Cross-Modality Person Re-Identification via Modality-aware Collaborative Ensemble Learning

Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen, *Senior Member, IEEE*

***Abstract*—Visible thermal person re-identification (VT-ReID) is a challenging cross-modality pedestrian retrieval problem due to the large intra-class variations and modality discrepancy across different cameras. Existing VT-ReID methods mainly focus on learning cross-modality sharable feature representations by handling the modality-discrepancy in feature level. However, the modality difference in classifier level has received much less attention, resulting in limited discriminability. In this paper, we propose a novel modality-aware collaborative ensemble (MACE) learning method with middle-level sharable two-stream network (MSTN) for VT-ReID, which handles the modality-discrepancy in both feature level and classifier level. In feature level, MSTN achieves much better performance than existing methods by capturing sharable discriminative middle-level features in convolutional layers. In classifier level, we introduce both modality-specific and modality-sharable identity classifiers for two modalities to handle the modality discrepancy. To utilize the complementary information among different classifiers, we propose an ensemble learning scheme to incorporate the modality sharable classifier and the modality specific classifiers. In addition, we introduce a collaborative learning strategy, which regularizes modality-specific identity predictions and the ensemble outputs. Extensive experiments on two cross-modality datasets demonstrate that the proposed method outperforms current state-of-the-art by a large margin, achieving rank-1/mAP accuracy 51.64%/50.11% on the SYSU-MM01 dataset, and 72.37%/69.09% on the RegDB dataset.**

I. Introduction

Person re-identification (Re-ID) is a specific pedestrian retrieval task, which aims at matching person images captured from different non-overlapping cameras [1]–[3]. It has gained increasing attention due to its importance in computer vision research community and practical video surveillance applications [4], [5]. Existing person Re-ID mainly focuses on single-modality module, where all the person images are captured by visible cameras in the daytime. Encouraging performance with deep neural networks has been achieved in both image-based [6]–[9] and video-based person Re-ID tasks [10]–[12], achieving more than 95% rank-1 recognition accuracy in most benchmarks. However, the general visible RGB cameras cannot capture valid appearance information under low-illumination environment, e.g., at night (Fig. 1 (a)). In comparison, many new-generation surveillance cameras can automatically switch to infrared mode to capture the person images at night [13]. Therefore, in this paper, we focus on the cross-modality visible thermal person re-identification (VT-ReID)[1] problem [16], [17], which plays an important role in practical night-time video surveillance applications.

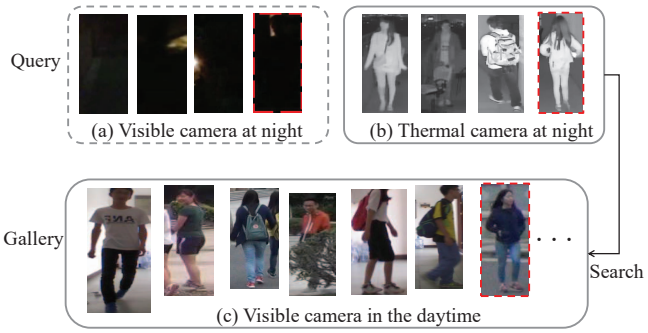


Fig. 1. Illustration of cross-modality visible-thermal person re-identification (VT-ReID). (a) The visible cameras usually cannot capture valid appearance information at night; The infrared(thermal) cameras could capture person images with rich appearance information at night; (c) The gallery images are usually collected by visible cameras in the daytime.

Given a query thermal image captured at night by a thermal/infrared camera, VT-ReID aims at searching out the corresponding visible images from a gallery set captured in the daytime, which represent the same identity as query. An illustration about the VT-ReID problem is shown in Fig. 1. To our best knowledge, VT-ReID has been rarely studied due to the significant visual differences between two modalities caused by modality discrepancy and large intra-class variations. The modality discrepancy is usually caused by different wavelength ranges in different camera modules, which results in huge visual difference in the visual information between two modalities. In addition, different viewpoints, poses and self-occlusions yield very large intra-class variations for cross-modality VT-ReID. In addition, the person images are usually captured in different environments, i.e., indoor or outdoor applications, which brings in additional difficulties. Related cross-modality matching problem has been extensively studied in VIS-NIR face recognition [18], [19]. However, the visual appearance variations of the person images in VT-ReID are much larger than that of face images, which makes their methods less competitive for VT-ReID task [14].

For VT-ReID, several pioneer works have been proposed

M. Ye and J. Shen are with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. E-mail: mangye16@gmail.com, shenjianbingcg@gmail.com. (*Corresponding Author*: Jianbing Shen)

X. Lan is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. E-mail: xiangyuanlan@life.hkbu.edu.hk

Q. Leng is with the Jiujiang University, Jiangxi, China. E-mail: lengqingming@126.com.

[1]We name the person images captured at night with special spectrum cameras (either infrared [14] or thermal [15] cameras) as thermal images.

to address the modality discrepancy and large intra-class variations. A deep zero-padding network is introduced to learn modality-sharable features by adaptively handling the modality input in [14]. cmGAN is proposed in [13] to simultaneously discriminate the identities and modalities with adversarial training. Ye *et al.* introduced a dual-constrained top-ranking loss with a two-stream network in [17], [20]. D^{2}RL [21] is the current state-of-the-art by handling the modality discrepancy in both pixel level and feature level. However, all of them usually learn the cross-modality feature representations with modality-sharable classifier [13], [14], [20], [21]. The modality discrepancy issue in classifier level is not well addressed in their methods, resulting in limited performance for VT-ReID [62]. In addition, previous works usually adopt two-stream network with shared embedding layer to learn modality-sharable features [17], [20], [22], [23], which can only capture high-level sharable information. The discriminative sharable information in middle-level convolutional layers is ignored.

To address above limitations, we introduce modality-aware collaborative ensemble (MACE) learning with a middle-level sharable two-stream network (MSTN) for VT-ReID. Our basic idea is to handle the modality discrepancy in both feature-level and classifier-level. Specifically, MSTN aims at learning modality-aware feature representations with partially shared network structures. The improvement mainly lies in the sharable convolutional blocks to capture discriminative middle-level features, not just high-level features. In classifier level, we introduce both modality-sharable and modality-specific classifiers to guide the feature learning. On one hand, the modality-sharable classifier aims at capturing the sharable information. On the other hand, the modality-specific classifiers learn two separate identity classifiers for two different modalities to handle the modality discrepancy. In addition, we introduce an ensemble learning strategy by combining all the prediction outputs of different classifiers to formulate an enhanced teacher ensemble. To facilitate knowledge transfer among different classifiers, we adopt the knowledge distillation technique introduced in [24] for collaborative learning. It improves the performance by utilizing the relationship between the modality-specific classifiers and the teacher ensemble with a consistency regularization.

The main contributions can be summarized as follows:

- We propose a novel modality-aware collaborative ensemble (MACE) learning method with an improved middle-level sharable two-stream network (MSTN) for cross-modality VT-ReID. We demonstrate that handling modality-discrepancy in both feature level and classifier level consistently is important for VT-ReID. And the proposed MSTN also greatly improves performance of other VT-ReID methods.
- We introduce a collaborative ensemble learning scheme to utilize the relationship among different classifiers. It enhances the discriminability with the ensemble outputs and their consistency.
- We outperform current state-of-the-arts by a large margin on two cross-modality person Re-ID datasets, which greatly accelerates the cross-modality Re-ID research.

A preliminary conference version has been published in [25]. We have made three major improvements in this journal version: Firstly, we present a middle-level sharable two-stream network structure to learn better multi-modality sharable features, which provides a strong baseline for the cross-modality person Re-ID task by learning sharable middle-level convolutional features. It also greatly improves previous state-of-the-art methods. Secondly, we introduce a collaborative ensemble learning strategy to improve the performance by facilitating the knowledge transfer among different classifiers. The ensemble provides a better comprehensive model learning guidance for different classifiers. Meanwhile, the improved method also contains fewer hyper parameters but achieves much better performance than our previous conference version. Finally, more comprehensive analysis is conducted to discuss the superiority and limitations of our proposed method.

