This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3001665, IEEE Transactions on Information Forensics and Security

Visible-Infrared Person Re-Identification via Homogeneous Augmented Tri-Modal Learning

Mang Ye, Jianbing Shen, *Senior Member, IEEE* and Ling Shao

***Abstract*—Matching person images between the daytime visible modality and night-time infrared modality (VI-ReID) is a challenging cross-modality pedestrian retrieval problem. Existing methods usually learn the multi-modality features in raw image space, ignoring the image-level discrepancy. Some methods apply GAN technique to generate the cross-modality images, but it destroys the local structure and introduces unavoidable noise. In this paper, we propose a Homogeneous Augmented Tri-Modal (HAT) learning method for VI-ReID, where an auxiliary grayscale modality is generated from their homogeneous visible images, without additional training process. It preserves the structure information of visible images and approximates the image style of infrared modality. Learning with the grayscale visible images enforces the network to mine structure relations across multiple modalities, making it robust to color variations. Specifically, we solve the tri-modal feature learning from both multi-modal classification and multi-view retrieval perspectives. For multi-modal classification, we learn a multi-modality sharing identity classifier with a parameter-sharing network, trained with a homogeneous and heterogeneous identification loss. For multi-view retrieval, we develop a weighted tri-directional ranking loss to optimize the relative distance across multiple modalities. Incorporated with two invariant regularizers, HAT simultaneously minimizes multiple modality variations. In-depth analysis demonstrates the homogeneous grayscale augmentation significantly improves the VI-ReID performance, outperforming the current state-of-the-art by a large margin. It provides a simple but effective strategy for future research in this task.**

I. Introduction

Person re-identification (Re-ID) has been widely studied as a specific pedestrian retrieval problem by retrieving person images across multiple non-overlapping surveillance cameras [1], [2]. With the advancement of deep neural networks, inspiring performance has been obtained in both image-based [3]–[5] and video-based person Re-ID tasks [6]–[8], where all the person images/videos are usually captured by visible cameras in the daytime [9]. However, visible cameras cannot capture enough discriminative information under poor lighting conditions, *e.g.*, at night [10]. This property limits the applicability of the single-modality person Re-ID for real surveillance scenarios. In contrast, this paper studies the cross-modality visible infrared person re-identification (VI-ReID) problem. This task aims at matching the daytime person images (captured by visible cameras) and nighttime infrared person images (either infrared [10] or thermal [11] cameras), which is imperative for night-time surveillance applications.

Mang Ye, Jianbing Shen and Ling Shao are with the Inception Institute of Artificial Intelligence, UAE. E-mail: mangye16@gmail.com, shenjianbingcg@gmail.com, ling.shao@inceptioniai.org
Corresponding Author: Jianbing Shen

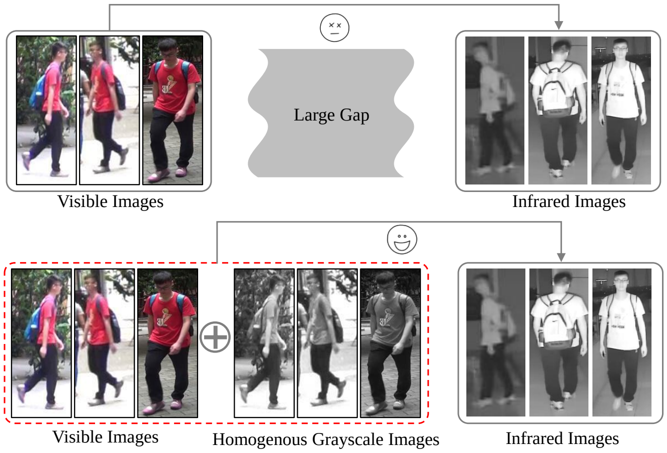
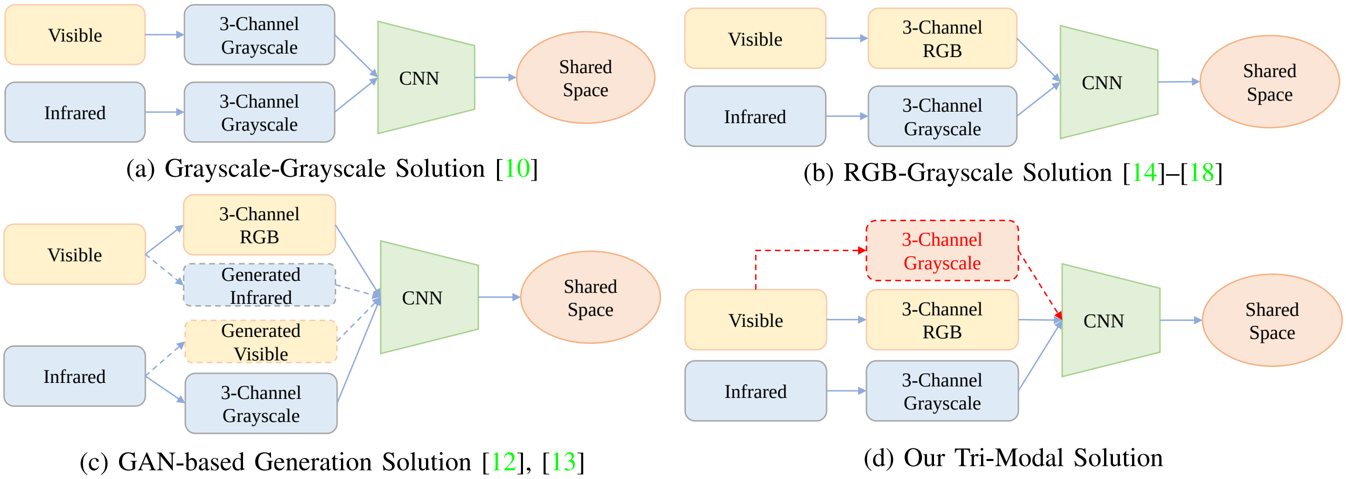


Fig. 1. Illustration of VI-ReID. There is a large domain gap between the visible and infrared images. This paper proposes to utilize the homogeneously augmented grayscale images to enhance the robustness against color variations. These augmented images preserves the structure information of the visible images.

The person images in VI-ReID are captured by cameras with different light spectrums under long time stamp. Different wavelengths result in large visual difference between the collected images from two modalities. Meanwhile, different viewpoints, poses variations and self-occlusions add additional difficulties. Thus, it is an extremely challenging cross-modality pedestrian retrieval problem with large domain gap and unconstrained variations, as shown in Fig. 1. The paradigm of VI-ReID is closely related to the widely-studied VIS-NIR face recognition [19]–[21], *i.e.,* matching between visible and near-infrared images. However, the visual variations of the person images are usually much larger than that of face images, which makes their methods incapable for VI-ReID task [10].

Bridging the modality gap is critical for modality invariant feature learning in VI-ReID, existing methods fall into three categories (Fig. 2): a) *Grayscale-grayscale Solution*: Since the infrared images do not contain the color information, all the person images are transformed into single-modality grayscale images to eliminate the color effects [10]. Wu *et al.* proposed a feature learning framework with a deep zero-padding network [10]. This strategy ignores the important color information in visible images. b) *RGB-grayscale Solution*: This approach directly learns the multi-modality representation using the original color channels in visible and infrared images, either by adversarial training [15] or feature alignment [14], [17]. They prove that the color information is beneficial for the VI-ReID. However, they ignore the large pixel-level gap between the 3-channel visible and 1-channel infrared images [12], *i.e.,* the



(a) Grayscale-Grayscale Solution [10]

(b) RGB-Grayscale Solution [14]–[18]

(c) GAN-based Generation Solution [12], [13]

(d) Our Tri-Modal Solution

Fig. 2. Illustration of four different solutions for VI-ReID. (a) *Grayscale-Grayscale Solution* [10]: All input images from both modalities are transformed into single-channel grayscale images. (b) *RGB-Grayscale Solution* [14]–[18]: All input images are kept with the original channels. (c) *GAN-based Generation Solution*: These methods generate cross-modality images before being fed into the network. (d) *Our Tri-Modal Solution*: The model is trained with two original (*Visible and Infrared*) modalities, and one homogeneous augmented grayscale images. All the single-channel (grayscale) images are expanded three times in channel dimension before feeding into the network for dimension consistency.

distributions of the pixel values are totally different. c) *GAN-based Generation Solution*: Two recent methods [12], [13] propose to generate the cross-modality images to eliminate the modality gap in pixel level. Nevertheless, training the image generator requires expensive computational cost and the important structural information for VI-ReID is easily destroyed by the unavoidable noise.

To address the above limitations, this paper presents a novel Homogeneous Augmented Tri-modal (HAT) learning solution, where an auxiliary grasycale modality is generated from the homogeneous visible images, as shown in Fig. 2 (d). These generated grayscale images preserve structure information of the visible images and approximate the infrared image style. The advantage of using homogeneous augmented grayscale images can be summarized as two folds: 1) The grayscale images preserve the structure information of visible images, and the structure information is crucial for cross-modality matching since the infrared images do not contain any color information. To approximate the image style of cross-modality images, existing GAN-based methods try to learn visible to infrared image generator, but the image generation process might introduce additional noise. In contrast, our homogeneous generation does not introduce any noise, and performs much better than existing methods. 2) Minimizing the modality gap between the grayscale visible images and infrared images can enforce the network to mine the structure relations between these two modalities, thereby making it robust to color variations and improving the visible-infrared matching performance. We solve the tri-modal feature learning from both multi-modal classification and multi-view retrieval perspectives. For multi-modal classification, a homogeneous and heterogeneous identification loss is designed by learning a multi-modality sharing identity classifier with a parameter-sharing network, achieving identity-invariant representation with sharable information mining. For multi-view retrieval, a weighted tri-directional ranking loss is developed to optimize the relative distance across multiple modalities. To further enhance the robustness, we propose two regularizers, enforcing the augmentation and cross-modality positive pair invariance. With in-depth analysis, we demonstrate that the homogeneous grayscale augmentation greatly improves the performance.

