Deep Multi-Patch Matching Network for Visible Thermal Person Re-Identification

Pingyu Wang, Zhicheng Zhao, Fei Su, Yanyun Zhao, Haiying Wang, Lei Yang, and Yang Li

***Abstract*—*Visible Thermal Person Re-Identification* (VTReID) is a cross-modality retrieval problem in computer vision. Accurate VTReID is very challenging due to large modality discrepancies. In this work, we design a novel *Multi-Patch Matching Network* (MPMN) framework to simultaneously mitigate the heterogeneity of coarse-grained and fine-grained visual semantics. In view of cross-modality matching, we verify that aligning modality distributions of the original features is likely to suffer from the selective alignment behavior, i.e., only focuses on easiest dimensions or subspaces. Inspired by adversarial learning, we propose a new *Multi-Patch Modality Alignment* (MPMA) loss to jointly balance and reduce the modality discrepancies of multi-patch features by mining hard subspaces and abandoning easy subspaces. Since multi-patch features are potentially complementary to each other, the semantic correlations between different patches should be exploited during training. Motivated by knowledge distillation, we put forward a new *Cross-Patch Correlation Distillation* (CPCD) loss to transfer the semantic knowledges across different patches. To balance multi-patch tasks, an effective *Patch-Aware Priority Attention* (PAPA) method is further introduced to dynamically prioritize hard patch tasks during training. This paper experimentally demonstrates the effectiveness of the proposed methods, achieving superior performance over the state-of-the-art methods on RegDB and SYSU-MM01 datasets.**

***Index Terms*—Multi-Patch Matching, Person Re-Identification, Modality Alignment, Correlation Distillation, Priority Attention.**

I. Introduction

PERSON Re-Identification (ReID) aims at matching person images of the same person across disjoint cameras. The key challenge lies in that person images usually suffer from significant intra-class variations, *e.g.*, views, poses and backgrounds. In order to address those issues, prior ReID works have broadly followed two main paradigms, i.e., representation learning [1]–[8] or metric learning [9]–[12]. They mainly concentrate on the *Visible Person Re-Identification* (VReID) task, where both training and testing images are captured by visible cameras. However, the VReID models might be out of service in night time or dark environment because visible cameras are unable to capture effective and reliable visual information under poor illuminations. In this case, thermal cameras that do not rely on visible light should be applied, which makes heterogenous person re-identification significant for public surveillance applications.

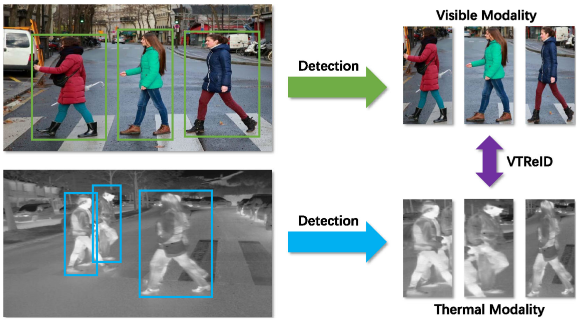


Fig. 1. The pipeline of visible thermal person re-identification in the real-world surveillance scenarios.

Against this issue, *Visible Thermal Person Re-Identification* (VTReID) is introduced for person matching between visible and thermal cameras, which is essentially a cross-modality retrieval problem and widely encountered in night-time surveillance scenarios. As shown in Fig. 1, the query and gallery images are captured by different modality cameras. Since the visible and thermal cameras operate under heterogenous illumination environments, the modality discrepancy becomes the key challenge of the VTReID task. Some VTReID works [13]–[22] have achieved remarkable cross-modality retrieval performances based on *Convolutional Neural Network* (CNN). They usually construct two separate stream networks for visible and thermal modalities to extract modality-specific features. Nevertheless, modality-unshared networks significantly enlarge the inference latencies and computational resources, with the comparison to modality-shared networks. Furthermore, existing VTReID works only consider the modality discrepancies of global coarse-grained features, but lose sight of the modality gaps of local fine-grained features. Therefore, although they are able to learn global modality-robust features, there is certainly no guarantee that local person features are robust to modality discrepancies, resulting in poor cross-modality generalization capability of VTReID models.

Manuscript received December 30, 2019; revised April 25, 2020 and May 19, 2020; accepted May 20, 2020. Date of publication June 1, 2020; date of current version May 26, 2021. This work was supported in part by the Chinese National Natural Science Foundation under Grants 61532018 and U1931202 and in part by MoE-CMCC "Artifical Intelligence" Project No. MCM20190701. The associate editor coordinating the review of this manuscript and approving it for publication was Lei Zhang. *(Corresponding author: Fei Su.)*

Pingyu Wang, Zhicheng Zhao, Fei Su, Yanyun Zhao, and Haiying Wang are with the Beijing Key Laboratory of Network System and Network Culture, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: applewangpingyu@gmail.com; zhaozc@bupt.edu.cn; sufei@bupt.edu.cn; zyy@bupt.edu.cn; why@bupt.edu.cn).

Lei Yang and Yang Li are with the China Mobile Research Institute, Beijing 100053, China (e-mail: yangleiyj@chinamobile.com; liyangyjy@chinamobile.com).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.2999180

In this work, we propose a novel VTReID framework named *Multi-Patch Matching Network* (MPMN) to simultaneously mitigate the coarse-grained and fine-grained heterogeneity of cross-modality person re-identification. For measuring the modality differences, one straightforward way is to leverage the Euclidean distance between visible and thermal features from the same person identity. While directly minimizing this Euclidean distance can largely relieve the modality discrepancies, the learned VTReID models are likely to suffer from the selective alignment behavior, i.e., only focuses on certain dimensions or subspaces which are the easiest ones to reduce the current modality gaps. Accordingly, the hard feature subspace may still have large modality discrepancies and the easy feature subspace exhibits the small modality discrepancies, which causes the imbalance of modality discrepancies. Inspired by adversarial learning [23], an effective *Multi-Patch Modality Alignment* (MPMA) loss is proposed to mitigate the selective alignment behavior by mining the hard subspace and abandoning the easy subspace. Specifically, we pay more attention on reducing modality discrepancies of the hard subspace than that of the easy one. The proposed MPMA loss is able to jointly align the modality distributions of coarse-grained and fine-grained features, which greatly enhances cross-modality generalization capability.

Aside from the modality alignment, we also consider the semantic correlations between different patch features. Generally, coarse-grained patch features with global person knowledges (*e.g.*, body, gender, clothes and trousers) are likely to be invariant to pose and occlusion but encode less discriminative identity information, while fine-grained patch features with local person knowledges (*e.g.*, bag, backpack, hair and face) encode discriminative identity information but are likely to be sensitive to pose and occlusion. Overall, coarse-grained features are robust but less discriminative, while fine-grained features are discriminative but less robust. Thus, we put forward a novel loss function called *Cross-Patch Correlation Distillation* (CPCD) loss to transfer the semantic knowledges of one patch to another patch by directly enforcing the cross-patch similarity correlations in training. In this way, we can utilize the complementary of coarse-grained and fine-grained patch features to boost the cross-modality representations.

In view of multi-patch feature learning, the imbalance of patch tasks can lead to unnecessary emphasis on easier patch tasks, thus neglecting and slowing progress on hard patch tasks. For patch balance learning, it is crucial to mine the relative contributions of each patch task and enable learning of all patch tasks with equal importance, without allowing easier patch tasks to dominate. Motivated by multi-task learning [24], [25], we propose a novel *Patch-Aware Priority Attention* (PAPA) method for multi-patch feature learning. It allows the MPMN to dynamically prioritize hard patch tasks during training, where task difficulty is positively proportional to the triplet error of the triplet loss. In our implementation, the PAPA automatically prioritizes more difficult patch tasks by adaptively adjusting the weights of loss functions from each patch task.

The main contributions of this work are as follows:

- We design a modality alignment loss function termed MPMA to simultaneously balance and reduce the modality discrepancies of both coarse-grained and fine-grained features.
- We propose a correlation distillation loss function called CPCD to transfer the semantic knowledges between different patches, boosting the cross-modality representations.
- We put forward a multi-task method named PAPA to dynamically adjust patch-level loss weights to continually prioritize difficult patch tasks.
- Without using modality-unshared networks, our one-stream VTReID framework named MPMN is practicable and efficient in real-world applications, achieving state-of-the-art performance on RegDB [26] and SYSU-MM01 [27] datasets.

The remainder of this paper is organized as follows: In Section II, some related works about person re-identification and cross-modality retrieval are discussed. In Section III, we introduce the proposed VTReID framework together with three novel components in detail. In Section IV, we compare the proposed methods with state-of-the-art methods and present extensive experiments for ablation study. In Section V, we conclude the main contents of this paper.

II. Related Works

A. Visible Thermal Person Re-Identification

Visible Person Re-Identification (VReID) addresses the problem of matching person images across disjoint visible cameras, which is widely used in video surveillance and public security. The VReID usually suffers from the large intra-class variations caused by different views, poses and occlusions. To mitigate those issues, prior VReID works have broadly adopted representation learning [1]–[8] or metric learning [9]–[12]. (1) The goal of representation learning is to adopt deep neural networks for learning robust and discriminative features. As global features learned from the full image intend to capture the coarse-grained clues of appearance, the global feature maps in [2], [3] are equally divided into multiple horizontal patches to exploit fine-grained local details. Based on PCB [3], some following works, i.e., MGN [28], PyramidNet [29] and HPM [2], extract both coarse-grained and fine-grained person representations by dividing convolutional feature maps horizontally into multi-grained patches. STA [7] fully exploits discriminative parts of one target person in both spatial and temporal dimensions for video-based person re-identification. AANet [30] leverages on a baseline model that uses body parts and integrates the key attribute information [31] in an unified learning framework. (2) Metric learning aims at making the positive pair have a relatively larger similarity than that of the negative pair. For example, the quadruplet loss [9] is designed to make the model output with larger inter-class variations and smaller intra-class variations compared to the triplet loss. Besides, the hard-aware point-to-set loss [12] adopts a soft hard-mining scheme to solve the limitation of traditional sampling methods. However, most of the prior works focus on the VReID task and visible cameras are unable to capture effective and reliable visual information in night time or dark environment. Therefore, VReID models may not perform well for the VTReID task, which limits the cross-modality generalization in practical surveillance applications.

