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.

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
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 grayscale 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
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, si-
multaneously handling the cross-modality and intra-modality
variations with dual-constrained top-ranking loss. The image
generation network for three different modalities, including the
RGB channels, and it preserves the structure information.

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 adversar-
ial 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 addi-
tional 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 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 \( T = \{ T^v, T^i \} \). In particular, \( T^v = \{ x^v_i | i = 1, 2, \cdots, N^v \} \) represents visible training set with \( N^v \) visible
images, where each visible image \( x_i^v \) is associated with identity label \( y_i \). \( T^r = \{ x_i^r | i = 1, 2, \cdots, N^r \} \) represents \( N^r \) infrared training images, where each element \( x_i^r \) is an infrared image. Due to the light spectrums in different cameras, \( x_i^r \) usually has three different color channels, i.e., R, G and B, and infrared image \( x_i^r \) 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 \( x_i^v \), its homogeneous grayscale modality image \( x_i^g \) is generated by

\[
L = g(R, G, B)
\]

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 \( x_i^v \), \( x_i^r \) and \( x_i^g \) are denoted by

\[
f_i^v = f(x_i^v), f_i^r = f(x_i^r), f_i^g = f(x_i^g).
\]

The learning objective of VI-ReID is to narrow the gap between \( f_i^v \) and \( f_i^r \) 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 \( \theta_0 \) for three different modalities. \( p(y_i|x_i^v; \theta_0) \) represents the predicted probability output of a visible image \( x_i^v \) being recognized as its identity label \( y_i \) using the classifier \( \theta_0 \). Assuming that each training batch contains \( n \) visible images,

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

\[
\text{Fig. 5. Illustration of the homogeneous invariant regularizer in Homogeneous and Heterogeneous Identification (HHI) loss, i.e., the features of the original visible image and its homogeneously augmented grayscale image are close in the learned feature space.}
\]
of abundant easy triplets [59]–[61]. We denote the Euclidean distance between two samples $x_i^v$ and $x_j^v$ as $D(x_i^v, x_j^v)$, 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 $x_i^v$ be an anchor visible sample, a triplet $\{x_i^v, x_j^v, x_k^g\}$ is selected if it satisfies the constraints:

$$P_{i,j}^{v,r} = \max_{y_j = y_i} D(x_i^v, x_j^r)$$

$$N_{i,j,k}^{v,g} = \min_{y_k \neq y_i} D(x_i^v, x_k^g)$$

For each anchor $x_i^v$, above strategy chooses the farthest visible-infrared positive pair and selects the closest visible-grayscale negative pair, formulating a mined informative triplet $\{x_i^v, x_j^r, x_k^g\}$ from cross-view ranking perspective. Generally, the triplet loss for visible-infrared-gray-scale\(^1\) using a margin parameter $\rho$ is then defined by

$$L_{v,r,g} = \frac{1}{n} \sum_{i=1}^{n} \max[\rho + D(x_i^v, x_i^r) - D(x_i^v, x_i^g), 0]$$

Similarly, we could mine the informative triplets for infrared-grayscale-visible relationship, denoted by $\{x_i^r, x_j^g, x_k^v\}_{i=1}^{n}$, and grayscale-visible-infrared relationship, represented by $\{x_i^g, x_j^r, x_k^v\}_{i=1}^{n}$ 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

$$L_{tdr} = L_{v,r,g} + L_{r,g,v} + L_{g,v,r}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \max[\rho + D(x_i^v, x_i^r) - D(x_i^v, x_i^g), 0]$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \max[\rho + D(x_i^r, x_i^g) - D(x_i^r, x_i^v), 0]$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \max[\rho + D(x_i^g, x_i^v) - D(x_i^g, x_i^r), 0]$$

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 \( \{x^+_i, x^+_j, x^-_k\} \), which is calculated by an exponential function of the relative difference,

\[
\tilde{w}^v_i = \exp(\beta(\rho + D(x^+_i, x^+_j) - D(x^+_i, x^-_k))_+). \tag{10}
\]

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 \( \{w^v_i\}_{i=1}^n \) and \( \{w^g_i\}_{i=1}^n \), 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

\[
\tilde{w}^v_i = \frac{3 \cdot \tilde{w}^v_i}{\sum_{i=1}^n \tilde{w}^v_i + \sum_{i=1}^n \tilde{w}^g_i + \sum_{i=1}^n \tilde{w}^g_i}. \tag{11}
\]