Our main contributions are summarized as follows:

- We propose a novel Homogeneous Augmented Tri-Modal (HAT) learning method for VI-ReID, which solves the problem from a new perspective with homogeneous grayscale augmentation to improve the performance.
- We introduce a Homo- and Heterogeneous Identification (HHI) loss for multi-modality classification with identity supervision. It enhances the discriminability with a homogeneous invariant regularizer.
- We develop a Weighted Tri-directional Ranking (WTDR) loss for multi-view retrieval, explicitly optimizing the cross-modality correlations across multiple modalities to reduce the modality variations.
- We present an in-depth analysis to validate the effectiveness of grayscale augmentation, and it also provides an effective increment for future VI-ReID research. We outperform the current state-of-the-arts by a large margin on two public cross-modality VI-ReID datasets, which sets a new baseline in this field.

II. Related Work

Single-Modality Person Re-ID. Person re-identification (a.k.a, pedestrian retrieval [22]) addresses the problem of retrieving person images across different video surveillance cameras [23], [24]. The primary challenges are large cross-camera variations for each identity, caused by different camera environments, poses changes and viewpoint variations [25]–[31]. To address these issues, existing methods either focus on extracting robust feature representations [32]–[34] or learning discriminative distance metrics [35]–[38]. With the advancement of deep neural networks, existing methods have achieved inspiring performance in single-modality person Re-ID, outperforming the human-level retrieval performance on the widely-used benchmarks [26], [39], [40]. However, these

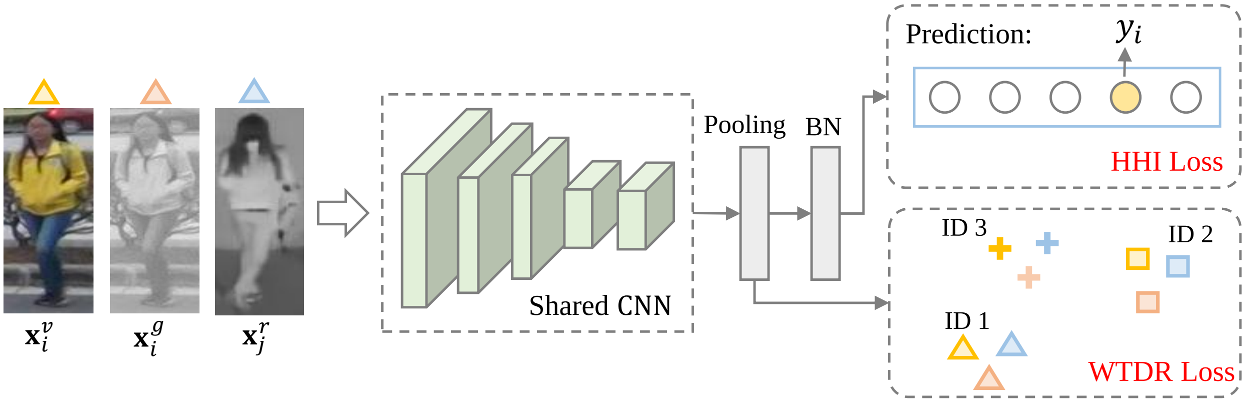


Fig. 3. The framework of the proposed HAT. Images from three modalities (visible, grayscale and infrared) are fed into the weights-sharing CNN network to learn multi-modality sharable feature representation. We solve the tri-modal feature learning from both multi-modal classification (HHI: homogeneous and heterogeneous identification loss) and multi-view retrieval (WTDR: weighted tri-directional ranking loss) perspectives. The former learns identity-invariant features while the latter optimizes the relationship among images from triple modalities.

methods are usually designed for single visible modality, *i.e.*, the person images are captured by general RGB cameras in the daytime [41]. In contrast, the nighttime scenarios with cross-modality images usually suffer from large modality discrepancy. Thus, the methods designed for single-modality person Re-ID are usually incapable of this task [10].

Cross-Modality Person Re-ID. It addresses person Re-ID by matching person images across different types, such as between visible and infrared images [10], [12], [42]. Some works also investigate the cross-modality matching between the images and non-visual text descriptions [43], [44].

For visible-infrared ReID (VI-ReID), a deep zero-padding scheme is proposed in [10] to adaptively capture the modality-specific information in a one-stream network, where the cross-modality representation is learned with the identity supervision. All the input images are transformed into single-grayscale images, this strategy avoids the interruption of the color information. Ye *et al.* [14] designed a two-stream network to learning the multi-modality sharable features, simultaneously handling the cross-modality and intra-modality variations with dual-constrained top-ranking loss. Dai *et al.* [15] introduced an adversarial learning framework to jointly discriminate the identity and modality with an improved triplet loss function. A hypersphere embedding with sharable feature learning was developed in [17]. In addition, some methods also leveraged the modality-specific classifiers to improve the feature learning [16], [18]. These methods achieve good performance by utilizing the original 3-channel color information in the visible images, which prove that the color information still improves the VI-ReID performance.

Two recent papers also investigated the GAN-generated images to bridge the gap between visible-infrared modalities [12], [13]. The reduce the modality gap in both feature-level and pixel-level, achieving state-of-the-art performance. These methods proved the effectiveness of pixel-level generation for cross-modality Re-ID. However, training a reliable image generator requires intensive computational cost, and the generated images usually contain unavoidable noise. In contrast, our solution is quite efficient, and we do not need any additional training for the image generation.

VI-ReID is also closely related to visible near-infrared face recognition [19], [20], [45]–[48]. For face recognition, multi-modal sharable features learning [20] or disentangled representations learning are the popular approaches [21]. Some works also studied the cross-modality matching models [49]. Compared to person Re-ID tasks, the visual difference is much smaller in this face recognition task, which greatly limited the applicability of their methods for VI-ReID task [42].

Image Generation in Person Re-ID. Generating person images has been widely explored in the literature, addressing the pose variations [50]–[52], cross-domain/cross-camera variations [53]–[55], or enriching the supervision [56], [57]. Generally, all of these methods adopt the generative adversarial networks for image generator training, which introduces additional pixel noise. In contrast, our generation can be efficiently obtained by using the linear accumulation of three RGB channels, and it preserves the structure information.

III. Proposed Method

This paper addresses the VI-ReID by leveraging an additional homogeneous augmented grayscale modality, which is obtained by a simple linear accumulation of three RGB color channels from visible images. An overview of the proposed model is shown in Fig. 3. We will firstly introduce the weights-sharing tri-modal feature learning in § III-A, and then the homo- and heterogeneous multi-modal classification and the weighted tri-directional ranking loss are presented in § III-B and § III-C, respectively. Finally, a joint learning framework with identity sample training strategy is provided in § III-D.

A. Parameter-sharing Tri-modal Feature Learning

This subsection presents our weights-sharing feature learning network for three different modalities, including the original visible and infrared modalities, and one generated grayscale modality. The learning target of VI-ReID is that the person images of the same person identity under different modalities are invariant. We denote the original cross-modality training set as $\mathcal{T} = \{\mathcal{T}^v, \mathcal{T}^r\}$. In particular, $\mathcal{T}^v = \{\mathbf{x}_i^v | i = 1, 2, \cdots, N^v\}$ represents visible training set with N^v visible

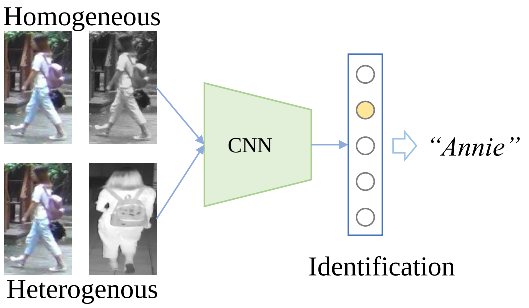


Fig. 4. Illustration of the modality-sharing classifier in Homogeneous and Heterogeneous Identification (HHI) loss. The homogeneous (visible-grayscale) and heterogeneous (visible-infrared) pair can be both correctly classified with the weights-sharing identity classifiers.

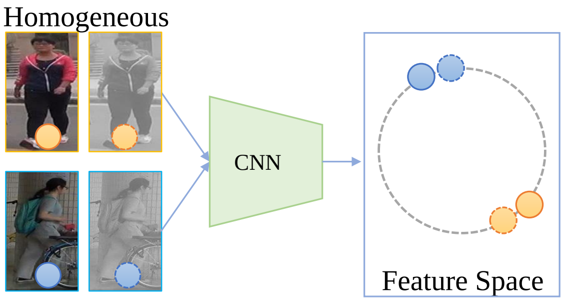


Fig. 5. Illustration of the homogeneous invariant regularizer in Homogeneous and Heterogeneous Identification (HHI) loss, *i.e.*, the features of the original visible image and its homogeneously augmented grayscale image are close in the learned feature space.

images, where each visible image $\mathbf{x}_i^v$ is associated with identity label y_i. $\mathcal{T}^r = \{\mathbf{x}_i^r | i = 1, 2, \cdots, N^r\}$ represents N^r infrared training images, where each element $\mathbf{x}_i^r$ is an infrared image. Due to the light spectrums in different cameras, $\mathbf{x}_i^v$ usually has three different color channels, *i.e.*, R, G and B, and infrared image $\mathbf{x}_i^r$ has single over-saturated grayscale channel.