B. Visible Thermal Person Re-Identification

Visible Thermal Person Re-Identification (VTReID) attempts to match visible and thermal images of the same person under cross-modality cameras. For the VTReID problem, in addition to views, poses and occlusions, the modality discrepancies between visible and thermal images need to be addressed. Zero-Padding [27] designs a one-stream network with deep zero-padding to automatically evolve modality-specific nodes in the network for cross-modality matching. TONE [14] utilizes a two-stream network to learn the multi-modality sharable feature representations for two heterogenous modalities and HCML [14] is introduced by jointly optimizing the modality-specific and modality-shared metrics. BDTR [13] uses a dual-path network for feature extraction and a bi-directional dual-constrained top-ranking loss for feature learning. To handle the lack of insufficient discriminative information, CM-GAN [15] designs a cutting-edge generative adversarial training based discriminator to learn discriminative feature representation from different modalities. HSME [18] learns a hypersphere manifold embedding and constrains the intra-modality and inter-modality variations on this hypersphere. AlignGAN [21] jointly performs pixel alignment and feature alignment to reduce the modality variations. However, most of the above methods mainly focus on relieving the modality discrepancies of global coarse-grained features but ignore the large modality variations of the local fine-grained features. Different from these methods, our proposed framework simultaneously reduces the modality discrepancies of coarse-grained and fine-grained person features, which boosts the consistency between the visible and thermal modalities.

C. Cross-Modality Retrieval

Cross-modality retrieval [32]–[34] refers to searching samples across different modality data, such as searching text in image dataset related to it semantically. Search between image and document [35], [36] is a representative cross-modality retrieval, which has attracted extensive attention in the past few years. The cross-modality methods include, but not limited to, traditional statistical correlation analysis [37], heterogenous face recognition [38], [39] and cross-modality data synthesis [40]. As for heterogenous face recognition, deep face features learned by CNNs have considered in [41] for NIR-VIS face recognition. Notably, our scenario is related to the above works, which however differs in person re-identification. Therefore such works cannot be directly applied. Specifically, the VTReID task faces large pose and viewpoint variations besides cross-modality variations compared with face recognition problem, which makes these methods unsuitable for VTReID. In addition, we jointly reduce coarse-grained and fine-grained modality discrepancies between visible and thermal person images, which contributes to the generalization capability of the VTReID models.

III. Proposed Method

A. Problem Definition

We define a mini-batch of training images from two different modalities by $\mathcal{X}^v = \{\boldsymbol{x}_i^v\}_{i=1}^{N/2}$ and $\mathcal{X}^t = \{\boldsymbol{x}_i^t\}_{i=1}^{N/2}$ with their corresponding identity labels $\mathcal{Y}^v = \{\boldsymbol{y}_i^v\}_{i=1}^{N/2}$ and $\mathcal{Y}^t = \{\boldsymbol{y}_i^t\}_{i=1}^{N/2}$. Accordingly, the total training images and identity labels of each mini-batch are denoted by $\mathcal{X} = \mathcal{X}^v \cup \mathcal{X}^t$ and $\mathcal{Y} = \mathcal{Y}^v \cup \mathcal{Y}^t$, respectively. Here, N is the number of training images. Our VTReID framework aims at learning a feature extractor to transform a visible image $\boldsymbol{x}_i^v$ to a visible feature $\boldsymbol{f}_i^v$ and a thermal image $\boldsymbol{x}_i^t$ to a thermal feature $\boldsymbol{f}_i^t$. The learning goal is that the two heterogeneous features $\boldsymbol{f}_i^v$ and $\boldsymbol{f}_i^v$ of the same person identity are close to each other.

B. Multi-Patch Matching Network

In this section, we introduce an effective cross-modality framework termed *Multi-Patch Matching Network* (MPMN) to learn coarse-grained and fine-grained visual semantics for the VTReID task.

Network Structure: As shown in Fig. 2, our MPMN framework consists of two main learnable modules. The first module is a stack of several convolutional blocks from ResNet-50 [42]. Given an input image $\boldsymbol{x}_i^r$, this module network outputs a convolutional feature map $\boldsymbol{F}_i^r \in \mathbb{R}^{C\times H\times W}$,

$$\boldsymbol{F}_i^r = \mathcal{F}\left(\boldsymbol{x}_i^r; \boldsymbol{\theta}_{\mathcal{F}}\right), \forall r \in \{v, t\}, \tag{1}$$

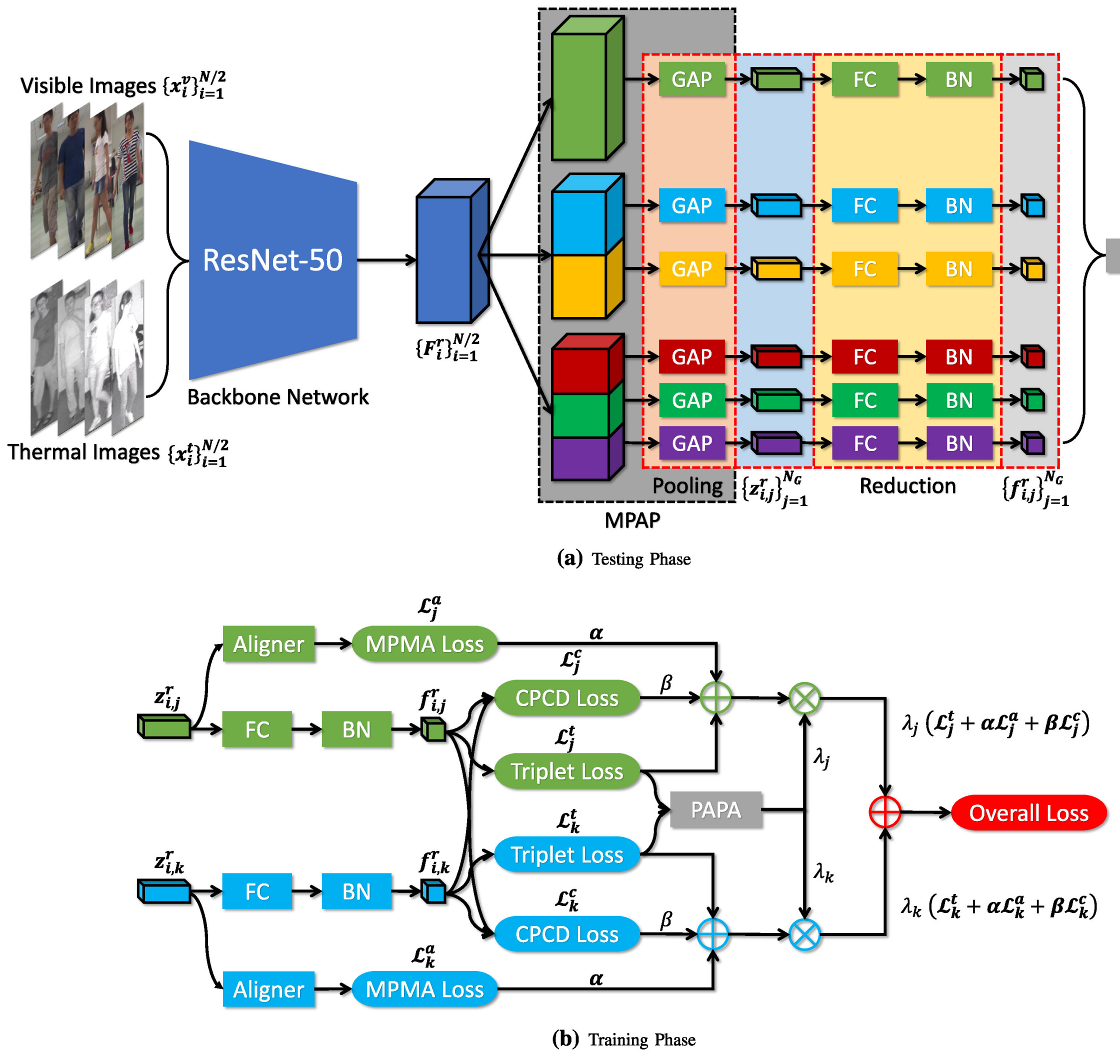
where r is the modality of the input image $\boldsymbol{x}_i^r$ and $\boldsymbol{\theta}_{\mathcal{F}}$ is the parameter of the first module $\mathcal{F}(\cdot;\boldsymbol{\theta}_{\mathcal{F}})$. C, H and W denote the channel, height and width dimension of the output feature maps, respectively. The second module termed *Multi-Patch Average Pooling* (MPAP) aims at learning multi-grained patch features without using body part annotations. Motivated by prior works [2], [3], [28], [29], we partition $\boldsymbol{F}_i^r$ into g horizontal stripe feature maps $\{\boldsymbol{F}_{i,j}^r\}_{j=1}^g$ to maintain the local fine-grained information. Then, all g stripe feature maps are individually averaged by a conventional *Global Average Pooling* (GAP) into g local feature vectors $\{\boldsymbol{z}_{i,j}^r\}_{j=1}^g$,

$$\boldsymbol{z}_{i,j}^r = \mathtt{GAP}\left(\boldsymbol{F}_{i,j}^r\right), \forall r \in \{v, t\}, \tag{2}$$

where $\mathtt{GAP}(\cdot)$ denotes the operator of the GAP layer. Analogously, we extract multi-patch features by varying g from 1 to G to excavate both coarse-grained and fine-grained visual semantics,

$$\begin{aligned}\{\boldsymbol{z}_{i,j}^r\}_{j=1}^{N_G} &= \mathtt{MPAP}\left(\boldsymbol{F}_i^r; \{1, 2, \ldots, G\}\right) \\ &= \left\{\mathtt{GAP}\left(\boldsymbol{F}_{i,j}^r\right)\right\}_{j=1}^{N_G}\end{aligned}, \forall r \in \{v, t\}, \tag{3}$$

where $\mathtt{MPAP}(\cdot)$ denotes the operator of the MPAP layer, $N_G = G(G+1)/2$ is the number of patch features and G is the number of patch groups. In order to reduce the feature dimension, all patch features $\boldsymbol{z}_{i,j}^r \in \mathbb{R}^C$ are individually transformed by a *Fully-Connected* (FC) layer and a *BatchNorm* (BN) layer [43], [44] into compressed features $\boldsymbol{f}_{i,j}^r \in \mathbb{R}^{D/N_G}$. Hence, the total dimension of all patch features is D, which is the same as the



(a) Testing Phase

(b) Training Phase

Fig. 2. The VTReID framework of the proposed MPMN. (a) The ResNet-50 backbone is split into six patches after the last residual block. During testing, all reduced features are concatenated together as the final representation of a person image. Note that the FC and BN layers for dimension reduction in each patch do not share weights with each other. (b) For simplicity, we use the j-th and k-th patches to introduce the training phase. The three loss functions, i.e., triplet loss, MPMA loss and CPCD loss, are used for training while PAPA dynamically provides weights of different patches on the basis of triplet errors.

single global feature ($G = 1$). In this way, we can strictly demonstrate that the performance improvement of the MPMN model is not brought by the additional computation costs and network parameters in the following experiments.