II. Related Work

Single-Modality Person Re-ID. Person re-identification (Re-ID) addresses the problem of matching person images across non-overlapping visible cameras [26]. The key challenges of person Re-ID task mainly lie in the large intra-class variation caused by different camera views, poses variations, illuminations changes and occlusions [28], [29]. Existing methods can be categorized into feature learning and metric learning methods. The former feature learning methods mainly focus on discriminative and robust feature representation learning by utilizing the human body structure [30]. The latter metric learning methods usually aim at learning discriminative distance measurements to make sure the positive distance is much smaller the negative distance [31]. Recently, Re-ID works have achieved inspiring performance with the deep end-to-end learning CNN network [32], and some of them have already outperformed the human-level performance on the widely-used datasets [33], [34]. However, most of existing methods are developed for single visible modality module, i.e., the person images are collected by RGB cameras in the day-time under well lighting conditions, and they usually cannot perform well for the night-time cross-modality person Re-ID task [14], which limits applicability in practical surveillance.

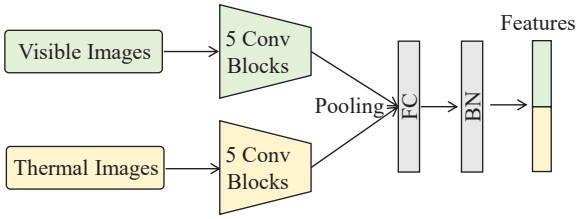
Multi-Modality Person Re-ID. Multi-modality person Re-ID has been extensively studied by combing the multiple modality information to improve the single modality person Re-ID [15], [35], [36]. Specifically, the depth information is adopted to improve the single RGB modality person Re-ID in [35], [36]. In addition, some researchers also try to combine the thermal information to provide additional appearance information under low-lighting conditions in [15]. Considering the semantic attributes as another modality cue, it has been widely used to improve the performance with single visual feature representation [7], [37]–[39]. Usually, combining additional modality information achieves better performance than solely using single modality information. However, it usually require additional cost to collect or mine the multi-modality information. In comparison, this paper mainly focuses on cross-modality person Re-ID. The main target is to match person images across different modalities rather than combing different modality information.

Cross-Modality Person Re-ID. Cross-modality person Re-ID matching person images across different modalities, i.e., text-to-image pedestrian retrieval [40]–[42] or visible-to-thermal matching [14], [21], [27]. Different from the text-to-image retrieval, the modality discrepancy in VT-ReID is totally different. Therefore, the methods designed from text-to-image retrieval are usually unsuitable for our VT-ReID problem.

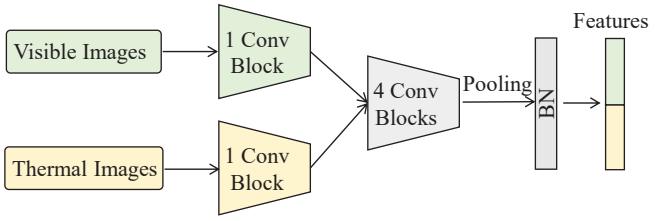
For VT-ReID, a zero-padding strategy with one-stream network is proposed to adaptively learn the cross-modality feature representations in [14]. Later on, a two stream network with dual-constrained top-ranking loss is introduced in [17], [20] to simultaneously handle the cross- and intra-modality variations. Besides, an adversarial learning framework with triplet loss is presented in [13], which jointly discriminates the identity information and the modality information. Recently, a dual-level discrepancy method is proposed to reduce the modality discrepancy in both feature level and image image level [21]. Meanwhile, some other papers also try to investigate a better loss function for this cross-modality person Re-ID task [43], [44]. Most of these methods have ignored the classifier discrepancy in different modalities, which limits their performance [49], [62]. Meanwhile, their baseline networks are not well designed for this task (usually less than 35% rank-1 accuracy on the large-scale SYSU-MM01 dataset). In this paper, we present a two-stream network with modality-aware learning in both feature and classifier level.

Heterogeneous Face Recognition. In a more general perspective, heterogeneous face recognition has been extensively studied in photo-to-sketch [18], [45] and NIR-VIS module [19], [46], [47]. To reduce the modality discrepancy, early research mainly focuses on learning modality-sharable or modality specific metrics or dictionaries [48]. With deep learning, most of them try to learn the modality-sharable feature representations or cross-modality matching models [19], [50]. Compared to the NIR-VIS face recognition problem, VT-ReID also shares the same module by matching visible and thermal images of the same identity [19]. However, VT-ReID suffers from much larger modality difference due to the different camera environments and the visual difference. The modality discrepancy is much more challenging than the NIR-VIS face recognition problem. Therefore, the methods designed for NIR-VIS face recognition usually have limited performance for our cross-modality person Re-ID task [14].

Collaborative Ensemble Learning. Collaborative learning aims at training an improved network with multiple classifiers, where these classifiers collaboratively improve the feature learning performance by using the same network structure [51], [52]. The output predictions of multiple classifiers can provide supplementary information for each other. Inspired by this idea, we propose to formulate a teacher ensemble by combing the outputs of modality-sharable classifiers and modality-specific classifiers. In addition, we introduce collaborative learning scheme to incorporate the the teacher ensemble with the modality-specific classifier output to improve the cross-modality person Re-ID performance.



(a) Two-stream network in [17], [20], [22]



(b) Our middle-level sharable two-stream network (MSTN)

Fig. 2. Comparison between the widely-used two-stream network in [17], [20], [22] and our improved middle-level sharable two-stream network (MSTN). We use the widely-used ResNet50 for illustration.

III. Proposed Method

A. Overview

Our proposed method mainly contains three parts: *1) Feature-Level Modality-aware Learning*, we introduce a middle-level sharable two-stream network for feature learning, which addresses the feature-level discrepancy with partially independent and sharable network structures. *2) Classifier-Level Modality-aware Learning*, we propose a modality-aware classifier learning strategy, which simultaneously uses the modality-sharable and modality-specific classifiers to handle the modality discrepancy in classifier level. *3) Collaborative Ensemble Learning*, we design a collaborative ensemble learning method to facilitate the feature learning by utilizing the relationship among different classifiers. Finally, we will present our overall loss function.

B. Feature-Level Modality-aware Learning

We firstly introduce the feature-level modality-aware learning with an improved two-stream network, termed as MSTN. To simultaneously handle the modality discrepancy and mine modality-sharable information at feature level, we use a two-stream CNN network with partially shared structures for feature learning. Specifically, the network parameters of shallow convolutional layers are specific to capture modality-specific low-level feature patterns. Meanwhile, the network parameters of deep convolutional layers are shared to learn modality-sharable middle-level feature representations. After the convolutional layers with adaptive pooling, a shared batch normalization layer is added to learn the shared feature embedding. Note that the output of shared batch normalization layer is used for the feature representation in testing process. In this manner, MSTN learns modality-sharable middle-level features while capturing the modality-specific low-level information.

Different from the two-stream network used in [17], [20], [22], [23], our proposed MSTN has two main modifications:

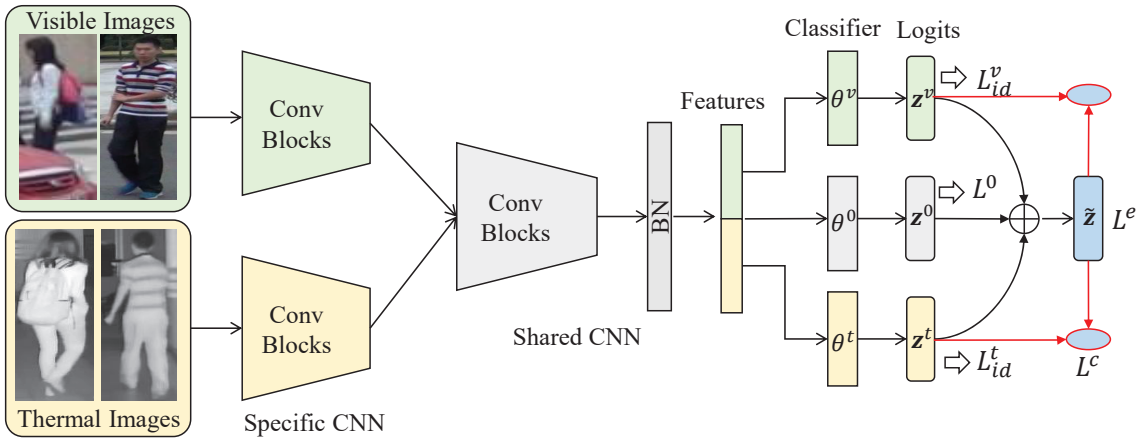


Fig. 3. The framework of our proposed method. The parameters of the first convolutional block are different to address the modality difference in feature level. Meanwhile, we use four shared convolutional blocks and one BN layer to learn modality-sharable middle-level feature. A modality sharable-classifier loss L^0 is adopted to guide the learning process. To further handle the modality discrepancy, we introduce two modality-specific classifier losses ($\mathcal{L}_{id}^v$ and $\mathcal{L}_{id}^t$). To facilitate knowledge transfer among different classifiers, we introduce a collaborative ensemble learning scheme, which contains an ensemble learning loss $\mathcal{L}^e$ and a consistency collaboration loss $\mathcal{L}^c$.