Homogeneous Grayscale Modality Generation. According to the imagery capturing characteristic, there is a large gap between the three color channel visible image and single-channel infrared image. To address this issue, each visible images is homogeneously transformed into single-channel gray-scale image in to approximate the style of single-channel infrared image. For each visible image $\mathbf{x}_i^v$, its homogeneous grayscale modality image $\mathbf{x}_i^g$ is generated by

$$L = g(R, G, B) \tag{1}$$

where $g(\cdot)$ is a grayscale transform function to perform the pixel-level accumulation of the original R, G and B channels. This generation can be efficiently performed in the deep learning platforms (*e.g.,* Pytorch, Tensorflow). The generated grayscale images provide self-supervision, and it is much more efficient than the GAN-based methods [12], [13], which needs complicated image reconstruction. Moreover, these GAN-based methods destroyed the original spatial structure information during the reconstruction process, introducing additional pixel noise. In contrast, our point-to-point pixel-wise transformation keeps the original structural information. Empirically, we show that these generated homogeneous grayscale images significantly improve the cross-modality Re-ID performance.

Parameter-sharing Feature Learning. To learn the multi-modality invariant features, we propose to apply a single stream weights-sharing network $f(\cdot)$ for three different modalities. In this manner, three different modalities are jointly projected in a shared common feature space. Specifically, the BNneck [58] designed for single-modality person Re-ID is adopted as the backbone. The output of the pooling layer is adopted as the feature representation and a batch normalization layer is added before the final classification layer. The feature representations for $\mathbf{x}_i^v$, $\mathbf{x}_i^g$ and $\mathbf{x}_i^r$ are denoted by

$$\mathbf{f}_i^v = f(\mathbf{x}_i^v), \ \mathbf{f}_i^g = f(\mathbf{x}_i^g), \ \mathbf{f}_i^r = f(\mathbf{x}_i^r). \tag{2}$$

The learning objective of VI-ReID is to narrow the gap between $\mathbf{f}_i^v$ and $\mathbf{f}_j^r$ when $y_i = y_j$. The generated grayscale modality acts as a bridge between the visible and infrared modality. Grayscale modality is directly generated by the linear accumulation of three color channels from the visible images, thus it preserves important structure information of the original visible image, while ignoring the color information. We assume that the modality-specific information can be automatically captured in the weights-sharing one-stream framework with abundant network parameters [10], and the single-stream is memory efficient compared to multiple-stream networks. It has two primary advantages: 1) By minimizing visible to grayscale variations with the weights-sharing network, the learned representation is more robust to the color variations by capturing the structure relations between the visible and grayscale images. This property greatly improves the visible-infrared person re-identification performance, since the color information is also lost in the infrared person images. 2) By minimizing the discrepancy between the grayscale and infrared images, the network can capture more discriminative texture/shape cues in the single channel image.

We propose to minimize the visible-grayscale-infrared modality variations in the following two different aspects:

- Multi-modal Classification It optimizes the network by formulating the tri-modality person re-identification as a multi-class classification problem in the training process, *i.e.,* images of the same identity from three different modalities as the same class. It aims at learning multi-modality identity invariant feature representations.
- Multi-view Retrieval: It learns the representation from the instance retrieval across multiple views, optimizing the relationship between different person images across three modalities with a modified ranking loss.

B. Homo- and Heterogeneous Multi-modal Classification

With the Homogeneous and Heterogeneous Identification (HHI) loss, both the homogeneous pair (visible-grayscale) and heterogeneous pair (visible-infrared) can be correctly classified with the learned sharing identity classifier. HHI loss contains two main components: modality-sharing identity classifier (Fig. 4) and homogeneous invariant regularizer (Fig. 5).

Modality-sharing Identity Classifier. It learns a shared classifier θ^0 for three different modalities. $p(y_i|\mathbf{x}_i^v; \theta^0)$ represents the predicted probability output of a visible image $\mathbf{x}_i^v$ being recognized as its identity label y_i using the classifier θ^0. Assuming that each training batch contains n visible images,

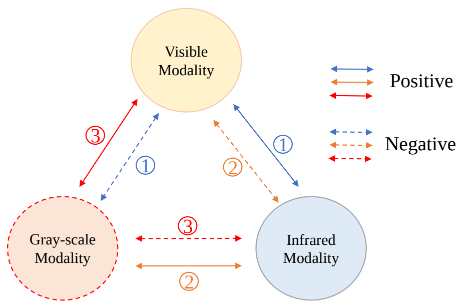


Fig. 6. Illustration of Tri-Directional Ranking (TDR) loss for multi-view retrieval. It aims at optimizing the relationship between visible-infrared, visible-grayscale and infrared-grayscale modalities, incorporated with a modified triplet-mining ranking loss. We use the format *A-B-C* to represent a triplet, where A represents the anchor modality, B and C denote the positive and negative modality, respectively. Specifically, ① for visible-infrared-grayscale triplet, ② for infrared-grayscale-visible triplet, and ③ for grayscale-visible-infrared triplet.

n homogeneously generated grayscale images and n infrared images, the HHI loss with softmax cross-entropy is represented by

$$\begin{aligned}\mathcal{L}_0 = &-\frac{1}{n}\sum_{i=1}^{n}\log(p(y_i|\mathbf{x}_i^v;\theta^0)) - \frac{1}{n}\sum_{i=1}^{n}\log(p^0(y_i|\mathbf{x}_i^g;\theta^0))\\ &-\frac{1}{n}\sum_{i=1}^{n}\log(p(y_i|\mathbf{x}_i^r;\theta^0)).\end{aligned} \tag{3}$$

Homogeneous Invariant Regularizer. To enhance the robustness against modality variations, we incorporate a homogeneous invariant regularizer. The main idea is that the features of the original visible image and homogeneously augmented grayscale image are invariant through the feature extraction network, *i.e.*, $\mathbf{f}_i^v - \mathbf{f}_i^g$ are minimized. Specifically, we adopt a smooth L1 loss as the regularization

$$\mathcal{L}_r = \sum_{i\in\mathcal{B}}\begin{cases}0.5(\mathbf{f}_i^v - \mathbf{f}_i^g)^2, & |\mathbf{f}_i^v - \mathbf{f}_i^g| < 1\\ |\mathbf{f}_i^v - \mathbf{f}_i^g|, & otherwise.\end{cases} \tag{4}$$

The overall HHI loss is the summarization of $\mathcal{L}_0$ and $\mathcal{L}_r$, denoted by

$$\mathcal{L}_{hhi} = \mathcal{L}_0 + \alpha\mathcal{L}_r, \tag{5}$$

where α controls the contribution of the invariant regularizer. The designed HHI loss has the following advantages: 1) The modality-sharing identity classifier ensures that the network optimization is conducted on the feature representations rather than classifiers. 2) The invariant regularizer enhances the robustness against color variations, making it discriminative to single-channel infrared image.

C. Weighted Tri-directional Ranking for Multi-view Retrieval

This subsection presents the designed Weighted Tri-directional Ranking (TDR) loss for multi-view retrieval, which optimizes the relationship with multi-view cross-modality retrieval (visible-infrared-grayscale). Different from previous triplets, each triplet consists mined samples with cross-view retrieval from three different modalities.

Informative Triplet Mining. We firstly select the informative triplets with cross-view retrieval, avoiding the domination of abundant easy triplets [59]–[61]. We denote the Euclidean distance between two samples $\mathbf{x}_i^v$ and $\mathbf{x}_j^r$ as $D(\mathbf{x}_i^v, \mathbf{x}_j^r)$, where the subscripts $\{i, j, k\}$ represent the image index and the superscripts $\{v, r, g\}$ denote the modality index. We firstly consider visible modality as the anchor modality, and then we search its positive from infrared modality and negative from grayscale modality. Formally, let $\mathbf{x}_i^v$ be an anchor visible sample, a triplet $\{\mathbf{x}_i^v, \mathbf{x}_j^r, \mathbf{x}_k^g\}$ is selected if it satisfies the constraints:

$$\mathcal{P}_{i,j}^{v,r} = \max_{\forall y_j = y_i} D(\mathbf{x}_i^v, \mathbf{x}_j^r) \tag{6}$$

$$\mathcal{N}_{i,k}^{v,g} = \min_{\forall y_k \neq y_i} D(\mathbf{x}_i^v, \mathbf{x}_k^g) \tag{7}$$

For each anchor $\mathbf{x}_i^v$, above strategy chooses the farthest visible-infrared positive pair and selects the closest visible-grayscale negative pair, formulating a mined informative triplet $\{\mathbf{x}_i^v, \mathbf{x}_j^{r+}, \mathbf{x}_k^{g-}\}$ from cross-view ranking perspective. Generally, the triplet loss for *visible-infrared-grayscale*[1] using a margin parameter ρ is then defined by

$$\mathcal{L}_{v,r,g} = \frac{1}{n}\sum_{i=1}^{n}\max[\rho + D(\mathbf{x}_i^v, \mathbf{x}_j^{r+}) - D(\mathbf{x}_i^v, \mathbf{x}_k^{g-}), 0] \tag{8}$$