Loss Function: A triplet loss is adopted for a VTReID model since it has better generalization ability than classification loss especially when training data are not large enough. Given a mini-batch of person features $\{\boldsymbol{f}_{i,j}\}_{i=1}^{N} = \{\boldsymbol{f}_{i,j}^v\}_{i=1}^{N/2} \cup \{\boldsymbol{f}_{i,j}^t\}_{i=1}^{N/2}$, we sample a feature triplet $(\boldsymbol{f}_{a,j}, \boldsymbol{f}_{p,j}, \boldsymbol{f}_{n,j})$ where $\boldsymbol{f}_{a,j}$ and $\boldsymbol{f}_{p,j}$ are of the same identity whilst $\boldsymbol{f}_{a,j}$ and $\boldsymbol{f}_{n,j}$ are from different identities. We expect the cosine similarities between $\boldsymbol{f}_{a,j}$ and $\boldsymbol{f}_{p,j}$ are larger than the cosine similarities between $\boldsymbol{f}_{a,j}$ and $\boldsymbol{f}_{n,j}$, which leads to the formulation of triplet loss,

$$\min_{\boldsymbol{\theta}_{\mathcal{F}}} \mathcal{L}_j^t = \min_{\boldsymbol{\theta}_{\mathcal{F}}} \sum_{a,p,n}^{N} \big[\langle \boldsymbol{f}_{a,j}, \boldsymbol{f}_{n,j} \rangle - \langle \boldsymbol{f}_{a,j}, \\ \boldsymbol{f}_{p,j} \rangle + m_t\big]_+ , \tag{4}$$

where $[\cdot]_+ = \max(\cdot, 0)$ represents a hinge loss and the similarity margin m_t is set as $m_t = 0.2$. $< \boldsymbol{v}_1, \boldsymbol{v}_2 >$ denotes the cosine similarity between two vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. For learning multi-patch features, we formulate the multi-patch triplet loss as the following:

$$\min_{\boldsymbol{\theta}_{\mathcal{F}}} \mathcal{L}^t = \sum_{j=1}^{N_G} \lambda_j \min_{\boldsymbol{\theta}_{\mathcal{F}}} \mathcal{L}_j^t, \tag{5}$$

where the loss weight λ_j controls the importance of different patch feature learning. During the testing phase, all patch features are concatenated as the final person representation, therefore both the coarse-grained and fine-grained semantics are mined to perfect the comprehensiveness for features.

C. Multi-Patch Modality Alignment

A major challenge in VTReID is that the distribution of the features from two modalities might be very different, resulting in poor generalization and slow convergence. A naive application of the modality alignment for the original feature space may not suffice. Specifically, directly adding modality alignment constraints after output features might be detrimental, since it is hard to know which dimension contains maximum modality discrepancies. Thus, it would cause the selective alignment behavior, i.e., only focuses on certain dimensions or subspaces which are the easiest ones to reduce the current distribution differences. Moreover, most of previous VTReID works only consider the modality discrepancies of global features and leave the local modality gaps out of consideration, so the modality distributions of local features may not be well aligned, resulting in a inferior cross-modality performance.

In order to jointly solve these problems, we propose a novel loss function named *Multi-Patch Modality Alignment* (MPMA) loss to simultaneously balance and reduce modality discrepancies of multi-patch features. Specifically, a light-weight modality aligner trained by the MPMA loss is constructed to mine a hard feature subspace with large modality discrepancies and then align the modality distribution on this subspace. In this way, we pay more attention on reducing the modality discrepancies of the hard feature subspace than that of the easy subspace, which is beneficial to mitigate the imbalance of modality discrepancies.

Maximize Subspace Discrepancy: In order to make the modality distributions more distinguishable, we utilize a FC layer for the modality aligner $\mathcal{A}_j$ to model subspace projection for the j-th patch feature. Therefore, the subspace feature $\hat{\boldsymbol{z}}_{i,j}^{r} \in \mathbb{R}^{P}$ is formulated by

$$\hat{\boldsymbol{z}}_{i,j}^{r} = \mathcal{A}_j\left(\boldsymbol{z}_{i,j}^{r}; \boldsymbol{\theta}_{\mathcal{A}_j}\right), \forall r \in \{v, t\}, \tag{6}$$

where $\boldsymbol{\theta}_{\mathcal{A}_j}$ denotes the projection parameter of the modality aligner $\mathcal{A}_j$ and the subspace dimension P is set as $P = C/4$ as default. In order to explore the hard feature subspace, one simple but effective way is to learn the optimal modality aligner $\mathcal{A}_j$ to maximize the subspace modality discrepancies by the following objective,

$$\max_{\boldsymbol{\theta}_{\mathcal{A}}} \mathcal{L}^{a} = \sum_{j=1}^{N_G} \lambda_j \max_{\boldsymbol{\theta}_{\mathcal{A}_j}} \mathcal{L}_j^{a}, \tag{7}$$

where

$$\mathcal{L}_j^{a} = \frac{1}{N_a \times P} \sum_{\boldsymbol{y}_k^{v} = \boldsymbol{y}_l^{t}} \left\| \hat{\boldsymbol{z}}_{k,j}^{v} - \hat{\boldsymbol{z}}_{l,j}^{t} \right\|_2^2 . \tag{8}$$

N_a is the number of positive feature pairs from different modalities, while $\boldsymbol{\theta}_{\mathcal{A}} = \{\boldsymbol{\theta}_{\mathcal{A}_j}\}_{j=1}^{N_G}$ is the parameter of all modality aligners. Note that the gradients for $\mathcal{L}^{a}$ in Eq. 7 are only backpropagated to $\boldsymbol{\theta}_{\mathcal{A}}$ and the backbone network $\boldsymbol{\theta}_{\mathcal{F}}$ is fixed. The subspace feature distribution of different modalities are made more distinguishable by maximizing Eq. 7, so any shift in the original modality distributions can be easily detected.

Minimize Subspace Discrepancy: After obtaining the optimal modality aligner, we need to reduce the modality discrepancies of the hard feature subspace by minimizing the following objective,

$$\min_{\boldsymbol{\theta}_{\mathcal{F}}} \mathcal{L}^{a} = \sum_{j=1}^{N_G} \lambda_j \min_{\boldsymbol{\theta}_{\mathcal{F}}} \mathcal{L}_j^{a}, \tag{9}$$

where the gradients for $\mathcal{L}^{a}$ are only backpropagated to the backbone network $\boldsymbol{\theta}_{\mathcal{F}}$ but do not update the modality aligner $\boldsymbol{\theta}_{\mathcal{A}}$. As a result, we minimize $\mathcal{L}^{a}$ with respect to $\boldsymbol{\theta}_{\mathcal{F}}$ to make the multi-grained features more modality-invariant.

Adversarial Subspace Learning: Finally, we jointly optimize Eq. 7 and Eq. 9 to learn modality-robust features. Since the optimization goals of these two objective functions are opposite, the process runs as a minimax game leading to an adversarial learning problem [23] as

$$\min_{\boldsymbol{\theta}_{\mathcal{F}}} \max_{\boldsymbol{\theta}_{\mathcal{A}}} \mathcal{L}^{a} = \sum_{j=1}^{N_G} \lambda_j \min_{\boldsymbol{\theta}_{\mathcal{F}}} \max_{\boldsymbol{\theta}_{\mathcal{A}_j}} \mathcal{L}_j^{a}. \tag{10}$$

Remark: The proposed MPMA is similar to adversarial learning of CM-GAN [15], ACME [16] and TIMAM [45], but differs in the following details. (1) The first distinction is related to the motivation of loss functions. The MPMA learns hard subspace based on Euclidean distance as modality adversary, while the three works confuse the binary modality classifier based on logistic regression. Compared with the MPMA, the three works only ensure that the features from different modalities cannot be linearly separated by the binary modality classifier, but the modality distribution may not be thoroughly aligned. (2) The second distinction is linked to the granularity of modality alignment. For instance, the three works only impose modality alignment on global patch features, therefore there is certainly no guarantee that local patch features are robust to modality discrepancies, resulting in poor cross-modality generalization performance. (3) The last distinction is associated with the supervision of modality discrepancies. The MPMA focuses on the modality discrepancies of intra-class features, while the three works consider the modality discrepancies of both intra-class and inter-class features. However, there exists the contradiction between reducing inter-class modality discrepancies and enlarging inter-class identity differences, the three works may not learn discriminative and robust person features.

D. Cross-Patch Correlation Distillation

Generally, coarse-grained features are robust but less discriminative, while fine-grained features are discriminative but less robust. Thus, these two features are potentially complementary to each other, if cross-patch correlations are effectively exploited. An straightforward way is to directly train each patch feature with an independent triplet loss and then concatenate all patch features for testing. However, the triplet loss is unable to regulate the correlations between different patches, which results in less discriminative and robust representations.

Inspired by knowledge distillation [46], we put forward a new loss function termed *Cross-Patch Correlation Distillation* (CPCD) loss to transfer the semantic knowledges of one patch to another patch by directly enforcing the cross-patch similarity correlations in training. Since the patch correlation is valuable information that defines a rich similarity structure over training data, our goal is to excavate the informative patch correlation among N_G patch features in a metric learning framework. We compose two types of correlations according to the relative similarity divergence of the two feature pairs from different patches.

Positive Cross-Patch Correlation: Suppose that we have two different patch pairs $(\boldsymbol{f}_{a,k}, \boldsymbol{f}_{b,k})$ and $(\boldsymbol{f}_{a,j}, \boldsymbol{f}_{b,j})$ where $a \neq b$ and $k \neq j$, the similarity divergence should be small because they both encode the same image pair $(\boldsymbol{x}_a, \boldsymbol{x}_b)$. Therefore, the correlation between the two patch pairs are positive. Mathematically, we expect that the similarity divergences of different patches from the same image pairs should be smaller than a margin $m_p > 0$, i.e.,

$$\min_{\boldsymbol{\theta}_\mathcal{F}} \mathcal{L}_j^p = \min_{\boldsymbol{\theta}_\mathcal{F}} \sum_{k\neq j}^{N_G} \sum_{a\neq b}^{N} \left[\left| \langle \boldsymbol{f}_{a,k}, \boldsymbol{f}_{b,k} \rangle - \langle \boldsymbol{f}_{a,j}, \boldsymbol{f}_{b,j} \rangle \right| - m_p \right]_+ , \tag{11}$$

where $|\cdot|$ denotes computing absolute values.