The normalized \( \{\tilde{w}^v_i\}_{i=1}^n \) and \( \{\tilde{w}^g_i\}_{i=1}^n \) 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 \( \tilde{w}^v_i = \frac{1}{n} \), \( \tilde{w}^g_i = \frac{1}{n} \) and \( \tilde{w}^g_i = \frac{1}{n} \), 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

\[
\mathcal{L}_{wtdr} = \sum_{i=1}^n \tilde{w}^v_i \cdot \max[\beta(\rho + D(x^+_i, x^+_j) - D(x^+_i, x^-_k)), 0] + \sum_{i=1}^n \tilde{w}^g_i \cdot \max[\beta(\rho + D(x^+_i, x^+_j) - D(x^+_i, x^-_k)), 0] + \sum_{i=1}^n \tilde{w}^g_i \cdot \max[\beta(\rho + D(x^+_i, x^+_j) - D(x^+_i, x^-_k)), 0]. \tag{12}
\]

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

\[
\mathcal{L}_{reg} = \frac{1}{n} \sum_{i=1}^n D(x^+_i, x^+_j) + D(x^+_i, x^+_j) + D(x^+_i, x^+_j). \tag{13}
\]

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.

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 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 adopted 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 α = 1 and β = 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. “B_i” 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). “WTDR,” 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.

1) Effectiveness of grayscale homogeneous augmentation: Compared to the baseline model (B), the performance with the grayscale augmented images is greatly 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 L_reg: When we further integrate with the invariant regularizer, 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 images. 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 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 MATCHING ACCURACY (%) AND mAP (%) ARE REPORTED.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>All Search</th>
<th>Indoor Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>r = 1</td>
<td>r = 5</td>
</tr>
<tr>
<td>B</td>
<td>45.30</td>
<td>83.52</td>
</tr>
<tr>
<td>B + H₀</td>
<td>47.80</td>
<td>86.56</td>
</tr>
<tr>
<td>B + HHI</td>
<td>48.23</td>
<td>87.26</td>
</tr>
<tr>
<td>B + HHI + TDR</td>
<td>54.99</td>
<td>91.25</td>
</tr>
<tr>
<td>B + HHI + WTDR</td>
<td>54.72</td>
<td>91.58</td>
</tr>
<tr>
<td>B + HHI + WTDR</td>
<td>55.29</td>
<td>92.14</td>
</tr>
</tbody>
</table>

![Graphs](image_url)

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 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 this term consistently improves the performance when α ∈ [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 β = 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.

Compared to other image filters in [66], we find that these
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. With our proposed HAT method, the features of each identity from three modalities are concentrated together in the learned embedding space.

**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.
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], D²RL [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 datasets 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>r = 1</th>
<th>r = 10</th>
<th>r = 20</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visible to Thermal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCML [42]</td>
<td>24.44</td>
<td>47.53</td>
<td>56.78</td>
<td>20.08</td>
</tr>
<tr>
<td>Zero-Pad [10]</td>
<td>17.75</td>
<td>34.21</td>
<td>43.35</td>
<td>18.90</td>
</tr>
<tr>
<td>eBDTR [14]</td>
<td>34.62</td>
<td>58.96</td>
<td>68.72</td>
<td>33.46</td>
</tr>
<tr>
<td>HSME [17]</td>
<td>50.85</td>
<td>73.36</td>
<td>81.66</td>
<td>47.00</td>
</tr>
<tr>
<td>D2R [12]</td>
<td>43.4</td>
<td>66.1</td>
<td>76.3</td>
<td>44.1</td>
</tr>
<tr>
<td>MAC [18]</td>
<td>36.43</td>
<td>62.36</td>
<td>71.63</td>
<td>37.03</td>
</tr>
<tr>
<td>MSR [16]</td>
<td>48.43</td>
<td>70.32</td>
<td>79.95</td>
<td>48.67</td>
</tr>
<tr>
<td>EDFL [69]</td>
<td>52.58</td>
<td>72.10</td>
<td>81.47</td>
<td>52.98</td>
</tr>
<tr>
<td>AlignGAN [13]</td>
<td>57.9</td>
<td>75.0</td>
<td>82.9</td>
<td>53.6</td>
</tr>
<tr>
<td>HAT (Ours)</td>
<td>71.83</td>
<td>87.16</td>
<td>92.16</td>
<td>67.56</td>
</tr>
</tbody>
</table>

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 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

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