Similarly, we could mine the informative triplets for *infrared-grayscale-visible* relationship, denoted by $\{\mathbf{x}_i^r, \mathbf{x}_j^{g+}, \mathbf{x}_k^{v-}\}_{i=1}^n$, and *grayscale-visible-infrared* relationship, represented by $\{\mathbf{x}_i^g, \mathbf{x}_j^{v+}, \mathbf{x}_k^{r-}\}_{i=1}^n$ from ranking perspective. An illustration of three different relationships is shown in Fig. 6. Summarizing loss of three different relationships, the tri-directional ranking (TDR) loss is

$$\begin{aligned}\mathcal{L}_{tdr} =& \mathcal{L}_{v,r,g} + \mathcal{L}_{r,g,v} + \mathcal{L}_{g,v,r}\\ =&\frac{1}{n}\sum_{i=1}^{n}\max[\rho + D(\mathbf{x}_i^v, \mathbf{x}_j^{r+}) - D(\mathbf{x}_i^v, \mathbf{x}_k^{g-}), 0]\\ &+\frac{1}{n}\sum_{i=1}^{n}\max[\rho + D(\mathbf{x}_i^r, \mathbf{x}_j^{g+}) - D(\mathbf{x}_i^r, \mathbf{x}_k^{v-}), 0]\\ &+\frac{1}{n}\sum_{i=1}^{n}\max[\rho + D(\mathbf{x}_i^g, \mathbf{x}_j^{v+}) - D(\mathbf{x}_i^g, \mathbf{x}_k^{r-}), 0]\end{aligned} \tag{9}$$

Above tri-directional ranking loss fully utilizes the cross-modality triplet-wise relationship in different views. It minimizes the relative difference between the farthest cross-modality positive pair distance and the closest negative pair distance, improving the robustness against modality variations. The informative triplet mining enhances the discriminability of the learned cross-modality features.

Triplet Global Weighting. The TDR loss treats each anchor sample equally, *all the mined triplet contributes equally to the overall loss*. However, due to the sample variety, the contribution of each triplet should be different, especially for the hard triplets. This part presents a simple but effective triplet weighting strategy, termed as *triplet global weighting*. Specifically, we use w_i^v represents the weight of the mined

[1] We use the format *A-B-C* to represent a triplet, where A represents the anchor modality, B and C denote the positive and negative modality, respectively.

triplet$\{\mathbf{x}_i^v, \mathbf{x}_j^{r+}, \mathbf{x}_k^{g-}\}$, which is calculated by an exponential function of the relative difference,

$$w_i^v = \exp([\rho + D(\mathbf{x}_i^v, \mathbf{x}_j^{r+}) - D(\mathbf{x}_i^v, \mathbf{x}_k^{g-})]_+). \tag{10}$$

where $[\cdot]_+ = \max(\cdot, 0)$. The above weight calculation does not introduce any additional pairwise comparison or hyper-parameters [62], which can be seamlessly incorporated in the original ranking loss. Similarly, we can calculate the triplet weights $\{w_i^r\}_{i=1}^n$ and $\{w_i^g\}_{i=1}^n$, using the calculated distance difference. To well balance the importance of each triplet in the training batch, we apply a normalization to each triplet weight, denoted by

$$\bar{w}_i^v = \frac{3 * w_i^v}{\sum_{i=1}^n w_i^v + \sum_{i=1}^n w_i^r + \sum_{i=1}^n w_i^g}. \tag{11}$$

The normalized $\{\bar{w}_i^r\}_{i=1}^n$ and $\{\bar{w}_i^g\}_{i=1}^n$ can also be obtained in a similar manner. The normalized triplet weight measures the importance of each mined triplet to the overall learning objective, where hard triplets will be assigned with large weights. The calculated weights update dynamically along with the network optimization. Meanwhile, this triplet weight measures the relationship of each triplet to all the other mined triplets, and it provides additional supervision of within triplet similarity [63]. Therefore, it results in a global optimization for each sampled batch. When $\bar{w}_i^v = \frac{1}{n}$, $\bar{w}_i^r = \frac{1}{n}$ and $\bar{w}_i^g = \frac{1}{n}$, the weighted TDR loss will be degenerated to the general tri-directional ranking loss (Eq. 9). With the normalized triplet weights, the weighted TDR loss is refined by

$$\begin{aligned}\mathcal{L}_{wtdr} = &\sum_{i=1}^n \bar{w}_i^v \cdot \max[\rho + D(\mathbf{x}_i^v, \mathbf{x}_j^{r+}) - D(\mathbf{x}_i^v, \mathbf{x}_k^{g-}), 0] \\ &+ \sum_{i=1}^n \bar{w}_i^r \cdot \max[\rho + D(\mathbf{x}_i^r, \mathbf{x}_j^{g+}) - D(\mathbf{x}_i^r, \mathbf{x}_k^{v-}), 0] \\ &+ \sum_{i=1}^n \bar{w}_i^g \cdot \max[\rho + D(\mathbf{x}_i^g, \mathbf{x}_j^{v+}) - D(\mathbf{x}_i^g, \mathbf{x}_k^{r-}), 0]\end{aligned} \tag{12}$$

Cross-modality Positive Pair Regularizer. To further minimize the cross-modality variations, we add a positive pair invariant regularizer to the WTDR loss. The main idea is to explicitly minimize each sampled cross-modality positive pair distance, which is formulated by

$$\mathcal{L}_{reg} = \frac{1}{n}\sum_{i=1}^n D(\mathbf{x}_i^v, \mathbf{x}_j^{r+}) + D(\mathbf{x}_i^r, \mathbf{x}_j^{g+}) + D(\mathbf{x}_i^g, \mathbf{x}_j^{v+}) \tag{13}$$

The regularization provides additional pairwise relationship optimization to the triplet constraint. The main benefit is that TDR loss optimizes the relative distance while this regularizer explicitly minimizes the cross-modality variations of each sampled positive pair. Indeed, an ideal case is that all the triplets satisfy the triplet margin constraint and the distances of all the cross-modality positive pairs are zero. But this condition is actually hard to be true for a large scale scenario, especially for the applications with large cross-modality variations. Therefore, a pairwise regularizer is essential to minimize the cross-modality variation. Extensive experiments demonstrate that this regularization improves the VI-ReID performance.

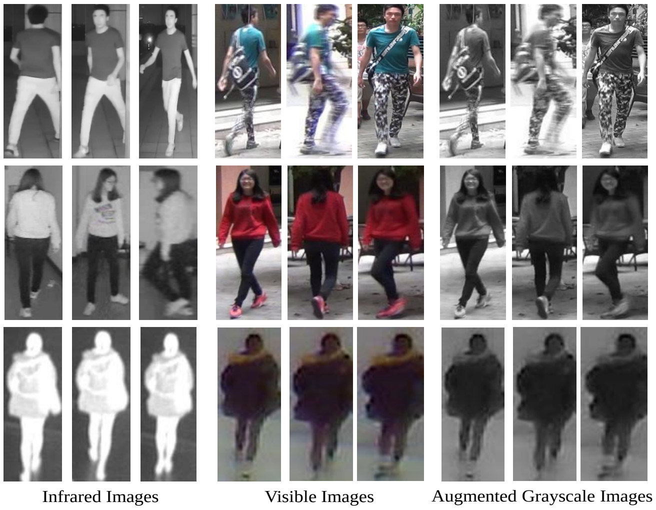


Fig. 7. Sampled images and the augmented grayscale images from SYSU-MM01 dataset [10] (first two rows) and RegDB dataset [11] (last row) . Each row represents the images of the same identity from three different modalities.

D. Joint Training with Identity Sampling

The total loss $\mathcal{L}_{hat}$ is then defined by the combination of $\mathcal{L}_{hhi}$ and $\mathcal{L}_{wtdr}$,

$$\mathcal{L}_{hat} = \mathcal{L}_{hhi} + \mathcal{L}_{wtdr} + \beta\mathcal{L}_{reg}, \tag{14}$$

where β controls the contribution of the cross-modality pair regularizer. The HHI loss $\mathcal{L}_{hhi}$ optimizes the parameter-sharing network with the identity supervision, which learns multi-modality identity-invariant feature. The WTDR loss $\mathcal{L}_{wtdr}$ provides the supervision to optimize the relative distance from triple views for retrieval. These two components are optimized jointly for the cross-modality Re-ID model learning.

Identity Sampling Strategy. To guarantee the informative triplet mining, we design an identity sampling strategy for training. At each training step, 8 identities are randomly selected, and then 4 visible and 4 infrared images are sampled for each identity, thus $n = 32$ in Eq. 3 and Eq. 9 in all the settings. With the augmented grayscale images, the total number of training images in each batch is 96. This strategy provides rich well-mined the positive and negative samples.

IV. Experimental Results

A. Experimental Settings

Datasets and Settings. Two publicly VI-ReID datasets (SYSU-MM01 [10] and RegDB [11]) are evaluated. The infrared images are captured by near-infrared cameras for SYSU-MM01 [10] and by thermal cameras for RegDB [11]. We also plot some example images from two datasets in Fig. 7, together with homogeneously augmented grayscale images.