Negative Cross-Patch Correlation: Suppose that we have two different patch pairs $(\boldsymbol{f}_{a,k}, \boldsymbol{f}_{p,k})$ and $(\boldsymbol{f}_{a,j}, \boldsymbol{f}_{n,j})$ where $\boldsymbol{y}_a = \boldsymbol{y}_p$, $\boldsymbol{y}_a \neq \boldsymbol{y}_n$ and $k \neq j$, the similarity of the former pair $(\boldsymbol{f}_{a,k}, \boldsymbol{f}_{p,k})$ should be larger than the latter pair $(\boldsymbol{f}_{a,l}, \boldsymbol{f}_{n,l})$. Therefore, the correlation between the two patch pairs are negative. In order to explore the negative correlation, we enforce that the similarity divergence between the two patch pairs is larger than a margin $m_n > 0$, i.e.,

$$\min_{\boldsymbol{\theta}_\mathcal{F}} \mathcal{L}_j^n = \min_{\boldsymbol{\theta}_\mathcal{F}} \sum_{k\neq j}^{N_G} \sum_{a,p,n}^{N} \left[\langle \boldsymbol{f}_{a,k}, \boldsymbol{f}_{n,k} \rangle - \langle \boldsymbol{f}_{a,j}, \boldsymbol{f}_{p,j} \rangle + m_n \right]_+ . \tag{12}$$

Transfer Overall Correlation: For modeling complete correlations, we jointly transfer positive and negative cross-patch correlations by combining Eq. 11 and Eq. 12, leading to the proposed CPCD loss function as following,

$$\min_{\boldsymbol{\theta}_\mathcal{F}} \mathcal{L}^c = \sum_{j=1}^{N_G} \lambda_j \min_{\boldsymbol{\theta}_\mathcal{F}} \mathcal{L}_j^c = \sum_{j=1}^{N_G} \lambda_j \min_{\boldsymbol{\theta}_\mathcal{F}} \left(\mathcal{L}_j^p + \mathcal{L}_j^n \right), \tag{13}$$

where we set $m_p = 0.1$ and $m_n = 0.1$ as default.

Remark: The proposed CPCD method differs with knowledge distillation [46] in aspects of both motivation and methodology. (1) For the motivation aspect, Hinton *et al.* [46] transfer knowledge from a pre-trained large teacher network to a smaller student network for network compression. The proposed CPCD aims at transferring cross-patch correlations from one patch feature to another patch feature for boosting multi-patch representational ability. (2) For the methodology aspect, Hinton *et al.* [46] use the class probabilities produced by the teacher model as "soft" targets for training the student model and do not update the teacher network during the student network training. Differently, the CPCD collaboratively learns multi-patch features from positive and negative cross-patch correlations based on the prior similarity divergence.

E. Patch-Aware Priority Attention

Finally, we joint the three loss functions, i.e., triplet loss, MPMA loss and CPCD loss, to train our MPMN end-to-end as following,

$$\begin{aligned} \min_{\boldsymbol{\theta}_\mathcal{F}} \max_{\boldsymbol{\theta}_\mathcal{A}} \mathcal{L} &= \min_{\boldsymbol{\theta}_\mathcal{F}} \max_{\boldsymbol{\theta}_\mathcal{A}} \left(\mathcal{L}^t + \alpha \mathcal{L}^a + \beta \mathcal{L}^c \right) \\ &= \sum_{j=1}^{N_G} \lambda_j \min_{\boldsymbol{\theta}_\mathcal{F}} \max_{\boldsymbol{\theta}_\mathcal{A}} \left(\mathcal{L}_j^t + \alpha \mathcal{L}_j^a + \beta \mathcal{L}_j^c \right), \end{aligned} \tag{14}$$

where we set $\alpha = 0.1$ and $\beta = 0.1$ in our experiments. In order to learn multi-patch features, a straightforward approach is to treat each patch task equally by enforcing $\lambda_k = \lambda_j$ $(k \neq j)$. However, the equal loss weights completely ignore the imbalance of each patch learning. Specifically, Eq. 14 should dynamically weight the relative contributions of each patch task to enable learning of all tasks with equal importance, without allowing easier tasks to dominate. Since manual tuning of weights is tedious, it is preferable to automatically learn weights via an attention mechanism.

Aiming to achieve the patch-aware weighting scheme, we propose a simple yet effective attention method, named *Patch-Aware Priority Attention* (PAPA), to prioritize the difficult patch tasks by adaptively assigning more weights to the harder patch tasks than the easies ones. For each patch task, the loss weight λ_j should be a meaningful metric for VTReID, so we use the triplet error $\mathcal{E}_j = N_e/N_t$ of the triplet loss $\mathcal{L}_j^t$ to represent the task difficulty. N_t is the number of all input triplets, while N_e is the number of the triplets which violate the similarity margin constraint. Accordingly, we formulate the loss weight by normalizing the triplet errors,

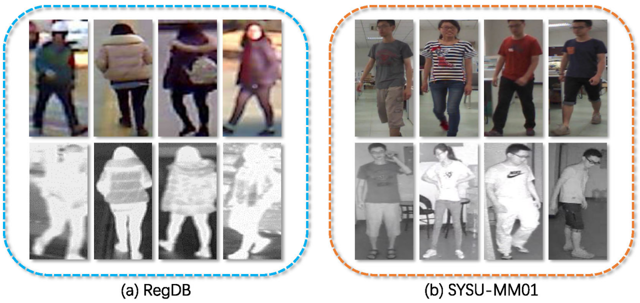
$$\lambda_j = N_G \times \frac{e^{\sigma(n)\mathcal{E}_j}}{\sum_{k=1}^{N_G} e^{\sigma(n)\mathcal{E}_k}}, \text{ where } \sigma(n) = \gamma \frac{n-1}{N_n - 1}. \tag{15}$$

N_n is the total training epochs, $n \in [1, N_n]$ is an epoch index and $\gamma > 0$ is a positive hyper-parameter. At the beginning of training, $\sigma(n)$ is small enough and all patch tasks are weighted equally. At the end of training, the large $\sigma(n)$ results in a more imbalanced weight distribution between different patch tasks and the hard tasks are assigned larger weights than the easy ones. The softmax operator, which is multiplied by N_G, ensures that $\sum_{j=1}^{N_G} \lambda_j = N_G$.

Remark: On the balancing of multi-task learning, there is extensive experimental analysis in [24], [25], with both papers arguing that different amounts of weighting tend to work best for different tasks. One example of weighting tasks appropriately is with the use of weight uncertainty [25], which modifies the loss functions in multi-task learning using task uncertainty. Another method is GradNorm [24], which manipulates gradient norms over time to control the training dynamics. Different from the previous works, the proposed PAPA method encourages prioritization of hard tasks directly using priority metrics such as triplet error. Compared with GradNorm which requires access to the network's internal gradients, our PAPA only requires the numerical triplet errors and therefore its implementation is far simpler.

Algorithm 1: Deep Multi-Patch Feature Learning

```
Input: Visible images $\mathcal{X}^v = \{\boldsymbol{x}_i^v\}_{i=1}^{N/2}$ and their labels
  $\mathcal{Y}^v = \{\boldsymbol{y}_i^v\}_{i=1}^{N/2}$; thermal images $\mathcal{X}^t = \{\boldsymbol{x}_i^t\}_{i=1}^{N/2}$ and their
  labels $\mathcal{Y}^t = \{\boldsymbol{y}_i^t\}_{i=1}^{N/2}$; hyperparameters $\alpha$, $\beta$ and $\gamma$;
  similarity margins $m_t$, $m_p$ and $m_n$; learning rate $r$.
Output: Network parameters $\boldsymbol{\theta}_{\mathcal{F}}$ and $\boldsymbol{\theta}_{\mathcal{A}}$
while not convergence do
    update parameters $\boldsymbol{\theta}_{\mathcal{F}}$ by descending their
    stochastic gradients:
      $\boldsymbol{\theta}_{\mathcal{F}} \leftarrow \boldsymbol{\theta}_{\mathcal{F}} - r \cdot \nabla_{\boldsymbol{\theta}_{\mathcal{F}}}(\mathcal{L}^t + \alpha\mathcal{L}^a + \beta\mathcal{L}^c)$
    update parameters $\boldsymbol{\theta}_{\mathcal{A}}$ by ascending their
    stochastic gradients throught GRL:
      $\boldsymbol{\theta}_{\mathcal{A}} \leftarrow \boldsymbol{\theta}_{\mathcal{A}} + r \cdot \nabla_{\boldsymbol{\theta}_{\mathcal{A}}}(\mathcal{L}^t + \alpha\mathcal{L}^a + \beta\mathcal{L}^c)$
end while
```



(a) RegDB

(b) SYSU-MM01

Fig. 3. Visualization of visible and thermal person images. The top row is visible images captured by visible cameras, while the bottom row is thermal images captured by thermal cameras. The left and right four columns show the samples from RegDB [26] and SYSU-MM01 [27] datasets, respectively. The two images of each column belong to the same identity.

F. Training Optimization

The process of learning the optimal feature representations in Eq. 14 is conducted by jointly optimizing minimum and maximum objectives. Since the optimization goals of these two objective functions are opposite, the process runs as a minimax game of the two concurrent sub-processes,

$$\hat{\boldsymbol{\theta}}_{\mathcal{F}} = \underset{\boldsymbol{\theta}_{\mathcal{F}}}{\operatorname{argmin}} \left(\mathcal{L}^t + \alpha\mathcal{L}^a + \beta\mathcal{L}^c\right),$$
$$\hat{\boldsymbol{\theta}}_{\mathcal{A}} = \underset{\boldsymbol{\theta}_{\mathcal{A}}}{\operatorname{argmax}} \left(\mathcal{L}^t + \alpha\mathcal{L}^a + \beta\mathcal{L}^c\right), \tag{16}$$

where $\hat{\boldsymbol{\theta}}_{\mathcal{F}}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$ denote the optimal solutions of $\boldsymbol{\theta}_{\mathcal{F}}$ and $\boldsymbol{\theta}_{\mathcal{A}}$, respectively. This minimax game can be implemented using a stochastic gradient descent optimization algorithm. As proposed in [47], minimax optimization can be performed efficiently by incorporating *Gradient Reversal Layer* (GRL), which is transparent when forward-propagating, but which multiples its values by -1 when back-propagating. If GRL is added before the modality aligner, the minimax optimization can be performed simultaneously, as shown in Algorithm 1.

IV. Experiments

A. Datasets

To evaluate the proposed method, we adopt two publicly available datasets, i.e., RegDB [26] and SYSU-MM01 [27], for evaluation. Example images from the two datasets are illustrated in Fig. 3.

RegDB [26]: It is collected from one visible or one thermal camera. It contains totally 412 persons and each person has 10 visible images and 10 thermal images captured by different cameras. We follow the evaluation protocol [13] to randomly split the dataset into two halves, which are used for training and testing respectively. The procedure is repeated for 10 times, then we compute the average results to achieve statistically stable results. Besides, we use the two query settings including `visible to thermal` and `thermal to visible`. Here, `visible to thermal` means that visible images are used for query set and thermal images are used for gallery set, and so on.