- *Shared convolutional blocks*. In [17], [20], all the network parameters in the convolution blocks are specific. However, this strategy cannot capture the shared middle-level feature patterns in convolutional layers. In comparison, we only utilize one single domain-specific convolutional block[2] to capture the modality-specific information and the left four residual blocks are shared in both modalities. Our improved MSTN learns better features by mining sharable information in middle-level convolutional blocks for cross-modality person Re-ID.
- *Feature embedding layer*. Similar to the Batch Normalization Neck (BNNeck) in introduced in [54] for single-modality person Re-ID, we directly add a batch normalization layer after the pooling layer as the feature embedding for cross-modality person Re-ID. Compared to the two-stream network with another fully connected layer for feature embedding learning in [17], [20], the improved structure also achieves better performance for the cross-modality person Re-ID task.

Experiments in Section IV-C demonstrate that our proposed MSTN has achieved quite competitive performance when configured with a simple baseline learning objective, using a modality-sharable classifier. Note that learning a sharable classifier is also widely used in existing cross-modality person Re-ID [13], [14], [20], [21]. Generally, we define a set of training images by X^v and X^t with identity labels $Y = \{y_i\}$. It contains images from visible modality (denoted by $X^v = \{x_i^v | i = 1, 2, \cdots, N_1\}$) and thermal modality (denoted by $X^t = \{x_i^t | i = 1, 2, \cdots, N_2\}$). N_1 (N_2) represents the number of visible (*thermal*) images in the training set.

We use the combination of triplet loss with hard mining [55] ($\mathcal{L}_{tri}^0$) and softmax identity loss [56] ($\mathcal{L}_{id}^0$) as the baseline learning objective function $\mathcal{L}^0$. Specifically, the triplet loss constrains the feature learning process by utilizing the bi-directional relationship (visible-to-thermal and thermal-to-visible [20]) among different person identities across two modalities. The identity loss aims at learning an identity invariant feature representation by treating the images of each identity captured from two different modalities as the same class. Mathematically, the baseline learning objective with modality-sharable classifier is a combination of two parts:

$$\mathcal{L}^0 = \mathcal{L}_{tri}^0 + \mathcal{L}_{id}^0. \tag{1}$$

The bi-directional triplet loss with hard mining is represented by

$$\begin{aligned}\mathcal{L}_{tri}^0 = &\sum_{i=1}^{n}[\rho + \min_{\forall y_j = y_i} D(f_i^v, f_j^t) - \min_{\forall y_i \neq y_k} D(f_i^v, f_k^t)]_+ \\ &+ \sum_{i=1}^{n}[\rho + \min_{\forall y_j = y_i} D(f_i^t, f_j^v) - \min_{\forall y_i \neq y_k} D(f_i^v, f_k^t)]_+,\end{aligned} \tag{2}$$

where $[\cdot]_+ = \max(\cdot, 0)$, ρ is the margin parameter, n is the number of visible (thermal) samples in each training batch. f_i^v (f_i^t) represents the extracted features of input visible (thermal) image x_i^v (x_i^t), and y_i is the corresponding identity label. $D(\cdot)$ represents the squared Euclidean distance between the extracted features of two samples [55]. Note that we also adopt a bi-directional training strategy as introduced in [20] to enhance the performance, which considers both visible-to-thermal and thermal-to-visible relationships.

The modality-sharable identity classifier learns the feature representation with sharable parameters θ^0 to calculate identity loss for two different modalities [13], [14], [20]. With the modality-sharable classifier θ^0, we calculate the probability $p^0(y_j|x_i^v)$ of a visible sample x_i^v being recognized as identity j. Mathematically, the probability is computed by a softmax function

$$p^0(y_j|x_i^v) = \frac{\exp(z_{i,j}^0)}{\sum_{k=1}^{C}\exp(z_{i,k}^0)}, j = 1, \cdots, C. \tag{3}$$

where $z_{i,j}^0$ represents the output classification logit of an input sample x_i^v being recognized as identity j through the

[2]We adopt the widely-used ResNet50 [53] as the backbone network.

modality-sharable classifier θ^0. C is the total number of identities. Similarly, we calculate the probability of an input thermal sample x_i^t being recognized as identity j, denoted by $p^0(y_j|x_i^t)$. With the calculated probabilities, the modality-sharable identity loss is denoted by

$$\mathcal{L}_{id}^0 = -\frac{1}{n}\sum_{i=1}^{n} log(p^0(y_i|x_i^v)) - \frac{1}{n}\sum_{i=1}^{n} log(p^0(y_i|x_i^t)), \quad (4)$$

where x_i^v (x_i^t) represents the input visible (thermal) image, and y_i is the corresponding label. n is the number of visible (*thermal*) images at each training batch. In our proposed model, we random select n-pair visible-thermal images to construct the batch, where each visible-thermal pair represents the same identity, as introduced in Section III-D.

Note that the modality-sharable classifier learns identity discriminative classifier for two different modalities with the same parameters θ^0, which is also widely used in [13], [14], [20], [21]. This strategy may lose modality-specific information in the classifier level, which cannot well reduce the cross-modality discrepancy. Therefore, it results in less discriminative cross-modality feature representations in the backward propagation learning process.

C. Classifier-Level Modality-aware Learning

To address above issue, we propose a novel modality-specific classifier learning strategy to improve the performance. Our basic idea is that two sets of modality-specific identity classifiers (θ^v for visible modality and θ^t for thermal modality) are learned for two different modalities, as illustrated in Fig. 3. Given the modality-specific identity classifier for visible modality represented by θ^v, the output logits of an input visible image x_i^v are calculated by $\mathbf{z}_i^v$. Correspondingly, we calculate the probability $p^v(y_i|x_i^v)$ of visible sample x_i^v being correctly recognized as identity i with the softmax function. Similar to the modality-sharable identity loss, the identity loss of visible modality-specific classifier is then calculated by

$$\mathcal{L}_{id}^v = -\frac{1}{n}\sum_{i=1}^{n} log(p^v(y_i|x_i^v)). \quad (5)$$

Similarly, we can compute the identity loss of the modality-specific identity classifier θ^t for thermal modality. We denote the corresponding output logits of an input thermal image x_i^t by θ^t. Meanwhile, the probability of thermal sample x_i^t being correctly recognized as identity i is represented by $p^t(y_i|x_i^t)$. The modality-specific identity loss for thermal modality is then calculated by

$$\mathcal{L}_{id}^t = -\frac{1}{n}\sum_{i=1}^{n} log(p^t(y_i|x_i^t)). \quad (6)$$

In summary, we define the modality-specific loss $\mathcal{L}^s$ as a combination of $\mathcal{L}_{id}^v$ and $\mathcal{L}_{id}^t$. Mathematically, it is represented by

$$\mathcal{L}^s = \mathcal{L}_{id}^v + \mathcal{L}_{id}^t. \quad (7)$$

Note that the modality-specific identity classifiers share the same structure with the modality-sharable identity classifier θ^0, but they are optimized separately to capture different modality-specific information in classifier level.

D. Collaborative Ensemble Learning

Above modality-specific classifiers (θ^v, θ^t) and modality-sharable identity classifier (θ^0) share most convolutional layers, but they are optimized separately to learn high-level semantic representations. This learning strategy may lose the complementary information among different classifiers. To address this issue, we introduce a collaborative ensemble learning scheme, which aims at collaboratively optimizing the feature learning with multiple classifiers. Motivated by teacher-ensemble model [52], we take the ensemble of different classifier output to generate an enhanced teacher model for n identities in each batch. We assume that different classifiers contribute equally in the ensemble. For each visible-thermal image pair x_i^v and x_i^t, we calculate the average prediction of all the classifiers as the ensemble $\mathbf{z}_i^e$, which is represented by

$$\mathbf{z}_i^e = \frac{1}{4}(\mathbf{z}_i^{0,1} + \mathbf{z}_i^{0,2} + \mathbf{z}_i^v + \mathbf{z}_i^t), i = 1, 2, \cdots, n, \quad (8)$$

where $\mathbf{z}_i^e$ is a C-dim vector, representing the calculated ensemble of each pair identity $\{x_i^v, x_i^t\}$. $\mathbf{z}_i^{0,1}$ and $\mathbf{z}_i^{0,2}$ represents the output logits of x_i^v and x_i^t with the shared classifier θ^0, respectively. We guide the ensemble training with the cross-entropy loss, which is represented by $\mathcal{L}^e$

$$\mathcal{L}^e = -\frac{1}{n}\sum_{i=1}^{n} log(p^e(y_i|x_i^v, x_i^t)), \quad (9)$$

where $p^e(y_i|x_i^v, x_i^t)$ is calculated with the softmax function in Eq. 4, representing the probability of pair $\{x_i^v, x_i^t\}$ being recognized as identity y_i with the teacher ensemble.