SYSU-MM01 dataset [10] is a large-scale dataset collected by 6 different cameras, including 4 general RGB cameras and 2 near-infrared cameras, captured in SYSU campus from both indoor and outdoor environment. This dataset contains 395 training identities, including 22,258 visible and 11,909 near-infrared images, where the images are captured in both indoor and outdoor environments. The testing set contains another 95 testing identities, with two different evaluation settings. In both settings, the query set is the same, containing 3803 images captured from two IR cameras. In *all-search* mode, the gallery set contains all the visible images captured from all four RGB cameras. In *indoor-search* mode, the gallery set only contains the visible images captured by two indoor RGB cameras. Generally, the *all-search* mode is more challenging than the *indoor-search* mode. We exactly follow existing methods to perform ten trials of the gallery set selection [12], [14], and the report the average retrieval performance. Details description of the evaluation protocol can be found in [10].

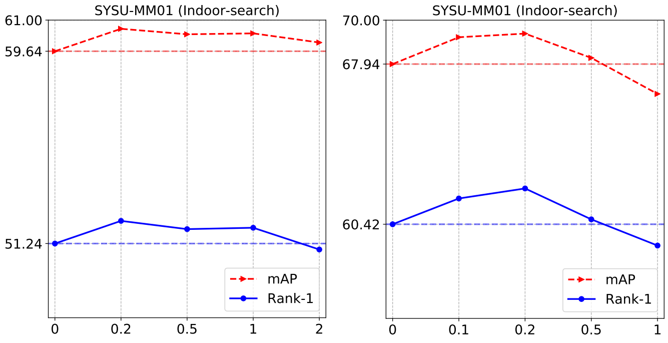
RegDB dataset [11] is a small-scale dataset captured by a dual-camera (one visible and one thermal camera) system. Thus, the visible and infrared images are quite similar, and it is less challenging for cross-modality person re-identification. Totally, this dataset contains 412 identities and each identity has ten visible and ten thermal images. Following the evaluation protocol in [42], we randomly select 206 identities (with 2,060 images) for training and the rest 206 identities (with 2,060 images) are used for testing. We evaluate two different retrieval settings, including both visible-thermal and thermal-visible retrieval performance. The average accuracy of ten randomly training/testing splits is reported [42].

Evaluation Metrics. Following existing works, Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) are aopdted as the evaluation metrics. CMC (Rank-k accuracy) measures the probability of a correct cross-modality person image occurs in the top-k retrieved results. mAP measures the retrieval performance when multiple matching images occur in the gallery set [64].

Implementation Details. The proposed algorithm is implemented on PyTorch framework. Following existing person Re-ID works, ResNet50 [65] is adopted as the backbone network for fair comparison, and the pretrained ImageNet parameters are adopted for the network initialization. Specifically, the stride of the last convolutional block is set to 1 to obtain fine-grained feature maps [58]. We adopt the default operation (*Grayscale(3)*) in Pytorch to get the homogeneously augmented grayscale image for each visible image. All the input images from three modalities are firstly resized to 288×144, and random crop with zero-padding together with random horizontal flipping are adopted for data argumentation. We adopt the stochastic gradient descent (SGD) optimizer for optimization, and the momentum parameter is set to 0.9. We set the initial learning rate as 0.1 for both datasets. The learning rate is decayed by 0.1 at 20 epoch and 0.01 at 50 epoch, with totally 60 training epochs on both datasets. We set the margin parameter ρ in the TDR loss to 0.3. We set $\alpha = 1$ and $\beta = 0.2$ by default. We use the output of the batch normalization (BN) layer for retrieval in the testing phase, and use the original visible image for feature extraction.

B. Self Evaluation

Evaluation of Each Component. We first evaluate each component in the proposed method on the large-scale SYSU-MM01 dataset under both *all-search* and *indoor-search* modes. Specifically, "**B**" represents the baseline performance by using the original 3-channel RGB visible images and 3-copied infrared channel, which is trained with an identity loss. "$\mathbf{H}_0$" represents the results by adding the homogeneously augmented grayscale images, trained with a modality-sharing identity classifier (Eq. 3). "**HHI**" represents the full version of HHI loss (Eq. 5) together with the homogeneous invariant regularizer. "**TDR**" represented the proposed tri-directional ranking loss (Eq. 9). "**WTDR**" represented the TDR loss with triplet global weighting (Eq. 12). "$\mathbf{WTDR}_r$" denotes the performance when cross-modality positive pair regularizer (Eq. 13) is aggregated. The results of adding/removing each component are shown in Table I.



(a) α in Eq. 5 (b) β in Eq. 14

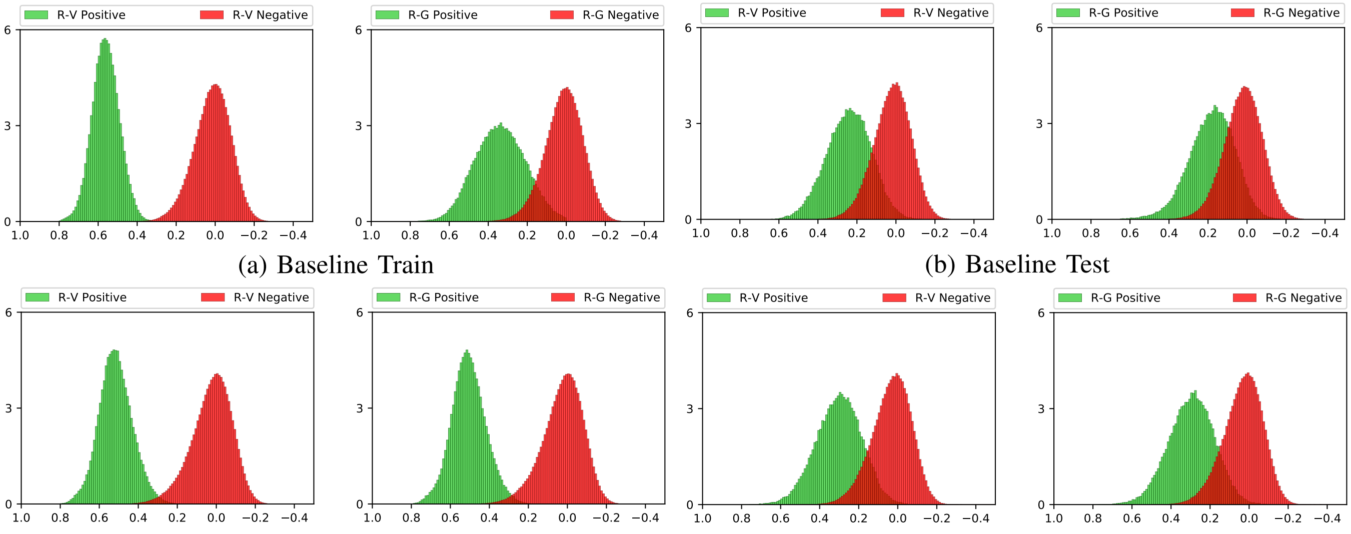
Fig. 8. Evaluation of different α in Eq. 5 (left) and different β in Eq. 14 (right) on SYSU-MM01 dataset, under the challenging *indoor-search* mode. The dotted lines represent the baseline in each setting. Rank-1 matching accuracy (%) and mAP (%) are reported.

1) Effectiveness of grayscale homogeneous augmentation: Compared to the baseline model (**B**), the performance with the grayscale augmented images is grealty improved, *i.e.,* the rank-1 accuracy changed from 45.30% to 47.80%. This experiment demonstrates the effectiveness of the grayscale image augmentation for cross-modality person re-identification. *2) Effectiveness of homogeneous invariant regularizer* $\mathcal{L}_r$: When we further intergrate with the invariant regurlarizer, the performance in both settings is further improved. This verifies that minimizing the homogeneously visible-grayscale variations can mine the useful information from color channels with single-channel grayscale image, improving the performance for visible-infrared person re-identification. *3) Effectiveness of TDR loss*: Integrating with the tri-directional ranking loss, the retrieval performance is greatly improved, which demonstrates the multi-view cross-modality ranking loss provides strong supervision to enhance the discriminability in the testing retrieval phase. *4) Effectiveness of triplet weighting*: When we re-weight the triplets according to the distance difference, the performance is slightly improved, which adaptively re-weight each mined triplet. *5) Effectiveness of cross-modality positive pair regularizer* $\mathcal{L}_{reg}$: this terms explicitly minimizes the cross-modality distances of the positive pair, which reinforces the robustness against the cross-modality variations for each mined hard positive pair. It complements the triplet constraint with relative distance optimization.

Evaluation of Regularization Parameters. We evaluate the effect of two hyperparameters in the proposed method, α in Eq. 5 and β in Eq. 14. The results on SYSU-MM01 dataset under indoor-search mode are shown in Fig. 8.

TABLE I
EVALUATION OF EACH COMPONENT IN OUR PROPOSED METHOD ON THE LARGE-SCALE SYSU-MM01 DATASET, INCLUDING BOTH ALL-SEARCH AND INDOOR-SEARCH MODES. RANK AT r MATCHING ACCURACY(%) AND MAP (%) ARE REPORTED.