SYSU-MM01 [27]: It is a large-scale dataset collected by four visible cameras and two thermal cameras. It contains in total 491 persons, and each person is captured by at least two different cameras. Our experiments follow the evaluation protocol [13] with two testing modes, i.e., `all-search` and `indoor-search` mode. For `all-search` mode, all images are used. For `indoor-search` mode, only indoor images captured by camera 1, 2, 3 and 6 are used. For both modes, the `single-shot` and `multi-shot` settings are adopted, where 1 or 10 images of a person are randomly selected to form the gallery set. The procedure is repeated for 10 times, then we compute the average results to achieve statistically stable results. Both modes use thermal images as query set and visible images as gallery set.

B. Implementation Details

Network Architecture: We implement the proposed VTReID framework based on `Pytorch` [48]. We take the ResNet-50 [42] initialized with the parameters pretrained on ImageNet [49] as the backbone network. Following the work [2], the last FC layer and GAP layer are removed and the stride of the last residual block *Conv4_1* is set from 2 to 1 for increasing the feature map size.

Data Processing: In order to obtain enough context information from person images and a proper size of feature maps, we first resize training images to 384 × 128. Then we randomly crop each training image with scale in the interval $[0.64, 1.0]$ and aspect ratio $[2, 3]$. Third, we resize these cropped images back to 384 × 128. Following the work [50], the training images are augmented with random erasing [51] and random horizontal flipping. Before sent to the network, each image is subtracted from the mean values and divided by the standard deviations according to normalization procedure when using the pretrained model on ImageNet [49]. Since visible and thermal images have different color channels, we transform both of them into gray images with a single channel. In view of the three-dimensional input channel of pre-trained models, we duplicate all gray images with three times to make the images have the same shape with the model input.

Training Configurations: Since triplet loss is used to learn person features, we need to adopt an appropriate triplet sampling strategy. To simplify this procedure, the $\mathcal{PK}$ sampling method [10] is applied to generate triplets, which randomly samples $\mathcal{P}$ identities and then randomly selects $\mathcal{K}/2$ visible images and $\mathcal{K}/2$ thermal images for each identity to form a mini-batch with the size $\mathcal{P} \times \mathcal{K}$. In a mini-batch, we use all possible $\mathcal{PK}(\mathcal{PK} - \mathcal{K})(\mathcal{K} - 1)$ combinations of triplets for triplet loss. For all datasets, $\mathcal{P}$ and $\mathcal{K}$ are set to 20 and 8, respectively. We use the SGD algorithm to minimize the overall loss function, where the initial learning rate, weight decay and momentum are set to 0.01, 2×10^{-4} and 0.9, respectively. The learning rate is decreased by a factor of 5 after every 40 epochs and all models are trained for 160 epochs. All experiments run on a sever with 2 Intel(R) Xeon(R) E5-2620 v4@2.10 GHz CPUs, 4 GeForce GTX 1080 Ti GPUs and 128G RAM.

Testing Configurations: We use standard *Cumulative Matching Characteristics* (CMC) curve and *Mean Average Precision* (mAP) to evaluate our algorithm. Specifically, we adopt the CMC at Rank1, Rank10 and Rank20. For the traditional VReID task, there are `single-qurey` and `multi-query` modes in evaluation. It is worth noting that all our results are obtained in a `single-query` mode, which is difficult than the `multi-query` mode in image retrieval.

C. Comparison with the State-of-the-Art Methods

Baseline Model: For baseline comparison, we survey VTReID works published at top conferences of past years. We use `V5` in Table III as our baseline model. Specifically, our baseline model outputs only global features ($G = 1$ in Fig. 2) and is trained only by the triplet loss. For other baselines [13]–[16], [18], [19], [21], [27], most of them build two separate stream networks with unshared parameters to capture modality-specific feature patterns, but such two-stream structure would largely increase inference budget and network parameters. Unlike the two-stream structure, our baseline shares a one-stream network for heterogeneous modalities to reduce computational costs, so it has widespread practical significance and application prospect in real-world cross-modality circumstances. As shown in Fig. 4, most of previous works are expanded on poor baseline models. On RegDB dataset, only two baseline models in 8 baseline models surpass **35%** Rank1 accuracy and **35%** mAP score. The Rank1 accuracies and mAP scores of the some baseline models even lower than **30%** and **25%**, respectively. On SYSU-MM01 dataset, all baseline models except our baseline do not surpass **30%** Rank1 accuracy or **40%** mAP score. In addition, we also find that some works are unfairly compared with other state-of-the-art methods. Specifically, the performance improvements are mainly from poor baseline models rather than methods themselves. Some training tricks are understated in the paper so that readers ignore them. It may make the effectiveness of the proposed method exaggerated. According to the above analysis, we think that a simple and strong baseline model is very important to promote the development of VTReID research, and we need to take into account the baseline models when commenting academic papers.

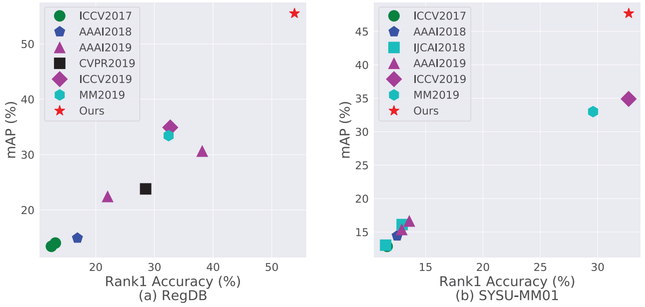


Fig. 4. The cross-modality retrieval performance of different baseline models on RegDB and SYSU-MM01 datasets. We compare our strong baseline model with other baseline published at top conferences of past years. Specifically, we use `V5` in Table III as our baseline model. Our baseline model outputs only global features ($G = 1$ in Fig. 2) and is trained only by the triplet loss. For other baselines [13]–[16], [18], [19], [21], [27], most of them build two separate stream networks with unshared parameters to capture modality-specific feature patterns. Note that we adopt the `visible to thermal` setting for RegDB dataset while the `all-search & single-shot` setting is used for SYSU-MM01 dataset.

Proposed Method: We compare the proposed method with the state-of-the-art VTReID methods on RegDB and SYSU-MM01 datasets. As shown in Table I and Table II, the proposed method outperforms the existing state-of-the-art methods by a large margin on RegDB and SYSU-MM01 datasets. Specifically, our VTReID framework significantly improves **29.31%** and **26.09%** in terms of `visible to thermal` and `thermal to visible` mAP scores with comparison to the second best method AlignGAN [21]. Compared with the other methods on SYSU-MM01 dataset under the `all-search` mode, our VTReID framework also achieves significant gains of at least **21.71%** and **18.00%** in terms of `single-shot` and `multi-shot` mAP accuracies. In view of network parameters, the proposed VTReID framework shares the backbone network between visible and thermal modalities, while some other works like TONE+HCML [14] learn modality-specific features with different parameters. Moreover, the proposed method does not introduce additional parameters to the baseline model during testing, which shows that our method has a widespread practical significance and application prospect in real-world circumstances.

D. Ablation Study for Baseline Methods

In order to explore the strong baseline models, we perform extensive ablation studies on RegDB and SYSU-MM01 datasets to validate the contributions of two components including image processing and layer setting. Note that we adopt the `visible to thermal` setting for RegDB dataset while the `all-search & single-shot` setting is used for SYSU-MM01 dataset.

Image Processing: As shown in Fig. 3, visible images usually have three color channels, while thermal images have only a single color channel. Directly implementing visible and thermal

TABLE I
ACCURACIES (%) ON REGDB CDATASET. "SHARE" WITH A BLUE MARKER $\surd$ REPRESENTS THE WEIGHTS OF MODALITY-SPECIFIC BACKBONE NETWORKS ARE SHARED; OTHERWISE THEY ARE NOT SHARED FOR DIFFERENT MODALITIES. "MP" DENOTES MULTI-PATCH FEATURES

Method	Publish	Share	Visible to Thermal				Thermal to Visible			
			Rank1	Rank10	Rank20	mAP	Rank1	Rank10	Rank20	mAP
Zero-Padding [27]	ICCV2017	✔	17.75	56.42	67.52	31.83	16.63	34.68	44.25	17.82
TONE + HCML [14]	AAAI2018	✘	24.44	47.53	56.78	20.80	21.70	45.02	55.58	22.24
BDTR [13]	IJCAI2018	✘	33.47	58.42	67.52	31.83	32.72	57.96	68.86	31.10
HSME [18]	AAAI2019	✘	41.34	65.21	75.13	38.82	40.67	65.35	75.27	37.50
D-HSME [18]	AAAI2019	✘	50.85	73.36	81.66	47.00	50.15	72.40	81.07	46.16
D^{2}RL [16]	CVPR2019	✔	43.40	66.10	76.30	44.10	-	-	-	-
AlignGAN [21]	ICCV2019	✔	57.90	-	-	53.60	56.30	-	-	53.40
MAC [19]	MM2019	✘	36.43	62.36	71.63	37.03	36.20	61.68	70.99	36.63
IPVT [17]	Access2019	✔	58.76	85.75	90.27	47.85	-	-	-	-
DGD + MSR [20]	TIP2019	✘	48.43	70.32	79.95	48.67	-	-	-	-
Baseline	-	✔	53.88	73.69	82.18	55.50	54.14	78.11	85.92	50.83
Baseline + MP	-	✔	75.44	89.27	94.22	73.09	74.12	90.07	94.98	71.54
MPMN	-	✔	**86.56**	**96.68**	**98.28**	**82.91**	**84.62**	**95.51**	**97.33**	**79.49**

TABLE II
ACCURACIES (%) ON SYSU-MM01 DATASET. "SEARCH" REPRESENTS THE `All-Search` OR `Indoor-Search` MODE. "MP" DENOTES MULTI-PATCH FEATURES