Collaborative Consistency. To facilitate knowledge transfer among different classifiers, we adopt the knowledge distillation technique introduced in [24] for collaborative learning. Following [24], we add a temperature parameter T to smooth the probability distributions for different classifiers. Mathematically, we compute smoothed probability of the teacher ensemble by

$$\tilde{p^e}(y_k|x_i^v, x_i^t) = \frac{\exp(\mathbf{z}_{i,j}^e/T)}{\sum_{k=1}^{C}\exp(\mathbf{z}_{i,k}^e/T)}, j = 1, \cdots, C. \quad (10)$$

Similarly, we could compute the smoothed probability of $\tilde{p^v}(y_k|x_i^v)$ and $\tilde{p^t}(y_k|x_i^t)$. Following [24], we set $T = 3$ in our experiments. Note that T controls the concentration level of the softened distributions [57].

To align the distributions between the modality-specific identity classifier and the teacher ensemble, we adopt the Kullback Leibler divergence to measure the distribution difference. It is formulated by

$$\begin{aligned}\mathcal{L}^c = &\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{C}\tilde{p^e}(y_k|x_i^v, x_i^t)\log\frac{\tilde{p^e}(y_k|x_i^v, x_i^t)}{\tilde{p^v}(y_k|x_i^v)} \\ &+ \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{C}\tilde{p^e}(y_k|x_i^v, x_i^t)\log\frac{\tilde{p^e}(y_k|x_i^v, x_i^t)}{\tilde{p^t}(y_k|x_i^t)}.\end{aligned} \quad (11)$$

n-pair Batch Sampling. This part introduces our n-pair batch sampling training strategy [58] for cross-modality person Re-ID, which is designed to match the rationale of collaborative ensemble learning. In particular, at each training batch,

(a) SYSU-MM01 (b) RegDB

Fig. 4. Sampled visible-thermal image pairs from SYSU-MM01 dataset [14] and RegDB dataset [15]. Each column represents the same identity from two different modalities. Note that night-time images are captured by near-infrared cameras on SYSU-MM01 dataset and by thermal cameras on RegDB dataset, respectively.

we firstly randomly select p person identities, and then we select k visible and k thermal images for each identity to feed into the two-stream network. It is easy to infer that n is equal to $p \times k$. Therefore, n-pair images are fed into the network at each step and each visible-thermal pair represents the same identity from two modalities.

On one hand, in our teacher ensemble, we learn a C-dim ensemble for each visible-thermal pair by combining the outputs of modality-sharable and modality-specific classifiers, and then the ensemble learning loss could guarantee that the ensemble is correctly classified. In this manner, we could learn an enhanced ensemble for each visible-thermal pair by considering the relationship between two images. In a random sampling mechanism, all the possible positive visible-thermal pairs are constrained to be correlated in the ensemble learning process, resulting in better performance. On the other hand, the collaborative consistency loss calculates the difference between the ensemble output and modality-specific classifier output. The collaborative consistency loss aims at transferring the learned information among multiple classifiers, which provides more reliable gradient information in back-propagation process. Experimentally, we demonstrate that both constraints improve the Re-ID performance consistently.

E. Overall Loss Function.

The total loss $\mathcal{L}$ of our modality-aware collaborative ensemble (MACE) learning is then defined by

$$\mathcal{L}^c = \mathcal{L}^0 + \lambda_1 \mathcal{L}^s + \mathcal{L}^e + w(t) \cdot T^2 \mathcal{L}^c, \tag{12}$$

where λ_1 is the coefficient to adjust the contribution of modality-specific classifier loss $\mathcal{L}^s$. Note that the gradient magnitudes of the collaborative consistency loss is scaled by $1/T^2$ due to the temperature T. Therefore, we multiply a factor T^2 for the collaborative consistency loss $\mathcal{L}^c$ to ensure that it shares similar contribution with the ensemble learning loss $\mathcal{L}^e$. $w(t)$ is a ramps up sigmoid function, where the weight value increases from zero to one gradually according to the training epoch t [59]. The main reason is that the initial predictions of different classifiers might be quite different and it is quite difficult to guarantee the predictions are consistent.

IV. Experimental Results

A. Experimental Settings

Datasets and settings. To evaluate our proposed method, we adopt two publicly available cross-modality person Re-ID datasets (SYSU-MM01 [14] and RegDB [15]) for experiments. We also plot some example visible-thermal image pairs randomly sampled from two datasets in Fig. 4.

SYSU-MM01 dataset [14] is a large-scale cross-modality person Re-ID dataset. It is collected by 4 general RGB cameras and 2 near-infrared cameras in SYSU campus. Note that this dataset contains images captured in both indoor and outdoor environment, which makes this dataset extremely challenging. SYSU-MM01 contains 491 person identities, and each identity appears in more than two different modality cameras. Specifically, it contains 395 identities for training and 96 identities for testing. Totally, the training set contains 22,258 visible and 11,909 near-infrared images for 395 identities, which are captured from both indoor and outdoor cameras. For testing, it contains two different evaluation settings, *all-search* mode and *indoor-search* mode. The query set contains 3803 images captured from IR camera 3 and 6 in both settings. The gallery set contains all the visible images captured from four RGB cameras in *all-search* mode, while the *indoor-search* mode only contains the images captured by two indoor cameras. Details description of the evaluation settings is in [14].

RegDB dataset [15] is a small-scale dataset collected by a dual-camera system, including one visible camera and one thermal camera. Totally, this dataset contains 412 person identities, in which each identity has 10 visible and 10 thermal images. Following the cross-modality pedestrian retrieval evaluation protocol in [16]. We randomly select 206 identities for training and the rest 206 identities are used for testing. Following [16], we use the images from visible modality as query and the images from thermal modality as gallery. Naturally, the query set contains 2,060 visible images and the gallery set contains 2,060 thermal images. The average performance of ten times randomly training/testing splits is reported following [16]. Note that we also evaluate the performance by changing the query setting to thermal (query) to visible (gallery).

Evaluation metrics. To evaluate our proposed method and competing methods, we use Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) as the evaluation metrics. CMC measures the matching probability of the groundtruth person occurs in the top-k retrieved results (Rank-k accuracy). mAP is adopted to measure the retrieval performance when multiple matching images occur in the gallery set for a given query image [60].

Implementation details. Our algorithm is implemented on PyTorch framework. Following most existing person Re-ID works, ResNet50 [53] is adopted as our backbone network for cross-modlaity feature learning. The stride of the last convolutional block is set to 1 following [34] to obtain fine-grained feature maps. We initialize the convolutional blocks of our two-stream network with the pre-trained ImageNet parameters, as done in [17]. All the input images are firstly resized to 288×144. We adopt random cropping with zero-padding and horizontal flipping for data argumentation. SGD

TABLE I
EVALUATION OF THE PROPOSED MSTN ON THE LARGE-SCALE SYSU-MM01 DATASET. NOTE THAT BOTH OF THEM UTILIZE THE COMBINATION OF SOFTMAX LOSS AND BI-DIRECTIONAL TRIPLET LOSS [20] AS BASELINE LEARNING OBJECTIVE. 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
Two-Stream [25]	30.67	60.73	74.76	87.57	32.90	33.50	67.28	81.64	93.38	44.83
MSTN (Ours)	45.22	74.22	84.51	93.15	45.79	49.53	78.75	88.56	95.49	58.15
One-Stream (Ours)	44.68	73.85	84.75	92.24	44.91	48.68	78.02	87.62	94.95	57.43