Datasets	*All Search*					*Indoor Search*				
Methods	$r = 1$	$r = 5$	$r = 10$	$r = 20$	mAP	$r = 1$	$r = 5$	$r = 10$	$r = 20$	mAP
B	45.30	73.47	83.52	92.29	44.82	48.85	79.23	90.18	96.87	57.84
B + H_0	47.80	76.29	86.56	94.12	45.99	51.24	82.18	91.52	97.26	59.64
B + HHI	48.23	77.05	87.26	94.79	46.18	51.93	82.53	92.25	98.26	60.42
B + HHI + TDR	54.99	83.08	91.25	96.41	53.26	60.01	88.45	95.13	98.45	67.44
B + HHI + WTDR	54.72	83.28	91.58	96.54	53.64	60.42	88.75	95.33	98.59	67.94
B + HHI + WTDR$_r$	55.29	83.74	92.14	97.36	53.89	62.10	89.35	95.75	99.20	69.37



(a) Baseline Train (b) Baseline Test

(c) HAT Train (d) HAT Test

Fig. 9. Visualization of Baseline (trained with identity+triplet loss, in the first row) and the proposed HAT method (second row). The x axis represents the cosine similarity score between two samples (positive or negative cross-modality matching pair), y axis is the statistical value for each cosine similarity score. Both strategies perform well in separating the infrared and RGB visible images (R-V positive and negative pair) in the training set. However, the proposed HAT performs much better than baseline in the testing set by incorporating with the homogeneously augmented grayscale images.

1) Effect of α: It controls the contribution of the invariant regularizer in the HHI loss. Results shown in Fig. 8 (a) demonstrate the this term consistently improves the performance when $\alpha \in [0, 1]$. This benefit is brought by mining the underlying structure relationship between the three-channel visible image and homogeneously augmented single-channel grayscale image. *2) Effect of β*: It provides additional pairwise constraint to the triplet relative distance optimization, which enhances the robustness against the cross-modality variations. Empirically, we set $\beta = 0.2$ in all our experiments.

C. In-depth Analysis

Evaluation of Grayscale Augmentation. We evaluate the grayscale augmentation with four different training strategies, as shown in Table II. In addition, we also evaluate the performance of using other pixel-to-pixel wise transform filters used in [66] for comparison, including the DoG filter, CSDN filter and Gaussian smoothing filter. All these strategies are trained with a modality-sharing identity classifier. Most existing methods adopt *RGB-Infrared* training strategy [14], [15], [17], where the original three RGB channels of the visible image are fed into the network. The performance is lower than the proposed tri-modal learning strategy. Even these grayscale images are just linear accumulation of the original RGB channels, the augmented grayscale images greatly improve the performance. When applied with the single-channel grayscale-infrared training strategy [10], the performance drops slightly. This experiment proves the effectiveness of the color information for visible-infrared person re-identification. Mixture-infrared represents the training with *RandomGrayscale(0.5)*, this strategy randomly optimizes the grayscale-infrared and visible-infrared relationships. *This experiment provides a good suggestion for future visible-infrared person re-identification research by applying this augmentation operation.*

TABLE II
EVALUATION OF GRAYSCALE AUGMENTATION WITH DIFFERENT TRAINING STRATEGIES, INCLUDING RGB-INFRARED, GRAYSCALE-INFARERD, MIXTURE-INFARED, TI-MODAL TRAINING. WE ALSO EVALUATE THE PIXEL-TO-PIXEL TRANSFORM FILTERS USED IN [66] FOR COMPARISON. RANK-1 ACCURACY (%) AND MAP (%) ON THE LARGE-SCALE SYSU-MM01 DATASET ARE REPORTED.

	All Search		*Indoor Search*	
Strategy	$r = 1$	mAP	$r = 1$	mAP
RGB-Infrared	45.30	44.82	48.85	57.84
Grayscale-Infarerd	43.76	43.12	46.56	55.98
Mixture-Infarerd	47.23	45.83	50.23	58.72
DoG [66]	46.28	45.23	49.62	58.42
CSDN [66]	45.82	45.12	50.12	58.02
Gaussian [66]	45.43	45.02	49.82	58.36
Tri-modal (Ours)	47.80	45.99	51.24	59.64
Tri-modal (Learn)	48.12	46.53	51.86	60.12

Compared to *other image filters* in [66], we find that these

TABLE III
EVALUATION OF DIFFERENT RETRIEVAL STRATEGIES BY CHANGING THE GALLERY IMAGES. RANK-1 ACCURACY (%) AND MAP (%) ON THE SYSU-MM01 DATASET ARE REPORTED. QUERY SET CONTAINS THE SINGLE-CHANNEL INFRARED IMAGES.

	All Search		*Indoor Search*	
Gallery	$r = 1$	mAP	$r = 1$	mAP
RGB	55.29	53.89	62.10	69.37
Gray	53.89	52.24	61.06	68.35
RGB + Gray (Feat)	54.83	53.34	61.75	69.10
RGB + Gray (Dist)	54.96	53.37	61.88	69.13

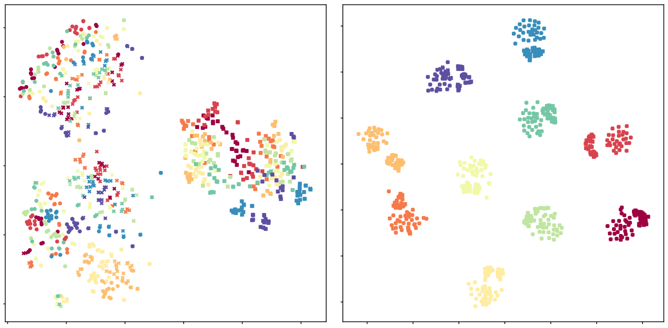
image filters can also slightly improve the performance, but the improvement is not significant as our proposed strategy. The main reason might be that the image filters perform well in the face recognition tasks, but they might not be directly applied to the person re-identification task. In our future work, we will explore how to design a more powerful transformation strategy explicitly for the VI-ReID.

Compared to *learnable linear transform*, we also evaluate the performance by learning a linear transformation function for the grayscale image generation, termed as *"Tri-modal (Learn)"*. We observe that the performance is slightly better than a simple transformation baseline.

Visualization Analysis. We visualize the cosine similarity distribution of positive/negative cross-modality matching pairs on both training and testing set in Fig. 9. We evaluate the proposed HAT and the baseline (trained identity + triplet loss, a strong baseline in [58]). Both strategies perform well in separating the infrared and RGB visible images (R-V positive and negative) in the training set. However, the proposed HAT performs much better than baseline in the testing set by incorporating with the homogeneously augmented grayscale images. Specifically, when the grayscale augmentation is used for tri-modal learning, the infrared-visible positive/negative difference of HAT is much larger than the baseline (R-V Positive v.s. R-V Negative). This experiment demonstrates that training with the homogeneously augmented grayscale images improves the generalizability on the testing set to discriminate the visible and infrared images. In terms of distinguishing the infrared and grayscale images (R-G Positive v.s. R-G Negative), HAT also shows stronger discriminability to discriminate the grayscale and infrared images. This experiment demonstrates that learning with the grayscale image results in better robustness against the color variations, *i.e.,* performing better to distinguish the grayscale and infrared images.

In summary, learning with the grayscale augmented images has two major advantages: 1) minimizing the grayscale-visible variations enhances the robustness against color variations. 2) minimizing the grayscale-infrared variations enforces the network mines the discriminative sharable structure cues within the person image, improving the discriminability for cross-modality visible-infrared person re-identification.

t-SNE Analysis. We plot the t-SNE map of 10 randomly selected identities on the SYSU-MM01 dataset in Fig. 10. We plot the features distribution of grayscale, visible and infrared images at the initial and the final stage. We observe that the grayscale modality is closer to the infrared modality at the initial stage than the original visible modality. And the features of the visible (grayscale) images and the infrared images are located in two different areas. With our proposed HAT method, the features of each identity from three modalities are concentrated in the learned embedding space.



(a) HAT Initial (b) HAT Train

Fig. 10. t-SNE visualization of ten randomly selected identities on the SYSU-MM01 dataset. Circle means visible modality, square means the infrared modality, and cross means the augmented grayscale modality. The color represents the identity. We observe that the grayscale modality is closer to the infrared modality at the initial stage than the original visible modality. With our proposed HAT method, the features of each identity from three modalities are concentrated together in the learned embedding space.

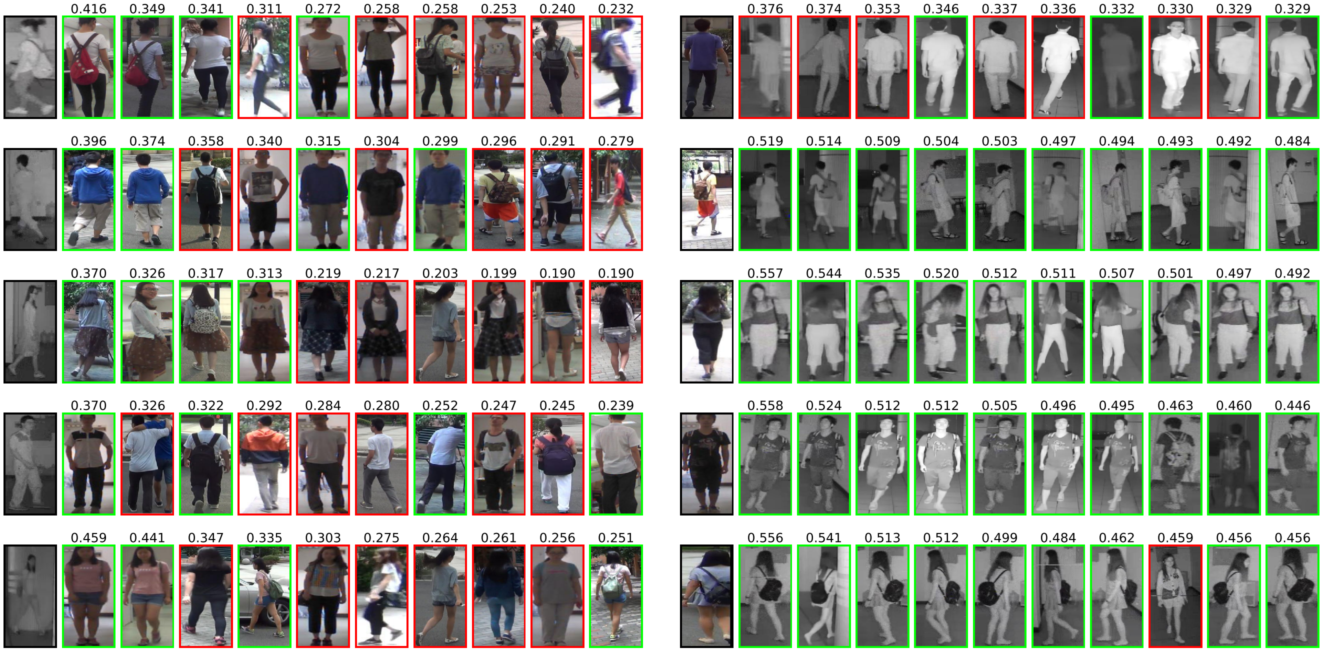
TABLE IV
COMPARISON WITH OTHER TRIPLET VARIANTS. RANK-1 ACCURACY (%) AND MAP (%) ON SYSU-MM01 DATASET ARE REPORTED.