Method	Publish	Search	Share	Single-Shot				Multi-Shot			
				Rank1	Rank10	Rank20	mAP	Rank1	Rank10	Rank20	mAP
Zero-Padding [27]	ICCV2017	all	✔	14.80	54.12	71.33	15.95	19.13	61.40	78.41	10.89
TONE + HCML [14]	AAAI2018	all	✘	14.32	53.16	69.17	16.16	-	-	-	-
BDTR [13]	IJCAI2018	all	✘	17.01	55.43	71.96	19.66	-	-	-	-
CM-GAN [15]	IJCAI2018	all	✔	26.97	67.51	80.56	27.80	31.49	72.74	85.01	22.27
HSME [18]	AAAI2019	all	✘	18.03	58.31	74.43	19.98	-	-	-	-
D-HSME [18]	AAAI2019	all	✘	20.68	62.74	77.95	23.12	-	-	-	-
D^{2}RL [16]	CVPR2019	all	✔	28.90	70.60	82.40	29.20	-	-	-	-
AlignGAN [21]	ICCV2019	all	✔	42.40	85.00	93.70	40.70	51.50	89.40	95.70	33.90
MAC [19]	MM2019	all	✘	33.26	79.04	90.09	36.22	-	-	-	-
IPVT [17]	Access2019	all	✔	23.18	51.21	61.73	22.49	-	-	-	-
DGD+MSR [20]	TIP2019	all	✘	37.35	83.40	93.34	38.11	43.86	86.94	95.68	30.48
Baseline	-	all	✔	32.71	79.32	90.00	47.69	42.91	78.64	87.65	34.51
Baseline + MP	-	all	✔	43.01	86.32	93.38	57.11	53.23	84.79	91.02	45.17
MPMN	-	all	✔	**48.98**	**90.33**	**97.13**	**62.41**	**60.88**	**88.70**	**94.06**	**51.90**
Zero-Padding [27]	ICCV2017	indoor	✔	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.64
CM-GAN [15]	IJCAI2018	indoor	✔	31.63	77.23	89.18	42.19	37.00	80.94	92.11	32.76
AlignGAN [21]	ICCV2019	indoor	✔	45.90	87.60	94.40	54.30	57.10	92.70	97.40	45.30
MAC [19]	MM2019	indoor	✘	33.37	82.49	93.69	44.95	-	-	-	-
DGD + MSR [20]	TIP2019	indoor	✘	39.64	89.29	97.66	50.88	46.56	93.57	98.80	40.08
Baseline	-	indoor	✔	42.75	88.47	95.84	57.98	53.68	83.25	90.85	45.93
Baseline + MP	-	indoor	✔	54.44	92.75	96.14	69.83	66.62	89.04	94.83	58.50
MPMN	-	indoor	✔	**64.89**	**96.85**	**99.22**	**76.47**	**74.42**	**92.93**	**96.41**	**66.98**

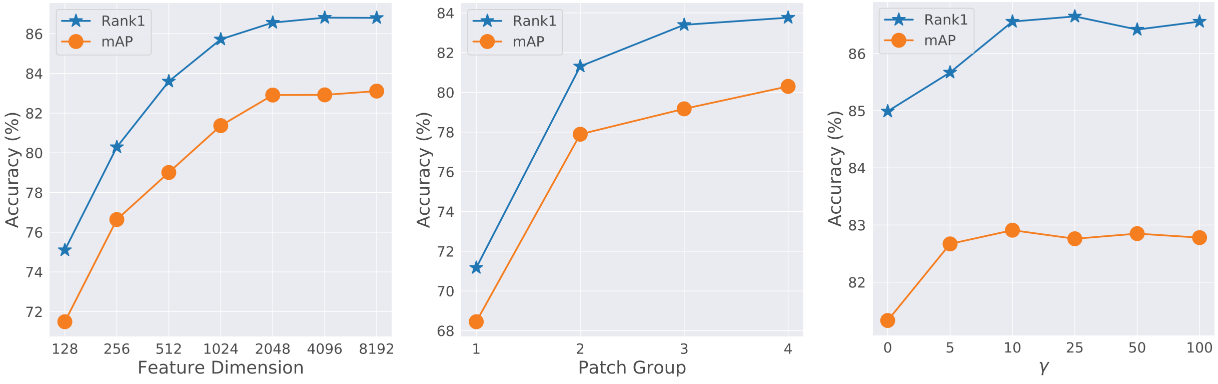
images becomes inevitably incompatible for modality-shared models because of different color channel dimensions. Thus, we introduce two image processing methods, i.e., Gray and RGB, in order to align color channels. Specifically, Gray denotes that visible images are transformed into gray images to make them have the same channel dimension as thermal images, while RGB denotes that we duplicate single-channel thermal images with three times to make them have the same shape with visible images. From `V1` to `V2` in Table III, we evaluate the impact of different image processing methods. Despite both of their results are probably close on RegDB dataset, Gray outperforms RGB with a significant margin on SYSU-MM01 dataset. According the above analysis, we recommend to use Gray in the following experiments.

Layer Settings: From `V3` to `V5` in Table III, we evaluate the impact of different layer settings. Compared with V1, changing the stride of the last downsampling layer *Conv4_1* from 2 to 1 (i.e., S1) can achieve higher retrieval accuracies, because enlarging the size of output features is beneficial to capture more detailed visual cues and then obtain informative person representations. Besides, we observe that using a BN layer input with extracted features can bring a significant improvement on VTReID performances, without increasing computational costs and memory consumption. That is to say, standardizing output features into a common Gaussian feature space is able to mitigate the modality misalignment problem to some extent. It is worth noting that the joint of S1 and BN can further enhance the generalization ability of the baseline models with the comparison to S1 or BN. Since S1 and BN provide the mutual promotion between discriminative and robust feature learning, we recommend to adopt both S1 and BN in the following experiments.

TABLE III

ABLATION STUDIES AND RETRIEVAL ACCURACIES (%) OF BASELINE MODELS ON REGDB AND SYSU-MM01 DATASETS. WE ANALYZE THE DIFFERENT COMPONENT COMBINATIONS GENERATED FROM TWO ASPECTS: IMAGE PROCESSING AND LAYER SETTINGS. IN THE FIRST COLUMN, THE FIVE VERSIONS OF BASELINE MODELS ARE MARKED FROM V1 TO V5. IN THE SECOND COLUMN, WE EXPLORE DIFFERENT IMAGE PROCESSING METHODS, I.E., GRAY AND RGB. IN THE THIRD COLUMN, WE STUDY THE COMBINATIONS OF THE TWO LAYER SETTINGS, I.E., BN AND S1. TO BE SPECIFIC, BN REPRESENTS THE BATCHNORM LAYER AFTER THE FEATURE EXTRACTOR, WHILE S1 MEANS THAT THE STRIDE OF THE LAST RESIDUAL BLOCK *CONV4_1* IS SET FROM 2 TO 1. SINCE V5 ACHIEVES RELATIVELY SUPERIOR PERFORMANCE ON REGDB DATASET AND THE BEST RESULTS ON SYSU-MM01 DATASET, WE ADOPT V5 AS THE FINAL BASELINE MODEL IN THE FOLLOWING EXPERIMENTS

Ver.	Gray	Layer Settings		RegDB				SYSU-MM01			
		BN	S1	Rank1	Rank10	Rank20	mAP	Rank1	Rank10	Rank20	mAP
V1	✘	✘	✘	51.25	70.89	80.43	51.90	29.98	74.04	87.44	44.57
V2	✔	✘	✘	51.70	71.11	80.53	52.09	31.27	76.63	89.21	46.80
V3	✔	✔	✘	**53.92**	**74.29**	**83.11**	54.82	31.34	76.90	89.59	47.00
V4	✔	✘	✔	52.17	72.43	82.07	52.33	32.05	78.87	89.93	47.65
V5	✔	✔	✔	53.88	73.69	82.18	**55.50**	**32.71**	**79.32**	**90.00**	**47.69**



(a) Feature Dimension **(b)** Patch Group **(c)** PAPA

Fig. 5. Accuracies (%) on RegDB dataset under the `Visibe to Thermal` setting. (a) Analyzing different dimensions of fused person features. (b) Analyzing different patch groups of fused person features. (c) The effect of hyperparameter γ of PAPA.

E. Ablation Study for Proposed Methods

We next analyze the contributions of proposed methods on RegDB and SYSU-MM01 datasets. Note that we adopt the `visible to thermal` setting for RegDB dataset while the `all-search & single-shot` setting is used for SYSU-MM01 dataset.

Analysis of Feature Dimension: To explore the sensitivity of the feature dimension D, we illustrate retrieval performances of different dimensions by varying D from 128 to 8192. For fair experiments, we set the number of patch groups as $G = 3$ in this experiment. As shown in Fig. 5(a), a larger feature dimension always benefits VTReID accuracies. The rank1 and mAP scores of RegDB dataset increase with feature dimensions until reaching a stable performance. Moreover, increasing the feature dimension ($D > 2048$) makes an unimpressive contribution to the accuracy improvement compared with $D = 2048$. Therefore, we recommend $D = 2048$ as it strikes a satisfactory balance between the computational efficiency and retrieval performance.

Analysis of Patch Group: In this part, we study the effectiveness of the number of patch groups $G \in \{1, 2, 3, 4\}$ on RegDB dataset. Here, $G = 1$ means that only global coarse-grained features are used for VTReID models, while $G > 1$ means that fused person representations contain both global coarse-grained and local fine-grained features. For experimental fairness, we set the feature dimension as $D = 2048$ for all multi-patch features. In this way, we can strictly demonstrate that the performance improvement of the MPMN model is not brought by the additional computation costs and network parameters in the following experiments. Besides, when $G = 1$, the proposed methods CPCD and PAPA are not applicable to the single global patch feature. Thus, CPCD and PAPA are not used for all multi-patch features in this experiment. As illustrated in Fig. 5(b), we observe that extracting more patch features always benefits cross-modality retrieval performances. This demonstrates that aggregating both coarse-grained and fine-grained features can enhance the cross-modality generalization capability of VTReID models. In the case of $G > 3$, increasing patch groups might reach performance saturation. On the bias of the above observations, we use $G = 3$ in the following experiments as default.

Analysis of MPMA: In view of modality discrepancies, we verify the impact of the proposed MPMA on RegDB dataset. As shown in Fig. 6, MPMA achieves the superior results than any other methods, which indicates that MPMA is greatly beneficial to learn modality-robust person representations. Compared with MPMA, we find that directly aligning modality on original feature space is unable to fully reduce the modality discrepancies

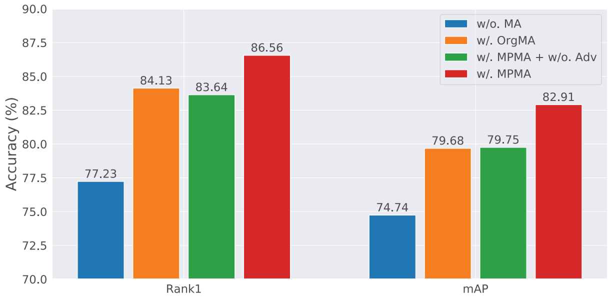


Fig. 6. Ablation study for MPMA on RegDB dataset. “w/o. MA” denotes that any modality alignment methods are not used. “w/. OrgMA” denotes that we directly align modality for original feature space. “w/. MPMA” denotes that the proposed MPMA method is adopted. “w/. MPMA + w/o. Adv” denotes we do not apply the adversarial learning of feature subspace and only minimize Eq. 14, resulting in ignoring the maximum process of Eq. 14.

and obtains inferior retrieval accuracies. In addition, without using adversarial learning, VTReID models achieve significantly inferior accuracies than MPMA. The main reason is that the adversarial learning aims to mine hard feature subspaces and abandon easy feature subspaces, which contributes to balancing the modality discrepancies. In addition, we study the impact of the feature level for modality alignment in Fig. 9. Specifically, we compare the performance of modality alignment between middle features $\boldsymbol{z}^r_{i,j}$ and final features $\boldsymbol{f}^r_{i,j}$. The results show that middle features achieve superior accuracies final features. This observation illustrates that middle features possess more modality variations than final features because the dimension of middle features is larger than final features. In Table V, we also evaluate the effectiveness of multi-patch modality alignment. For experimental fairness, we extract identical multi-patch features but apply the modality alignment loss for different patch groups. Compared with the modality alignment for the single patch group, MPMA makes the absolute improvement in terms of rank1 and mAP accuracies. The results indicate the advantage of coarse-grained and fine-grained modality alignment in modality-robust feature learning to enhance cross-modality generalization of VTReID models.