TABLE II
EVALUATION OF EACH COMPONENT ON THE LARGE-SCALE SYSU-MM01 DATASET. "**B**" REPRESENTS THE BASELINE LEARNING OBJECTIVE WITH THE COMBINATION OF IDENTITY LOSS AND TRIPLET LOSS $\mathcal{L}^0$. "**S**" MEANS RESULTS WITH THE MODALITY-SPECIFIC CLASSIFIER LOSS $\mathcal{L}^s$. "**E**" MEANS RESULTS WITH THE ENSEMBLE LEARNING LOSS $\mathcal{L}^e$. '**C**" MEANS THE COLLABORATIVE LEARNING WITH CONSISTENCY REGULARIZATION $\mathcal{L}^c$. 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	45.22	74.22	84.51	93.15	45.79	49.53	78.75	88.56	95.49	58.15
B + S	49.79	78.28	87.77	94.87	48.54	54.66	83.11	91.59	97.25	62.22
B + S + E	50.38	78.52	88.46	95.29	49.68	55.04	84.44	92.50	97.37	62.62
B + S + C	50.48	79.31	88.25	94.87	49.35	56.05	85.26	92.65	97.31	63.98
B + S + C + E	51.64	78.24	87.25	94.44	50.11	57.35	85.67	93.02	97.47	64.79

optimizer is adopted to optimize the network, 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 30 epoch with totally 60 epochs on both datasets. We set the margin parameter ρ in Eq.2 to 0.3, following existing Re-ID methods. By default, we randomly select $p = 8$ identities in each training step, and then random select $k = 4$ visible and $k = 4$ thermal images for each training step in all our experiments. It means that each training batch contains 32 visible and 32 thermal images. The training batch size is 64. We set the weighting coefficient of the modality-aware identity classifier as $\lambda_1 = 5$ on both datasets. The collaborative consistency loss is added to the total loss with a ramps up sigmoid function[3]. Mathematically, it is represented by $w(t) = exp(-5.0 * (1 - \frac{t}{t_m})^2)$, where t is the current epoch number and t_m is set to 100 in our experiments. This function aims at increasing the weights from zero to one gradually [59].

B. Self Evaluation

Evaluation of MSTN. In this subsection, we firstly evaluate the effectiveness of our improved two-stream network, MSTN. We compare our performance with the two-stream network used in [17], [20], [25]. Specifically, both methods use the combination of softmax loss and bi-directional triplet loss [20] as our baseline learning objective. The results on the large-scale SYSU-MM01 dataset are shown in Table I.

We observe that our MSTN achieves much better performance than the widely-used two-stream network baselines [17], [20], [22], [25]. The main improvement is brought by the shared convolutional blocks. The main reason is that the middle-level convolutional blocks usually capture the middle-level features, which is quite important for fine-grained recognition task. Therefore, our MSTN learns sharable middle-level feature representations rather than optimizing them independently for cross-modality person Re-ID. This modification results in significant improvment for VT-ReID task, even outperforming most of the current state-of-the-art [21]. In addition, we also compare the performance with one-stream network under our settings as shown in Table I. Results show that our two-stream network achieves slightly better performance than the one-stream network. This experiment demonstrates that modeling modality discrepancy in feature level is also quite important for VT-ReID.

Evaluation of Each Component. We evaluate the effectiveness of each component on the SYSU-MM01 dataset. The results of adding/removing each component are shown in Table II. Specifically, "**B**" represents the baseline results by combing the identity loss and triplet loss $\mathcal{L}^0$. "**S**" denotes the modality-specific identity loss $\mathcal{L}^s$. "**E**" means the ensemble learning loss $\mathcal{L}^e$. '**C**" demonstrates the collaborative learning loss $\mathcal{L}^c$.

1) Effectiveness of $\mathcal{L}^s$: Compared to the baseline model (**B**), the proposed modality-specific classifier loss (**S**) greatly improves the performance on both query settings. The improvement is about 10% for rank-1 accuracy and 6% for mAP on this large-scale dataset. This experiment demonstrates that handling the modality discrepancy in classifier level is important for VT-ReID. *2) Effectiveness of $\mathcal{L}^e$*: When we further combine the modality-specific loss with the ensemble learning loss, the performance is further improved by about 2% for rank-1 accuracy. It shows the importance of learning a teacher ensemble based on the outputs of different classifiers, which enhances the similarity between the two images in each visible-thermal image pair. *3) Effectiveness of $\mathcal{L}^c$*: We also evaluate the collaborative consistency loss. We observe that facilitating the knowledge transfer between different classifiers also consistently improves the performance. After combing all the terms together, the final performance is further improved, which shows that all these components work well together. Finally, we achieve rank-1/mAP accuracy 51.64%/50.11 for the challenging single-shot all search on SYSU-MM01 dataset. This experiment verifies the effectiveness of the proposed modality-aware collaborative ensemble learning.

[3]Available at https://github.com/benathi/fastswa-semi-sup

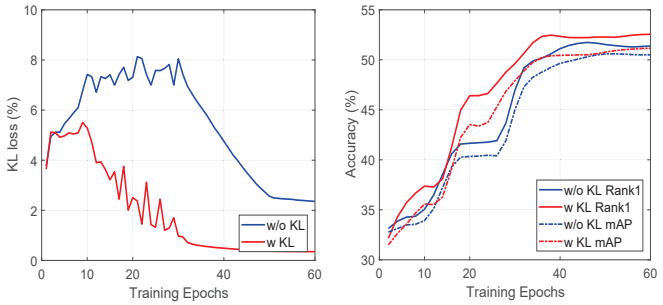


Fig. 5. Evaluation of the collaborative consistency loss $\mathcal{L}^c$ on the large-scale SYSU-MM01 dataset (*single-shot all search mode*). We calculate the KL loss between the teacher ensemble and the modality-specific classifier output at different epochs (*left*). We also report the person Re-ID performance at different epochs (*right*).

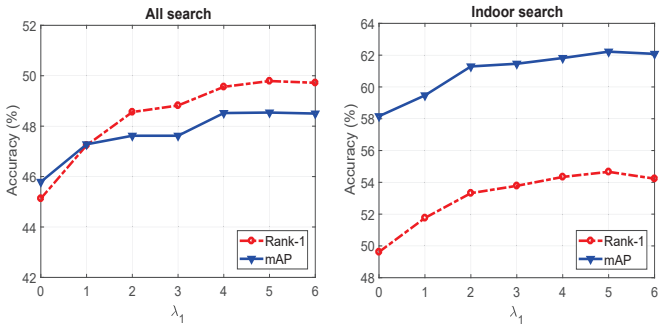


Fig. 6. Evaluation of the weighting parameter λ_1 SYSU-MM01 dataset (*single-shot all search/indoor search mode*). Note that we only combine the baseline loss $\mathcal{L}^0$ and the evaluated component ($\mathcal{L}^s$ for demonstration. Rank-1 matching accuracy (%) and mAP (%) are reported.

TABLE III
EVALUATION OF THE PROPOSED MACE WITH THE ONE-STREAM NETWORK BASELINE ON THE LARGE-SCALE SYSU-MM01 DATASET. RANK AT r MATCHING ACCURACY(%) AND MAP (%) ARE REPORTED.

Datasets	*All Search*		*Indoor Search*	
Methods	$r = 1$	mAP	$r = 1$	mAP
without ramp	50.42	49.48	55.24	62.84
Full Model	51.64	50.11	57.35	64.79

Analysis of Collaborative Consistency Loss. We also evaluate the collaborative consistency loss on the large-scale SYSU-MM01 dataset under the challenging single-shot all search mode (*w KL loss* or *w/o KL loss*). The results are shown in Fig. 5. We calculate the KL loss between the teacher ensemble and the modality-specific classifier output at different epochs (*left*). We also report the cross-modality person Re-ID performance at different epochs (*right*).

Results shown in Fig. 5 demonstrate that the divergence between the teacher ensemble and the modality-specific classier outputs is very large if without the collaborative consistency loss $\mathcal{L}^c$. Specifically, the KL divergence drops dramatically when combined with $\mathcal{L}^c$. Meanwhile, we observe that the rank-1 accuracy and mAP at different epochs also perform better than the baseline results. It also achieves faster learning speed in terms of the VT-ReID accuracy. This experiment verifies the idea to constrain the consistency between different classifiers. It facilitates the knowledge transfer among different classifiers, which is similar to knowledge distillation [24].

Analysis of Ramps Up Sigmoid Function. We evaluate the effect of the ramp up function by simply setting the weight of the collaborative consistency loss as 1. The results are shown in Table III. We observe that the performance is slightly lower than our full model. Meanwhile, it performs closely to the results when this loss is not included in the model. The reason is that the predictions of different classifiers at the early stage are different due to the random initialization of the classifier weights. Forcibly including this constraint too early may lead a trivial solution, *i.e.,*, all the classifiers model the same information. Thus we progressively add this constraint in the overall learning process, ensuring that modality-specific information is captured by the classifiers at the early stage.