	All Search		*Indoor Search*	
Strategy	$r = 1$	mAP	$r = 1$	mAP
Baseline	48.23	46.18	51.93	60.42
Triplet	51.33	49.32	56.82	63.18
Triplet(Hard) [36]	53.32	51.78	58.43	64.32
Ours	54.99	53.26	60.01	67.44

Evaluation of Different Retrieval Strategies. We also evaluate different retrieval strategies, as shown in Table III. By default, the single-channel grayscale images formulate the query set. For the gallery images, we test the performance with RGB, grayscale images and their combination. We observe that the combination of RGB and grayscale does not improve the performance. The main reason is that the proposed HAT already learns the modality invariant features. Meanwhile, it also demonstrates the importance of the color information for visible-infrared person re-identification.

Retrieved Examples. We also plot some retrieved results on SYSU-MM01 dataset, including both visible-infrared and infrared to visible query settings. The retrieved results of 10 randomly selected query examples with the calculated cosine similarity score are shown in Fig. 11. The results demonstrate that HAT retrieve good results when the person appearance preserves rich structure cues (e.g., bags or stripes). Interestingly, some person images are even difficult for human, but the proposed method can still retrieve the correct results.

Comparison with Other Triplet Variants. We now evaluate the different triplet variants, as shown in IV. The baseline represents the tri-modal learning with HHI loss. Consistent improvements can be obtained by integrating with the triplet loss, which provides the relative distance optimization. Compared to the triplet loss with hard mining [36], the proposed solution achieves higher performance by explicitly optimizing the cross-modality relationships in different views.



(a) Infrared to Visible.

(b) Visible to Infrared

Fig. 11. Ten randomly selected query examples and their top-ten retrieved results on the SYSU-MM01 dataset, including two different query settings, visible to infrared and infrared to visible. Corrected retrieved samples are in green boxes and wrong results are in red boxes. Cosine similarity score is reported for each image pair (best viewed in color.)

TABLE V
COMPARISON WITH THE STATE-OF-THE-ARTS ON SYSU-MM01 DATASET. RANK AT r ACCURACY (%) AND MAP (%) ARE REPORTED.

Settings			*All Search*				*Indoor Search*			
Method	Venue	Type	$r = 1$	$r = 10$	$r = 20$	mAP	$r = 1$	$r = 10$	$r = 20$	mAP
One-stream [10]	ICCV17	(a)	12.04	49.68	66.74	13.67	16.94	63.55	82.10	22.95
Two-stream [10]	ICCV17	(a)	11.65	47.99	65.50	12.85	15.60	61.18	81.02	21.49
Zero-Pad [10]	ICCV17	(a)	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
TONE [42]	AAAI18	(b)	12.52	50.72	68.60	14.42	20.82	68.86	84.46	26.38
HCML [42]	AAAI18	(b)	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
cmGAN [15]	IJCAI18	(b)	26.97	67.51	80.56	31.49	31.63	77.23	89.18	42.19
eBDTR [14]	TIFS19	(b)	27.82	67.34	81.34	28.42	32.46	77.42	89.62	42.46
HSME [17]	AAAI19	(b)	20.68	32.74	77.95	23.12	-	-	-	-
D^{2}RL [12]	CVPR19	(c)	28.9	70.6	82.4	29.2	-	-	-	-
MAC [18]	MM19	(b)	33.26	79.04	90.09	36.22	36.43	62.36	71.63	37.03
MSR [16]	TIP19	(b)	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
AlignGAN [13]	ICCV19	(c)	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3
HPILN [67]	arXiv19	(b)	41.36	84.78	94.31	42.95	45.77	91.82	98.46	56.52
LZM [68]	arXiv19	(b)	45.00	89.06	-	45.94	49.66	92.47	-	59.81
HAT (Ours)	-	(d)	**55.29**	**92.14**	**97.36**	**53.89**	**62.10**	**95.75**	**99.20**	**69.37**

D. Comparison with State-of-the-Arts

This section conducts the comparison with the state-of-the-art VI-ReID methods, including eBDTR [14], HSME [17], D^{2}RL [12], MAC [18], MSR [16] and AlignGAN [13]. These methods are published in recent two years. AlignGAN [13], published in ICCV 2019, achieves the state-of-the-art performance by aligning the cross-modality representation in both the feature level and pixel level with GAN generated images. In addition, some arXiv papers are also included for comparison, including EDFL [69], HPILN [67] and LZM [68]. We also mark the method types (a, b, c and d) according to the four different learning solutions in Fig. 2. The results on two datasets are shown in Tables V and VI.

Results on two datsets demonstrate that the proposed method outperforms the current state-of-the-art by a large margin. We set a new baseline for this task, achieving 55.29%/53.89% rank-1 accuracy/mAP on the challenging SYSU-MM01 dataset. Except for the type “a” solution in early years, we observe that RGB-Grayscale (type “b”) and GAN-generated (type “c”) are two most popular approaches, but the former type ignores the image level discrepancy and the latter type introduces unavoidable noise in the image generation process. Compared to AlignGAN [13], our method achieves much higher accuracy on both datasets, and doesn’t need the complicated cross-modality image generation. In contrast, the homogeneously augmented grayscale images in our method can be efficiently generated with the linear accumulation of three RGB channels. In addition, our method does not need the adversarial training [12], [13], [15], which is hard to train. Moreover, the single-stream network also contains much less parameters compared to the two-stream network methods [14], [18], which is more suitable for applications with limited data. This experiment further verifies the advantage of our proposed type “d” solution.

TABLE VI
COMPARISON WITH THE STATE-OF-THE-ARTS ON REGDB DATASET. RANK AT r ACCURACY (%) AND MAP (%) ARE REPORTED.

Method	$r = 1$	$r = 10$	$r = 20$	mAP
Visible to Thermal				
HCML [42]	24.44	47.53	56.78	20.08
Zero-Pad [10]	17.75	34.21	44.35	18.90
eBDTR [14]	34.62	58.96	68.72	33.46
HSME [17]	50.85	73.36	81.66	47.00
D^{2}RL [12]	43.4	66.1	76.3	44.1
MAC [18]	36.43	62.36	71.63	37.03
MSR [16]	48.43	70.32	79.95	48.67
EDFL [69]	52.58	72.10	81.47	52.98
AlignGAN [13]	57.9	-	-	53.6
HAT (Ours)	**71.83**	**87.16**	**92.16**	**67.56**
Thermal to Visible				
HCML [42]	21.70	45.02	55.58	22.24
Zero-Pad [10]	16.63	34.68	44.25	17.82
eBDTR [14]	34.21	58.74	68.64	32.49
HSME [17]	50.15	72.40	81.07	46.16
MAC [18]	36.20	61.68	70.99	36.63
EDFL [69]	51.89	72.09	81.04	52.13
AlignGAN [13]	56.3	-	-	53.4
HAT (Ours)	**70.02**	**86.45**	**91.61**	**66.30**

The experiment on RegDB dataset (Table VI) demonstrates that HAT achieves the best performance in both query settings, usually by a large margin, achieving rank-1/mAP 71.83%/67.56% for the visible to thermal query setting. This experiment suggests that HAT can learn better cross-modality sharing feature representations by leveraging the grayscale images. The parameter-sharing one-stream network could well capture the relationship across three modalities to learn multi-modality sharing feature representations.

V. CONCLUSIONS

This paper presents a homogeneously augmented tri-modal (HAT) learning for cross-modality person re-identification. With the augmented grayscale images, we propose a homogeneous and heterogeneous identification loss to learn a modality sharing classifier for three modalities, training with a parameter sharing one-stream network. To mine the structure information between the 3-channel color image and the single-channel grayscale image, we design a homogeneous invariant regularizer. Furthermore, we introduce a weighted tri-directional ranking loss to optimize the relative distance across cross-modality positive and negative triplets. This strategy incorporates with an informative triplet mining scheme, explicitly optimizing the cross-modality relationships in different views. Integrated with a cross-modality positive pair invariant regularizer, the cross-modality re-identification performance is further improved. Extensive experiments demonstrate that the proposed method significantly outperforms the current state-of-the-art, on both visible-infrared and visible-thermal person re-identification tasks.