Analysis of CPCD: Then, we analyze the contributions of the proposed CPCD on RegDB dataset. CPCD incorporates two cross-patch prior knowledges, i.e., positive and negative cross-patch correlations, into cross-modality metric learning. The patch-specific information is transferred to other patches and thus contributes to generating robust representations. From Fig. 7, we can observe two interesting phenomena. (1) VTReID models with using either positive or negative correlations outperform models without using any cross-patch correlations, achieving a significant improvement in rank1 and mAP accuracies. (2) Transferring both positive and negative correlations further brings an additional improvement on retrieval accuracies with the comparison to using a single cross-patch correlation. These observations illustrate that learning multi-patch features from positive and negative cross-patch correlations is able to enhance the modality-robustness capability of VTReID models.

Analysis of PAPA: Since multi-patch feature learning can be viewed as a specific multi-task learning problem, we focus on analyzing the effectiveness of the proposed PAPA. In Fig. 5(c), we plot a performance curve to visually explore the sensitivity of hyperparameter $\gamma \geq 0$. When $\gamma = 0$, all patch tasks are assigned identical loss weights, i.e., $\lambda_i = \lambda_j = 1$. When $\gamma > 0$, PAPA adaptively assigns more weights to the harder patch tasks than the easies ones to prioritize the difficult patch tasks. According to the results in Fig. 5(c), we observe that the large γ always performs better than the small one, which indicates that hard patch tasks are more important than easy patch tasks for discriminative representation learning. When $\gamma > 10$, increasing γ may bring an insignificant contribution for VTReID models, resulting in performance saturation. Based on the above observations, we recommend $\gamma = 10$ in the following experiments. Moreover, we also plot loss weight curves for different patch tasks in Fig. 11. During the training phase, the fine-grained patch tasks receive more average weights than coarse-grained patch tasks. Coarse-grained patch features might contain more person body information than fine-grained patch features, therefore learning fine-grained features is more harder than coarse-grained features.

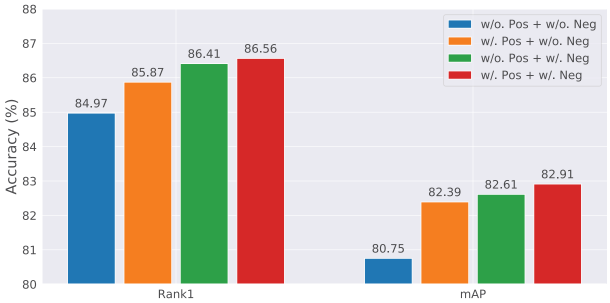


Fig. 7. Ablation study for CPCD on RegDB dataset. “Pos” and “Neg” represent positive and negative cross-patch correlations, respectively.

Analysis of Overall Framework: Table IV illustrates the effectiveness of the overall framework. The MPMA is proved to be efficacious for enhancing the cross-modality retrieval performance of the VTReID model. Some conclusions mentioned above can also be verified in Fig. 6 and Table V. For example, adversarial learning is conducive to mining hard subspaces and balancing modality discrepancies. The joint of coarse-grained and fine-grained modality alignment greatly enhances the cross-modality generalization capability of VTReID models. Additionally, We demonstrate that imposing cross-patch correlations by the CPCD into VTReID models is able to promote modality-robust feature learning. The similar observations can also be concluded in Fig. 7. Furthermore, the PAPA further boosts cross-modality retrieval performance by encouraging prioritization of hard patch tasks. Notably, the overall VTReID framework with MPMA, CPCD and PAPA reaches the best accuracies among all compositional models.

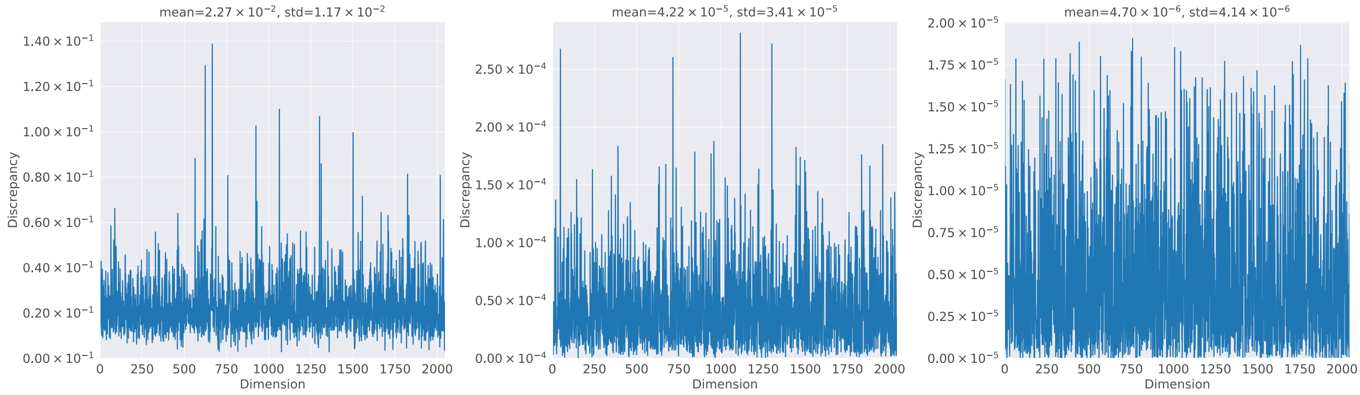
F. Visualization Analysis

Finally, we give a microscopic interpretation from the perspective of visualization analysis, which is a strong justification of our method. Besides, it also reveals the reason why mining hard subspaces is useful to modality-robust feature learning.

TABLE IV

ABLATION STUDIES AND RETRIEVAL ACCURACIES (%) OF PROPOSED METHODS ON REGDB AND SYSU-MM01 DATASETS. WE ANALYZE THE DIFFERENT COMPONENT COMBINATIONS FROM FOUR ASPECTS INCLUDING MP, MPMA, CPCD, AND PAPA. TL REPRESENTS TRIPLET LOSS AND MP DENOTES MULTI-PATCH FEATURES

TL	MP	MPMA	CPCD	PAPA	RegDB				SYSU-MM01			
					Rank1	Rank10	Rank20	mAP	Rank1	Rank10	Rank20	mAP
✓	✗	✗	✗	✗	53.88	73.69	82.18	55.50	32.71	79.32	90.00	47.69
✓	✓	✗	✗	✗	75.44	89.27	94.22	73.09	43.01	86.32	93.38	57.11
✓	✓	✓	✗	✗	83.40	94.17	95.58	79.17	45.79	88.70	95.68	60.07
✓	✓	✓	✓	✗	84.99	96.02	96.87	81.33	47.82	89.21	96.35	61.20
✓	✓	✓	✓	✓	**86.56**	**96.68**	**98.28**	**82.91**	**48.98**	**90.33**	**97.13**	**62.41**



(a) Without Modality Alignment **(b)** Modality Alignment for Original Space **(c)** Modality Alignment for Hard Subspace

Fig. 8. Modality discrepancies between query and gallery images on RegDB dataset. In each figure, the modality distribution can be viewed on the two-dimensional coordinates with the modality discrepancy of each dimension as the ordinate axis and the dimension index as the abscissa axis. The modality discrepancy of the i-th dimension is computed by $\boldsymbol{d}_i = (\boldsymbol{f}_v[i] - \boldsymbol{f}_t[i])^2$. $\boldsymbol{f}_v[i]$ and $\boldsymbol{f}_t[i]$ represent the value of the i-th dimension of visible and thermal features. The "mean" and "std" of each subfigure title represent the mean and standard deviation of all dimension discrepancies $\{\boldsymbol{d}_i\}_{i=1}^{2048}$. For simplifying experiments, we only visualize the modality discrepancy of the first patch feature after the GAP layer.

TABLE V

ACCURACIES (%) OF DIFFERENT PATCH GROUPS USED FOR MPMA. "0" REPRESENTS MODALITY ALIGNMENT IS NOT APPLIED FOR ANY PATCHES, WHILE "1" MEANS THAT MODALITY ALIGNMENT IS ONLY ADOPTED FOR THE FIRST PATCH GROUP. "1, 2, 3" DENOTES ALL PATCH FEATURES ADOPT MODALITY ALIGNMENT

Patch Group	RegDB			
	Rank1	Rank10	Rank20	mAP
0	77.23	90.97	94.76	74.74
1	80.83	94.22	96.21	80.45
2	84.51	94.42	96.41	79.90
3	84.37	94.13	96.55	79.57
1, 2	86.31	95.78	97.52	82.13
1, 3	84.66	94.64	96.73	80.38
2, 3	86.46	95.68	97.14	81.85
1, 2, 3	**86.56**	**96.68**	**98.28**	**82.91**

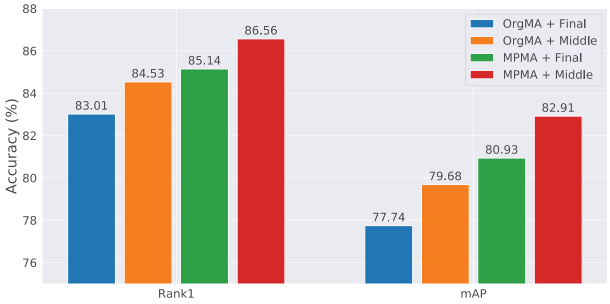
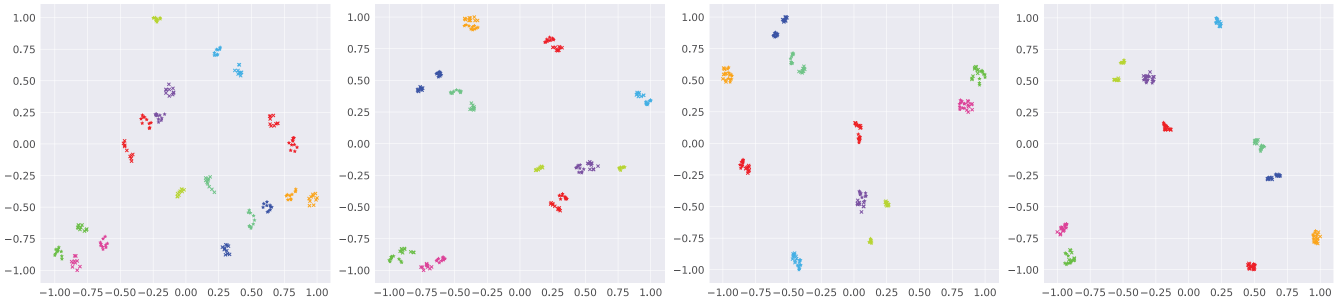


Fig. 9. Ablation study for the feature level on RegDB dataset. "Middle" and "Final" denote that the modality alignment is performed on the middle and final features, respectively.