Parameter Analysis. We also evaluate the weighting parameters λ_1 of modality-specific classifier loss $\mathcal{L}^s$ in the proposed method. Note that we only have one hyper-parameter in our collaborative ensemble learning method, which is suitable for real applications. Specifically, we only adopt the baseline loss $\mathcal{L}^0$ to evaluate the performance to better illustrate the influence. The rank-1 accuracy and mAP on the large-scale SYSU-MM01 dataset with different λ_1 are reported in Fig. 6.

Fig. 6 demonstrates that integrating $\mathcal{L}^s$ with $\mathcal{L}^0$ consistently improves the cross-modality person Re-ID performance. The improvement is obvious under both query settings. This experiment verifies the importance of addressing the modality discrepancy in classifier level. We also observe that we achieve the best performance when λ_1 is close to 5. When we keep on increasing λ_1, the performance is almost unchanged.

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

Method	$r = 1$	$r = 10$	$r = 20$	mAP
Setting	*Visible to Thermal*			
TONE + HCML [16]	24.44	47.53	56.78	20.08
Zero-Padding [14]	17.75	34.21	44.35	18.90
BDTR [20]	33.56	58.61	67.43	32.76
eBDTR [17]	34.62	58.96	68.72	33.46
HSME [23]	50.85	73.36	81.66	47.00
D^{2}RL [21]	43.4	66.1	76.3	44.1
MAC [25]	36.43	62.36	71.63	37.03
EDFL†[22]	52.58	72.10	81.47	52.98
MSR [62]	48.43	70.32	79.95	48.67
DFE [61]	70.13	86.32	91.96	69.14
MACE (Ours)	**72.37**	**88.40**	**93.59**	**69.09**
Setting	*Thermal to Visible*			
TONE + HCML [16]	21.70	45.02	55.58	22.24
BDTR [20]	32.92	58.46	68.43	31.96
Zero-Padding [14]	16.63	34.68	44.25	17.82
eBDTR [17]	34.21	58.74	68.64	32.49
HSME [23]	50.15	72.40	81.07	46.16
MAC [25]	36.20	61.68	70.99	36.63
EDFL†[22]	51.89	72.09	81.04	52.13
DFE [61]	67.99	85.56	91.41	66.70
MACE (Ours)	**72.12**	**88.07**	**93.07**	**68.57**

† Arxiv papers, not yet published.

C. Comparison with the State-of-the-arts

In this subsection, we compare our proposed method (MACE) with the state-of-the-art methods on two different

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE SYSU-MM01 DATASET. ACCURACY(%) AT RANK r AND MAP (%).

Datasets	*All Search*				*Indoor Search*			
Methods	$r=1$	$r=10$	$r=20$	mAP	$r=1$	$r=10$	$r=20$	mAP
HOG	2.76	18.25	31.91	4.24	3.22	24.7	44.6	7.25
MLBP	2.12	16.23	28.32	3.86	3.43	26.42	45.36	7.72
LOMO [30]	1.75	14.14	26.63	3.48	2.24	22.52	41.53	6.64
GSM [63]	5.29	33.71	52.95	8.00	9.46	48.98	72.06	15.57
One-stream [14]	12.04	49.68	66.74	13.67	16.94	63.55	82.10	22.95
Two-stream [14]	11.65	47.99	65.50	12.85	15.60	61.18	81.02	21.49
Zero-Padding [14]	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
TONE [16]	12.52	50.72	68.60	14.42	20.82	68.86	84.46	26.38
HCML [16]	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
cmGAN [13]	26.97	67.51	80.56	31.49	31.63	77.23	89.18	42.19
BDTR [17]	27.32	66.96	81.07	27.32	31.92	77.18	89.28	41.86
eBDTR [17]	27.82	67.34	81.34	28.42	32.46	77.42	89.62	42.46
HSME [23]	20.68	32.74	77.95	23.12	-	-	-	-
D^2RL [21]	28.9	70.6	82.4	29.2	-	-	-	-
MAC [25]	33.26	79.04	90.09	36.22	36.43	62.36	71.63	37.03
EDFL†[22]	36.94	84.52	93.22	40.77	-	-	-	-
HPILN† [44]	41.36	84.78	94.31	42.95	45.77	91.82	**98.46**	56.52
LZM† [43]	45.00	**89.06**	-	45.94	49.66	92.47	-	59.81
MSR [62]	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
DFE [61]	48.71	88.86	95.27	48.59	52.25	89.86	95.85	59.68
MACE (Ours)	**51.64**	87.25	**94.44**	**50.11**	**57.35**	**93.02**	97.47	**64.79**

† Arxiv papers, not yet published.

datasets. All existing cross-modality VT-ReID methods are included for comparison, including the one-stream **Zero-Padding** [14] network in ICCV 2017, **TONE + HCML** [16] with two-stage learning in AAAI 2018, bi-directional dual-constrained top-ranking (**BDTR**) [20] in IJCAI 2018, center-constrained top-ranking (**eBDTR**) [17] in TIFS 2019, cross-modality GAN (**cmGAN**) [13] in IJCAI 2018, Hypersphere Manifold Embedding (**HSME**) [23] in AAAI 2019, dual-level discrepancy learning (**D^{2}RL**) [21] in CVPR 2019, modality-aware collaborative learning (**MAC** [25]) in ACM MM2019, (**DFE** [61]) in ACM MM2019 and (**MSR** [62]) in TIP 2020. In addition, we also compare some unpublished arXiv papers, including EDFL [22], HPILN [44] and LZM [43]. Note that the numbers of these methods are all taken from their original papers. The results are shown in Table IV and V.

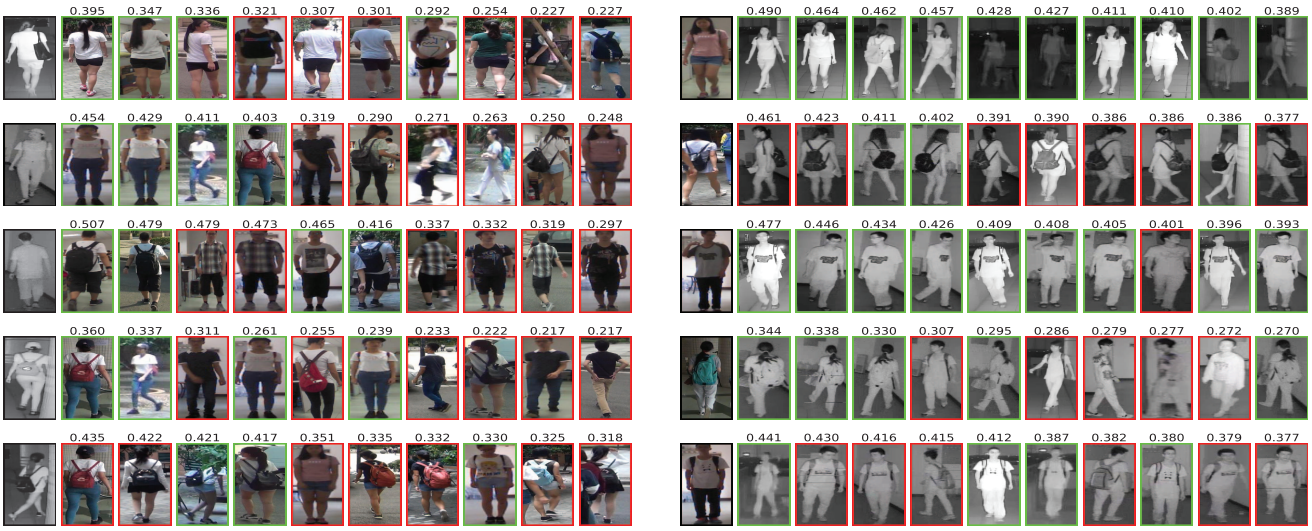
We have the following observations in Table IV and V: 1) Compared to traditional hand-crafted features learning methods, we achieve much better performance for the cross-modality person Re-ID. The main reason is that the domain knowledge in cross-modality person Re-ID is ignored in their methods. Meanwhile, we observe that deep learning also performs much better than the hand-crafted features and dictionary/metric learning methods [16], [63]. 2) Our proposed method outperforms the current state-of-the-art cross-modality person Re-ID method D^{2}RL [21] by a large margin on both datasets. Note that D^{2}RL [21] needs to generate visible-to-thermal and thermal-to-visible images for unified training, and the augmented training set is three times larger than the original dataset. Meanwhile, their testing process also needs to generate more images for feature extraction. In comparison, our proposed method does not need any cumbersome image generation process but achieves much better performance. In addition, compared to the cmGAN method [13], it trains *more than 2,000* epochs for adversarial modality discrimination to achieve good performance, while our proposed method only needs 60 epochs. We achieve much better performance in a more efferent and simpler way. Compared to MSR [62], which also adopts a similar idea with modality-specific classifier learning, we achieve much higher accuracy than MSR in most settings. The comparison demonstrates the effectiveness of our proposed MACE method, which is more suitable for real applications. This experiment shows the superiority by simultaneously handling the modality discrepancy in feature level and classifier level with modality-aware collaborative ensemble learning. 3) In addition, we also observe that our MSTN achieves quite competitive performance when only configured with the baseline learning objective, outperforming most counterparts.