REFERENCES

[1] X. Zhu, X.-Y. Jing, X. You, W. Zuo, S. Shan, and W.-S. Zheng, "Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix," *IEEE TIFS*, vol. 13, no. 3, pp. 717–732, 2018.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020.

[3] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *CVPR*, vol. 6, 2017, p. 7.

[4] A. R. Lejbølle, K. Nasrollahi, B. Krogh, and T. B. Moeslund, "Person re-identification using spatial and layer-wise attention," *IEEE TIFS*, 2019.

[5] F. Ma, X.-Y. Jing, X. Zhu, Z. Tang, and Z. Peng, "True-color and grayscale video person re-identification," *IEEE TIFS*, vol. 15, pp. 115–129, 2019.

[6] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *ECCV*, 2016.

[7] W. Zhang, B. Ma, K. Liu, and R. Huang, "Video-based pedestrian re-identification by adaptive spatio-temporal appearance model," *IEEE TIP*, vol. 26, no. 4, pp. 2042–2054, 2017.

[8] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE TIP*, vol. 28, no. 6, pp. 2872–2881, 2019.

[9] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE TCSVT*, 2019.

[10] A. Wu, W.-s. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *ICCV*, 2017, pp. 5380–5389.

[11] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[12] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *CVPR*, 2019, pp. 618–626.

[13] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *ICCV*, 2019, pp. 3623–3632.

[14] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE TIFS*, vol. 15, pp. 407–419, 2020.

[15] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training." in *IJCAI*, 2018, pp. 677–683.

[16] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE TIP*, vol. 29, pp. 579–590, 2020.

[17] Y. Hao, N. Wang, J. Li, and X. Gao, "Hsme: Hypersphere manifold embedding for visible thermal person re-identification," in *AAAI*, 2019, pp. 8385–8392.

[18] M. Ye, X. Lan, and Q. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," in *ACM MM*, 2019, pp. 347–355.

[19] S. P. Mudunuri, S. Venkataramanan, and S. Biswas, "Dictionary alignment with re-ranking for low-resolution nir-vis face recognition," *IEEE TIFS*, vol. 14, no. 4, pp. 886–896, 2019.

[20] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *AAAI*, 2018, pp. 1679–1686.

[21] X. Wu, H. Huang, V. M. Patel, R. He, and Z. Sun, "Disentangled variational representation for heterogeneous face recognition," in *AAAI*, 2019, pp. 9005–9012.

[22] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *ICCV*, 2017, pp. 3800–3808.

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

[24] X. Zhu, B. Wu, D. Huang, and W.-S. Zheng, "Fast open-world person re-identification," *IEEE TIP*, vol. 27, no. 5, pp. 2286 – 2300, 2018.

[25] S. Bai, P. Tang, P. H. Torr, and L. J. Latecki, "Re-ranking via metric fusion for object retrieval and person re-identification," in *CVPR*, 2019, pp. 740–749.

[26] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, "Video person re-identification by temporal residual learning," *IEEE TIP*, vol. 28, no. 3, pp. 1366–1377, 2018.

[27] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *CVPR*, 2019, pp. 9317–9326.

[28] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *CVPR*, 2018, pp. 1179–1188.

[29] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *CVPR*, 2019, pp. 393–402.

[30] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *CVPR*, 2019, pp. 1389–1398.

[31] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, “Pyramidal person re-identification via multi-loss dynamic training,” in *CVPR*, 2019, pp. 8514–8522.

[32] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in *CVPR*, 2015, pp. 2197–2206.

[33] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, “Hierarchical gaussian descriptor for person re-identification,” in *CVPR*, 2016, pp. 1363–1372.

[34] K. Liu, B. Ma, W. Zhang, and R. Huang, “A spatio-temporal appearance representation for viceo-based pedestrian re-identification,” in *ICCV*, 2015, pp. 3810–3818.

[35] S. Liao and S. Z. Li, “Efficient psd constrained asymmetric metric learning for person re-identification,” in *ICCV*, 2015, pp. 3685–3693.

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

[37] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Deep metric learning for person re-identification,” in *ICPR*, 2014, pp. 34–39.

[38] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond triplet loss: a deep quadruplet network for person re-identification,” in *CVPR*, 2017, pp. 403–412.

[39] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, “Alignedreid: Surpassing human-level performance in person re-identification,” *arXiv preprint arXiv:1711.08184*, 2017.

[40] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling,” in *ECCV*, 2018, pp. 480–496.

[41] Z. Zhong, L. Zheng, S. Li, and Y. Yang, “Generalizing a person retrieval model hetero-and homogeneously,” in *ECCV*, 2018, pp. 172–188.

[42] M. Ye, X. Lan, J. Li, and P. C. Yuen, “Hierarchical discriminative learning for visible thermal person re-identification,” in *AAAI*, 2018, pp. 7501–7508.

[43] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang, “Improving deep visual representation for person re-identification by global and local image-language association,” in *ECCV*, 2018, pp. 54–70.

[44] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, “Identity-aware textual-visual matching with latent co-attention,” in *ICCV*, 2017, pp. 1890–1899.

[45] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, “Cross-modal retrieval with cnn visual features: A new baseline,” *IEEE TCYB*, vol. 47, no. 2, pp. 449–460, 2017.

[46] R. He, X. Wu, Z. Sun, and T. Tan, “Learning invariant deep representation for nir-vis face recognition.” in *AAAI*, 2017, pp. 2000–2006.

[47] C. Peng, N. Wang, J. Li, and X. Gao, “Re-ranking high-dimensional deep local representation for nir-vis face recognition,” *IEEE TIP*, 2019.

[48] N. Wang, X. Gao, L. Sun, and J. Li, “Bayesian face sketch synthesis,” *IEEE TIP*, vol. 26, no. 3, pp. 1264–1274, 2017.

[49] M. S. Sarfraz and R. Stiefelhagen, “Deep perceptual mapping for cross-modal face recognition,” *International Journal of Computer Vision*, vol. 122, no. 3, pp. 426–438, 2017.

[50] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, “Disentangled person image generation,” in *CVPR*, 2018, pp. 99–108.

[51] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, “Pose-normalized image generation for person re-identification,” in *ECCV*, 2018, pp. 650–667.

[52] L. Zheng, Y. Huang, H. Lu, and Y. Yang, “Pose invariant embedding for deep person re-identification,” *IEEE Transactions on Image Processing*, 2019.

[53] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, “Camera style adaptation for person re-identification,” in *CVPR*, 2018, pp. 5157–5166.

[54] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, “Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification,” in *CVPR*, 2018, pp. 994–1003.

[55] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, “Camstyle: A novel data augmentation method for person re-identification,” *IEEE TIP*, vol. 28, no. 3, pp. 1176–1190, 2018.

[56] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in *ICCV*, 2017, pp. 3754–3762.

[57] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, “Joint discriminative and generative learning for person re-identification,” in *CVPR*, 2019, pp. 2138–2147.

[58] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, “A strong baseline and batch normalization neck for deep person re-identification,” *arXiv preprint arXiv:1906.08332*, 2019.

[59] B. Harwood, B. Kumar, G. Carneiro, I. Reid, T. Drummond *et al.*, “Smart mining for deep metric learning,” in *ICCV*, 2017, pp. 2821–2829.

[60] Y. Wang, J. Choi, V. Morariu, and L. S. Davis, “Mining discriminative triplets of patches for fine-grained classification,” in *CVPR*, 2016, pp. 1163–1172.

[61] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, “Mancs: A multi-task attentional network with curriculum sampling for person re-identification,” in *ECCV*, 2018, pp. 365–381.

[62] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, “Multi-similarity loss with general pair weighting for deep metric learning,” in *CVPR*, 2019, pp. 5022–5030.

[63] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, “Deep metric learning via lifted structured feature embedding,” in *CVPR*, 2016, pp. 4004–4012.

[64] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *ICCV*, 2015, pp. 1116–1124.

[65] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.

[66] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li, “Multiple representations-based face sketch–photo synthesis,” *IEEE TNNLS*, vol. 27, no. 11, pp. 2201–2215, 2015.

[67] J.-W. Lin and H. Li, “Hpiln: A feature learning framework for cross-modality person re-identification,” *arXiv preprint arXiv:1906.03142*, 2019.

[68] E. Basaran, M. Gokmen, and M. E. Kamasak, “An efficient framework for visible-infrared cross modality person re-identification,” *arXiv preprint arXiv:1907.06498*, 2019.

[69] H. Liu and J. Cheng, “Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification,” *arXiv preprint arXiv:1907.09659*, 2019.

Mang Ye received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2013 and 2016, respectively. He received the Ph.D degree in Computer Science from Hong Kong Baptist University, Hong Kong, in 2019. He is currently a Research Scientist at Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests focus on multimedia retrieval, computer vision and pattern recognition.

Jianbing Shen (M’11-SM’12) is currently acting as the Lead Scientist at the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. He is also an adjunct Professor with the School of Computer Science, Beijing Institute of Technology. He has published more than 150 journal and conference papers, fifteen papers are selected as the ESI Hightly Cited. His current research interests include deep learning and computer vision. He serves as an Associate Editor for *IEEE Trans. on Image Processing*, *IEEE Trans. on Neural Networks and Learning Systems*, *Pattern Recognition* and other journals.

Ling Shao (Senior Member, IEEE) is currently the Executive Vice President and Provost of the Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates. He is also the CEO and the Chief Scientist of the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi. His research interests include computer vision, machine learning, and medical imaging. He is a fellow of IAPR, IET, and BCS.