Modality Discrepancy: In this part, we visualize the modality discrepancy of each dimension in Fig. 8. Compared with Fig. 8(a), Fig. 8(b) illustrates that aligning modality distributions is conducive to learning modality-invariant person representations. However, simply applying modality alignment for original feature space might suffer from the selective alignment behavior, i.e., only focuses on certain dimensions or subspaces which are the easiest ones to reduce the current distribution gap. For example, the modality discrepancies reach the peak from the 750-th to 1250-th dimension, while the remainder of dimensions possess relatively small modality discrepancies. We also find an interesting phenomenon that the modality discrepancies of Fig. 8(c) approximately follow uniform distribution pattern with very small mean and deviation values. This shows that the MPMA not only reduces the modality discrepancies with a significant margin, but also relieves the problem of the selective alignment behavior. Accordingly, mining the hard feature subspace by the MPMA is able to greatly enhance the robustness to the gap of modality distribution.



(a) MP

(b) MP + MPMA

(c) MP + MPMA + CPCD

(d) MP + MPMA + CPCD + PAPA

Fig. 10. The t-SNE [52] visualization of features on RegDB dataset. We randomly select 10 classes of testing data and extract person descriptors from the feature extractor. The classes are marked by different colors while the × and ⋆ markers represent visible and thermal modalities, respectively.

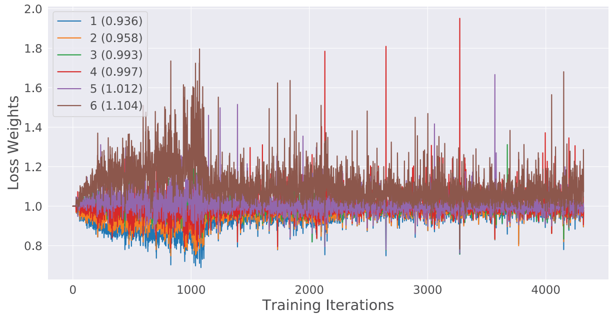


Fig. 11. The loss weights of different patch tasks on RegDB dataset. "1 (0.936)" represents that the average loss weight of the first patch is 0.936.

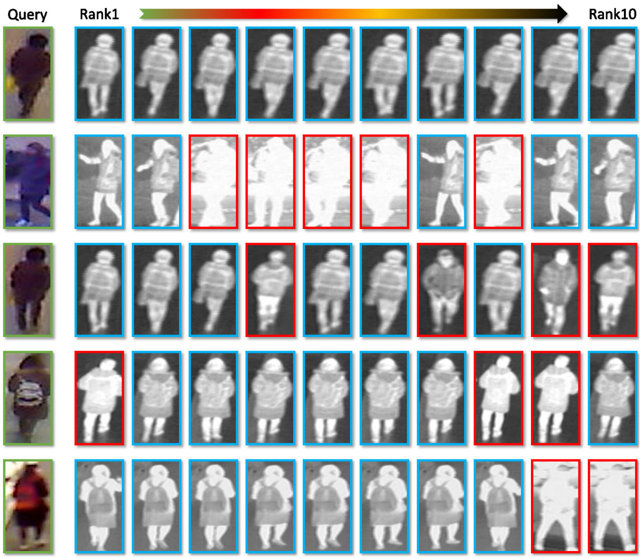


Fig. 12. Five query samples and their corresponding top-10 retrieval results on RegDB dataset (`visible to thermal`). The corrected retrieval samples are in blue boxes and wrong retrieval samples are in red boxes.

Feature Distribution: In this part, we analyze the deep features of visible and thermal modalities. Since it is hard to visualize high-dimensional vectors, we use t-SNE [52] to transform these feature vectors into two-dimensional vectors, as shown in Fig 10. Here, we can observe two interesting phenomena from Fig 10(a), (b), (c) and (d). For the first observation, most of classes in Fig 10(a) possess relatively large modality discrepancies without using the MPMA loss. It indicates that the backbone network is able to reduce intra-modality variations, but inter-modality variations require additional alignment algorithms. For the second observation, the joint of the CPCD loss and PAPA further relieves intra-modality and inter-modality variations.

Retrieval Ranks: We also visualize some retrieval results with five randomly selected query samples on RegDB dataset. The top-10 retrieval results together with query samples are shown in Fig. 12. The results show that the proposed method can achieve good performance if the person has rich structure (*e.g.*, bags or stripes). However, there are still some matching errors, which can be easily filtered by human beings.

V. Conclusion

In this work, we design a novel neural network framework named MPMN for cross-modality person re-identification. Apart from the heterogeneity of global coarse-grained semantics, we demonstrate that the joint of global coarse-grained and local fine-grained modality alignment is essential for a good cross-modality VTReID model. Specifically, we propose a new loss function termed MPMA to jointly balance and reduce the modality discrepancies of multi-patch features by mining hard subspaces and abandoning easy subspaces. Motivated by knowledge distillation, we put forward a new CPCD loss to transfer the semantic knowledges across different patches, which incorporates the positive and negative correlations from different patch features to improve retrieval performance. In addition, an PAPA method is further introduced to dynamically prioritize hard patch tasks during training, contributing to balancing multi-patch tasks. We have achieved the best performance on two public cross-modality person datasets and usually outperform the state-of-the-art methods by a significant margin. It provides new insights for the VTReID task in practical video surveillance systems.

References

[1] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 384–393.

[2] Y. Fu *et al.*, "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 8295–8302.

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

[4] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 402–419.

[5] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Trans. Multimedia* vol. 21, no. 6, pp. 1412–1424, Jun. 2019.

[6] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global–local-alignment descriptor for scalable person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, Apr. 2018.

[7] Y. Fu, X. Wang, Y. Wei, and T. Huang, "STA: Spatial-temporal attention for large-scale video-based person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8287–8294.

[8] H. Luo *et al.*, "AlignedReID++: Dynamically matching local information for person re-identification," *Pattern Recognit.*, vol. 94, pp. 53–61, 2019.

[9] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 403–412.

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

[11] S. Zhou *et al.*, "Large margin learning in set-to-set similarity comparison for person reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 593–604, Mar. 2017.

[12] R. Yu *et al.*, "Hard-aware point-to-set deep metric for person re-identification," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 188–204.

[13] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1092–1099.

[14] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/rt/captureCite/16734/0/BibtexCitationPlugin

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

[16] 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 *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 618–626.

[17] J. K. Kang, T. M. Hoang, and K. R. Park, "Person re-identification between visible and thermal camera images based on deep residual cnn using single input," in *IEEE Access*, vol. 7, pp. 57 972–57 984, 2019.

[18] Y. Hao, N. Wang, J. Li, and X. Gao, "HSME: Hypersphere manifold embedding for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 8385–8392.

[19] M. Ye, X. Lan, and Q. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 347–355.

[20] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 579–590, 2019.

[21] G. Wang *et al.*, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 3623–3632.

[22] P. Wang *et al.*, "Deep hard modality alignment for visible thermal person re-identification," in *Proc. Pattern Recognit. Lett.*, 2020, vol. 133, pp. 195–201.

[23] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[24] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 794–803.

[25] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7482–7491.

[26] D. Nguyen, H. Hong, K. Kim, and K. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, 2017, pp. 605–633.

[27] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5380–5389.

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

[29] F. Zheng *et al.*, "Pyramidal person re-identification via multi-loss dynamic training," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 8514–8522.

[30] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 7134–7143.

[31] P. Wang, F. Su, and Z. Zhao, "Joint multi-feature fusion and attribute relationships for facial attribute prediction," in *Proc. IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4.

[32] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 154–162.

[33] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 686–701.

[34] N. Murrugarra-Llerena and A. Kovashka, "Cross-modality personalization for retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 6429–6438.

[35] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 4654–4662.

[36] Z. Wang *et al.*, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 5764–5773.

[37] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5427–5440, Nov. 2016.

[38] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, "Heterogeneous face recognition: A common encoding feature discriminant approach," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2079–2089, May 2017.

[39] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein CNN: Learning invariant features for nir-vis face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1761–1773, Jul. 2018.

[40] R. Vemulapalli, H. Van Nguyen, and S. Kevin Zhou, "Unsupervised cross-modal synthesis of subject-specific scans," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 630–638.

[41] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/rt/captureCite/16241/0/BibtexCitationPlugin

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

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[44] P. Wang *et al.*, "Deep class-skewed learning for face recognition," in *Neurocomputing*, vol. 363, pp. 35–45, 2019.

[45] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 5814–5824.

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

[47] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[48] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Advances Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[49] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

[50] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 365–381.

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

[52] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

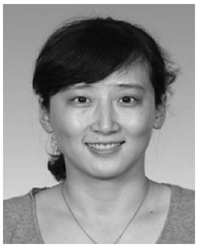
Pingyu Wang is currently a Ph.D. candidate with the Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include attribute classification, face recognition, person reidentification and computer vision.

Zhicheng Zhao is an associate professor with the Beijing University of Posts and Telecommunications. He was a visiting scholar with the School of Computer Science, Carnegie Mellon University from 2015 to 2016. His research interests are computer vision, image and video semantic understanding and retrieval. He has authored and coauthored more than 60 journal and conference papers.

Fei Su received the Ph.D. degree majoring in communication and electrical Systems from BUPT in 2000. She is a female professor in the multimedia communication and pattern recognition laboratory, school of information and telecommunication, Beijing university of posts and telecommunications. She was a visiting scholar at electrical computer engineering department, Carnegie Mellon University from 2008 to 2009, Her current interests include pattern recognition, image and video processing and biometrics. She has authored and co-authored more than 70 journal and conference papers and some textbooks.

Yanyun Zhao received the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2009. She is a female associate professor in the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. Her research interests include pattern recognition, image and video processing. She has authored and coauthored more than 60 journal and conference papers and some textbooks.

Haiying Wang received the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2009. She is a female associate professor in the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. Her current interests include image and video processing, computer vision. She has authored and coauthored more than 60 journal and conference papers and some textbooks.



Lei Yang received the Ph.D. degree majoring in pattern recognition and intelligent system from Xian Jiaotong University in 2010. She is a senior researcher in China Mobile Research Institute, China Mobile Communications Corporation. Her current interests include intelligent system and service of image and video.

Yang Li received the master.s degree in instruments science and technology from Tsinghua University in 2014. She is a researcher of Carrier Services Research Department, China Mobile Research Institute. Her current research interests include image and video processing and computer vision.