In addition, we find that our proposed method performs much better than the other counterparts on the RegDB dataset, usually about 40% improvement for the rank-1 accuracy. Table V also demonstrates that MACE is robust to different query settings. The main reason is that we can learn much better modality-sharable middle-level feature representations with our proposed framework.

D. Further Analysis

Retrieved Examples. We also visualize some retrieved results on the large-scale SYSU-MM01 dataset. Two different searching modes are demonstrated: thermal to visible search and visible to thermal search. For each query setting, five query samples are randomly selected and their corresponding top ten retrieved cross-modality results are visualized in Fig. 7. Note that we use the all-search gallery for visualization. Meanwhile, we also report the cosine similarity scores.

The results demonstrate that our method method can get good retrieval results when the person appearance has rich structure information (e.g., bags or stripes) or conspicuous part (e.g., logo). This observation is also consistent with the cross-modality person Re-ID task since the thermal images at



(a) Thermal to visible (b) Visible to thermal

Fig. 7. Retrieved results visualization. Two different query settings: visible to thermal and thermal to visible. For each setting, we randomly select five query examples and visualize their corresponding top-10 retrieved results (*All-search Mode*) from SYSU-MM01 dataset. Corrected retrieved samples are in green boxes and wrong matchings are in red boxes (best viewed in color.)

TABLE VI
EVALUATION OF THE PROPOSED MSTN CONFIGURED WITH PEER METHODS ON THE LARGE-SCALE SYSU-MM01 DATASET. RANK AT r MATCHING ACCURACY(%) AND MAP (%) ARE REPORTED.

Settings	*All Search*					*Indoor Search*				
Methods	$r=1$	$r=5$	$r=10$	$r=20$	mAP	$r=1$	$r=5$	$r=10$	$r=20$	mAP
eBDTR [17]	27.82	58.32	67.34	81.34	28.42	32.46	66.72	77.42	89.62	42.46
eBDTR (Ours)	47.32	76.24	87.02	94.26	47.92	51.26	81.32	91.22	96.73	60.22
Improvements	**70.09**↑	**30.73**↑	**29.22**↑	**15.88**↑	**68.61**↑	**57.92**↑	**21.88**↑	**17.82**↑	**7.93**↑	**41.83**↑
MAC [25]	33.26	65.10	79.04	90.09	36.22	33.37	67.02	82.49	93.69	44.95
MAC (Ours)	50.08	78.46	88.04	94.16	48.96	55.06	84.65	92.32	97.02	62.72
Improvements	**50.57**↑	**20.52**↑	**11.39**↑	**4.52**↑	**35.17**↑	**65.00**↑	**26.31**↑	**11.92**↑	**3.65**↑	**39.53**↑

night cannot capture the color information but they preserve rich texture information. Interestingly, we find that when some persons changing their clothes (e.g., 3rd example in the left), we can still get the correct results by mining the discriminative visual cues, maybe the T-shirt or the shorts. Another interesting observation is that using the visible-to-thermal query setting usually performs better than that of thermal-to-visible query setting. The main reason is that using the visible images as query provides richer appearance information for the query image, which is useful for cross-modality person Re-ID.

However, there are still many errors and the performance is still far from the requirements in real applications for cross-modality person Re-ID. In addition, we observe that the similarity scores are distributed differently for different query examples, it would be interesting to study how to define a similarity threshold for the VT-ReID problem.

MSTN for Other Methods. In this subsection, we evaluate the performance of two state-of-the-art methods when configured with our proposed MSTN. Two methods are selected for evaluation, including eBDTR [17] and MAC [25]. The results on the large-scale SYSU-MM01 dataset under both query settings are shown in Table VI.

We observe that both methods are significantly improved when using our MSTN as the backbone network. We achieve nearly 50%-70% rank-1 accuracy improvement under both query settings on the large-scale SYSU-MM01 dataset. We can draw two important conclusions according to the results in Table VI: 1) The middle-level sharable features play an important role in cross-modality person Re-ID to bridge the modality gap. By learning middle-level sharable feature representations with MSTN, we can achieve much better VT-ReID performance than learning high-level sharable features in the final embedding layers. 2) Our improved MSTN works well for other counterparts, which provides an important insight for researchers in this field. It can greatly accelerate the cross-modality Re-ID research, which is very important for practical person Re-ID applications.

Cross-Dataset Evaluation. In this subsection, we conduct the cross-dataset evaluation experiments, which is ignored in previous cross-modality person Re-ID works. Specifically, we use the trained model on the large-scale SYSU-MM01 dataset and test it on the small-scale RegDB dataset. We evaluate our baseline method and the proposed MACE method. The results under two different query settings are shown in Table VI.

Although we find that we perform better than the baseline method, a thought-provoking observation is that the performance drops dramatically under the cross-dataset evaluation setting. The main reason is that these two datasets use different light spectrums to capture the night-time person images (infrared camera on the SYSU-MM01 dataset and thermal camera

TABLE VII
CROSS-DATASET EVALUATION. THE MODELS ARE TRAINED ON SYSU-MM01 DATASET AND TESTED ON THE REGDB DATASET. RANK AT r MATCHING ACCURACY(%) AND MAP (%) ARE REPORTED.

Settings	*Visible to Thermal*					*Thermal to Visible*				
Methods	$r=1$	$r=5$	$r=10$	$r=20$	mAP	$r=1$	$r=5$	$r=10$	$r=20$	mAP
Baseline	3.12	7.01	10.31	15.22	4.29	2.72	5.66	8.33	12.24	4.11
MAC [25]	3.42	7.89	11.23	17.62	4.62	3.28	7.62	10.75	16.32	4.78
MACE (Ours)	**4.43**	**8.99**	**12.96**	**19.09**	**5.57**	**4.44**	**9.13**	**12.63**	**19.07**	**5.30**

TABLE VIII
EVALUATION OF THE PROPOSED MSTN ON THE LARGE-SCALE SYSU-MM01 DATASET WITH DIFFERENT BACKBONE NETWORKS. NOTE THAT ALL OF THEM UTILIZE THE COMBINATION OF SOFTMAX LOSS AND BI-DIRECTIONAL TRIPLET LOSS AS BASELINE LEARNING OBJECTIVE. RANK AT r MATCHING ACCURACY(%) AND MAP (%) ARE REPORTED.

Settings	*All Search*					*Indoor Search*				
Methods	$r=1$	$r=5$	$r=10$	$r=20$	mAP	$r=1$	$r=5$	$r=10$	$r=20$	mAP
Baseline [25] (AlexNet)	20.42	56.48	62.42	76.42	19.88	28.84	60.52	73.42	88.23	34.52
MSTN (AlexNet)	30.32	61.22	73.58	86.72	31.48	32.88	66.82	80.38	92.02	41.26
MACE (AlexNet)	38.43	65.12	76.42	88.14	36.11	42.62	71.03	84.42	94.28	48.42
Baseline [25] (ResNet50)	30.67	60.73	74.76	87.57	32.90	33.50	67.28	81.64	93.38	44.83
MSTN (ResNet50)	45.22	74.22	84.51	93.15	45.79	49.53	78.75	88.56	95.49	58.15
MACE (ResNet50)	51.64	78.24	87.25	94.44	50.11	57.35	85.67	93.02	97.47	64.79
Baseline [25] (DenseNet121)	31.82	61.78	75.84	88.47	33.42	34.56	69.02	82.42	93.84	46.51
MSTN (DenseNet121)	46.52	76.03	85.82	93.08	46.92	50.42	79.63	89.42	95.28	57.72
MACE (DenseNet121)	52.72	78.92	88.03	94.28	52.08	58.48	86.73	93.28	97.06	66.32
One-stream Comparison										
One-Stream (Baseline)	44.68	73.85	84.75	92.24	44.91	48.68	78.02	87.62	94.95	57.43
One-Stream (MACE)	49.63	77.12	86.54	93.16	48.21	54.26	83.01	91.42	96.76	61.82
MSTN (MACE)	51.64	78.24	87.25	94.44	50.11	57.35	85.67	93.02	97.47	64.79

on the RegDB dataset), and the collected night-time person images are totally visually different. However, this would be the practical scenario in real application by applying a trained model in different environments, but importance of this cross-dataset evaluation setting is ignored in previous works. Our observation in this experiment provides an important but unexplored direction for future research in VT-ReID.

