., "Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3180–3195, Dec. 2020.

[7] Q. Yang, A. Wu, and W.-S. Zheng, "Person re-identification by contour sketch under moderate clothing change," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2029–2046, Jun. 2021.

[8] P. Hong, T. Wu, A. Wu, X. Han, and W.-S. Zheng, "Fine-grained shape-appearance mutual learning for cloth-changing person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10513–10522.

[9] J. Chen et al., "Deep shape-aware person re-identification for overcoming moderate clothing changes," *IEEE Trans. Multimedia*, vol. 24, pp. 4285–4300, 2022.

[10] J. Chen et al., "Learning 3 D shape feature for texture-insensitive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8146–8155.

[11] Y. Huang, Q. Wu, J. Xu, Y. Zhong, and Z. Zhang, "Clothing status awareness for long-term person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11895–11904.

[12] X. Qian et al., "Long-term cloth-changing person re-identification," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 71–88.

[13] C. Eom, W. Lee, G. Lee, and B. Ham, "IS-GAN: Learning disentangled representation for robust person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, pp. 8975–8991, 2021.

[14] W. Xu et al., "Adversarial feature disentanglement for long-term person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 1201–1207.

[15] Y.-J. Li, X. Weng, and K. M. Kitani, "Learning shape representations for person re-identification under clothing change," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 2432–2441.

[16] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–15.

[17] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global–local-alignment descriptor for scalable person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, Apr. 2019.

[18] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 346–363.

[19] Z. Zhuang et al., "Rethinking the distribution gap of person re-identification with camera-based batch normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 140–157.

[20] X. Jia et al., "Patching your clothes: Semantic-aware learning for cloth-changed person re-identification," in *Proc. Int. Conf. Multimedia Model.*, 2022, pp. 121–133.

[21] X. Gu et al., "Clothes-changing person re-identification with RGB modality only," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1060–1069.

[22] Z. Cui, J. Zhou, Y. Peng, S. Zhang, and Y. Wang, "DCR-ReID: Deep component reconstruction for cloth-changing person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4415–4428, Aug. 2023.

[23] F. Wan, Y. Wu, X. Qian, Y. Chen, and Y. Fu, "When person re-identification meets changing clothes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 830–831.

[24] S. Yu et al., "COCAS: A large-scale clothes changing person dataset for re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3400–3409.

[25] P. Xu and X. Zhu, "DeepChange: A long-term person re-identification benchmark," *CoRR*, pp. 1–13, 2021, *arXiv:2105.14685*.

[26] X. Shu et al., "Large-scale spatio-temporal person re-identification: Algorithms and benchmark," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4390–4403, Jul. 2022.

[27] J. Wang et al., "Generalizing to unseen domains: A survey on domain generalization," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 8, 2023, pp. 8052–8072, doi: 10.1109/TKDE.2022.3178128.

[28] X. Yue et al., "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2100–2110.

[29] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5400–5409.

[30] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Feature alignment and restoration for domain generalization and adaptation," *CoRR*, pp. 1–15, 2020, *arXiv:2006.12009*.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[32] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1501–1510.

[33] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3143–3152.

[34] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, pp. 1–13, 2018.

[35] X. Wang, H. Zhang, W. Huang, and M. R. Scott, "Cross-batch memory for embedding learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6388–6397.

[36] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[37] S. Li, D. Chen, B. Liu, N. Yu, and R. Zhao, "Memory-based neighbourhood embedding for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6102–6111.

[38] L. Zheng et al., "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

[39] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2016, pp. 1–6.

[40] M. Chen, Z. Wang, and F. Zheng, "Benchmarks for corruption invariant person re-identification," in *Proc. Neural Inf. Process. Syst. Datasets Benchmarks Track.*, 2021, pp. 1–18.

[41] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13001–13008.

[42] R. Hou et al., "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9317–9326.

[43] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 274–282.

[44] C. Luo, Y. Chen, N. Wang, and Z. Zhang, "Spectral feature transformation for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4976–4985.

[45] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3702–3712.

[46] G. Wang et al., "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6449–6458.

[47] G. Wang, S. Gong, J. Cheng, and Z. Hou, "Faster person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 275–292.

[48] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3997–4005.

[49] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1025–1034.
[50] Z. Tang et al., "Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 211–220.
[51] Y. Bai et al., "Disentangled feature learning network for vehicle re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 474–480.
[52] Z. Zheng et al., "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2138–2147.
[53] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11165–11172.
[54] L. V. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008, pp. 1–27.

Zhiwei Zhao received the B.S. degree from the Hefei University of Technology, Hefei, China, in 2019. He is currently working toward the Ph.D. degree with the School of Cyber Science and Technology, University of Science and Technology of China, Hefei. His current research interests include person re-identification, multi-modal learning, and computer vision.

Bin Liu (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1998 and 2001, respectively, and the Ph.D. degree in electrical engineering from Syracuse University, Syracuse, NY, USA, in 2006. He is currently an Associate Professor with the School of Cyber Science and Technology, University of Science and Technology of China. His research interests include computer vision and AI security.

Yan Lu received the B.Eng. degree from Anhui Polytechnic University, Wuhu, China, and the M.S. degree from the University of Science and Technology of China, Hefei, China, in 2020. He is currently with the Shanghai AI Laboratory as a Research Intern. His current research interests include computer vision, machine learning, and AI for Science.

Qi Chu (Member, IEEE) received the B.S. degree in electronic engineering and the Ph.D. degree in information and communication engineering from the University of Science and Technology of China, Hebei, China, in 2014 and 2019, respectively. He is currently an Associate Research Fellow with the School of Cyber Science and Technology, University of Science and Technology of China. His research interests include computer vision, deep learning, and AI security.

Nenghai Yu received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2004. He is currently a Full Professor with the University of Science and Technology of China. He was a Visiting Scholar with the Institute of Production Technology, Faculty of Engineering, University of Tokyo, Tokyo, Japan, in 1999, and did cooperative research as a Senior Visiting Scholar with the Department of Electrical Engineering, Columbia University, New York, NY, USA, from April to October 2008. His research interests include image processing and video analysis, computer vision, multimedia communication, media content security, Internet information retrieval, data mining and content filtering, network communication and security.

Chang Wen Chen (Fellow, IEEE) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 1983, the M.S.E.E. degree from the University of Southern California, Los Angeles, CA, USA, in 1986, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1992. He is currently a Chair Professor of visual computing with The Hong Kong Polytechnic University, Hong Kong. He was an Empire Innovation Professor of computer science and engineering with the University at Buffalo, State University of New York, Buffalo, NY, USA, from 2008 to 2021. He also was the Dean of the School of Science and Engineering, Chinese University of Hong Kong, Shenzhen, China, from 2017 to 2020. He was an Allen Henry Endow Chair Professor with the Florida Institute of Technology, Melbourne, FL, USA, from 2003 to 2007. He was on the Faculty of electrical and computer engineering with the University of Missouri-Columbia, Columbia, MO, USA, from 1996 to 2003 and on the Faculty of Electrical and computer engineering with the University of Rochester, Rochester, NY, USA, from 1992 to 1996. He was the Editor-In-Chief of the IEEE TRANSACTIONS ON MULTIMEDIA from January 2014 to December 2016. He was also the Editor-in-Chief for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from January 2006 to December 2009.