Results with Different Backbone Networks. We evaluate the proposed MSTN and MACE by applying AlexNet, ResNet50 and DenseNet121 as backbones on the large-scale SYSU-MM01 dataset. Similar to our design for ResNet50, the first convolutional block (convolutional layer for AlexNet, Dense block for DenseNet121) is specific for modality-specific feature learning, while the rest layers are shared for modality-sharable feature learning. Other training parameters are exactly the same with ResNet50 as described in Sec IV-A. The results are shown in Table VIII. We also include the baseline performance by using the two-stream network in [25]. We observe that the proposed methods (MSTN and MACE) perform well in improving the accuracy on different backbone networks, and consistently outperform the two-stream network structure in [25]. This experiment further verifies the flexibility of our proposed method for different backbone networks.

MACE with One-stream Network Baseline. We have applied our method to the one-stream network, and the results are shown in Table VIII. We observe that the baseline performance of one-stream network is also consistently improved by our MACE method. However, the improvement is not as significant as our proposed MSTN network backbone. The main reason is that the one-stream network provides limited ability to mine the modality-specific information in feature level, since all the network parameters are the same for both modalities to extract the features. In comparison, our method addresses the modality discrepancy in both feature level and classifier level under a collaborative ensemble learning framework, resulting in better performance.

V. CONCLUSIONS

In this paper, we propose a modality-aware collaborative ensemble learning (MACE) method with an improved middle-level sharable two-stream network (MSTN) for cross-modality VT-ReID. We firstly introduce MSTN for modality-aware feature learning, which learns modality-sharable features in middle-level convolutional layers. It achieves much better performance compared than current state-of-the-arts when using a simple combination of softmax loss and triplet loss. Experiments also demonstrate that our proposed MSTN also greatly improves the performance of other methods. Besides the feature-level modality discrepancy with MSTN, we also propose to handle the modality difference in classifier-level. Extensive experiments demonstrate that the modality-specific classifier is essential for a good cross-modality person Re-ID system. To incorporate different classifiers, we introduce a collaborative ensemble learning scheme to further improve the performance. By facilitating the knowledge transfer among different classifiers, we outperform the state-of-the-arts by a large margin on two public cross-modality person Re-ID datasets. It provides new insights and greatly accelerates the cross-modality Re-ID research, which is very important for real applications.

REFERENCES

[1] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2019.

[2] H.-X. Yu, A. Wu, and W.-S. Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[3] M. Ye and J. Shen, "Probabilistic Structural Latent Representation for Unsupervised Embedding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] X. Dong, J. Shen, W. Wang, L. Shao, H. Ling, and F. Porikli, "Dynamical hyperparameter optimization via deep reinforcement learning in tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[5] W. Wang, J. Shen, X. Lu, S. C. Hoi, and H. Ling, "Paying attention to video object pattern understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[6] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2530–2539.

[7] W. Jingya, Z. Xiatian, G. Shaogang, and L. Wei, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2275–2284.

[8] M. Ye and P. C. Yuen, "Purifynet: A robust person re-identification model with noisy labels," *IEEE Transactions on Information Forensics and Security (TIFS)*, 2020.

[9] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 7, pp. 3472–3483, 2018.

[10] 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 *European Conference on Computer Vision (ECCV)*, 2016, pp. 868–884.

[11] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 6, pp. 2976–2990, 2019.

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

[13] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training." in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 677–683.

[14] A. Wu, W.-s. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5380–5389.

[15] 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.

[16] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[17] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Transactions on Information Forensics and Security (TIFS)*, 2019.

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

[19] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[20] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 1092–1099.

[21] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, no. 3, 2019, p. 4.

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

[23] Y. Hao, N. Wang, J. Li, and X. Gao, "Hsme: Hypersphere manifold embedding for visible thermal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 8385–8392.

[24] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

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

[26] 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.

[27] M. Ye, J. Shen and L. Shao, "Visible-Infrared Person Re-Identification via Homogeneous Augmented Tri-Modal Learning," *IEEE Transactions on Information Forensics and Security (TIFS)*, 2020.

[28] Z. Liang and J. Shen, "Local semantic siamese networks for fast tracking," *IEEE Transactions on Image Processing (TIP)*, 2019.

[29] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing (TIP), volume=27, number=5, pages=2368–2378, year=2017,*.

[30] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2197–2206.

[31] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3685–3693.

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

[33] 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.

[34] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *European Conference on Computer Vision (ECCV)*, 2018.

[35] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with rgb-d sensors," in *European Conference on Computer Vision Workshops (ECCVW)*, 2012, pp. 433–442.

[36] A. Wu, W.-S. Zheng, and J.-H. Lai, "Robust depth-based person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 6, pp. 2588–2603, 2017.

[37] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *arXiv preprint arXiv:1703.07220*, 2017.

[38] Z. Wang, X. Bai, M. Ye, and S. Satoh, "Incremental deep hidden attribute learning," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 72–80.

[39] V. Ponce-López, T. Burghardt, S. Hannunna, D. Damen, A. Masullo, and M. Mirmehdi, "Semantically selective augmentation for deep compact person re-identification," in *European Conference on Computer Vision Workshops (ECCVW)*, 2018, pp. 551–561.

[40] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1890–1899.

[41] M. Ye, C. Liang, Z. Wang, Q. Leng, J. Chen, and J. Liu, "Specific person retrieval via incomplete text description," in *International Conference on Multimedia Retrieval (ICMR)*, 2015, pp. 547–550.

[42] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 54–70.

[43] 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.

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

[45] N. Wang, X. Gao, L. Sun, and J. Li, "Bayesian face sketch synthesis," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 3, pp. 1264–1274, 2017.

[46] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for nir-vis face recognition." in *AAAI Conference on Artificial Intelligence (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 Transactions on Image Processing (TIP)*, 2019.

[48] 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 Transactions on Cybernetics (TCYB)*, vol. 47, no. 2, pp. 449–460, 2017.

[49] M. Ye, Y. Cheng, X. Lan, and H. Zhu, "Improving night-time pedestrian retrieval with distribution alignment and contextual distance," *IEEE Transactions on Industrial Informatics (TII)*, 2019.

[50] 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.
[51] G. Song and W. Chai, "Collaborative learning for deep neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 1837–1846.
[52] X. Lan, X. Zhu, and S. Gong, "Knowledge distillation by on-the-fly native ensemble," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 7528–7538.
[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[54] 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.
[55] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.
[56] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1367–1376.
[57] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6210–6219, 2019.
[58] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2016, pp. 1857–1865.
[59] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.
[60] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124.
[61] Y. Hao, N. Wang, X. Gao, J. Li, and X. Wang, "Dual-alignment feature embedding for cross-modality person re-identification," in *ACM Multimedia (ACM MM)*, 2019, pp. 57–65.
[62] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 579–590, 2020.
[63] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 6, pp. 1089–1102, 2017.

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

Xiangyuan Lan received the B.Eng. degree in computer science and technology from the South China University of Technology, China, in 2012, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong in 2016. He is currently a Research Assistant Professor with Hong Kong Baptist University. His current research interests include intelligent video surveillance and biometric security.

Qingming Leng received the B.S degree in life science from Nanchang University, Nanchang, China, in 2007, M.S degree in International School of Software from Wuhan University, Wuhan, China, in 2009, and the Ph.D degree in National Engineering Research Center for Multimedia Software from Wuhan University, Wuhan, China, in 2014. He is currently working as a lecturer at School of Information Science and Technology, Jiujiang University, China. His research interests include person re-identification, image retrieval and machine learning.

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, Beijing, China. He has published more than 100 journal and conference papers, eight 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*, *Neurocomputing* and other journals.