WARWICK
THE UNIVERSITY OF WARWICK

Manuscript version: Author's Accepted Manuscript
The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/178261

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

© 2023 Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/.



Publisher's statement:
Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

warwick.ac.uk/lib-publications

On Exploring Pose Estimation as an Auxiliary Learning Task for Visible-Infrared Person Re-identification

Yunqi Miao, Nianchang Huang, Xiao Ma, Qiang Zhang, and Jungong Han

***Abstract*—Visible-infrared person re-identification (VI-ReID) has been challenging due to the existence of large discrepancies between visible and infrared modalities. Most pioneering approaches reduce intra-class variations and inter-modality discrepancies by learning modality-shared and ID-related features. However, an explicit modality-shared cue, *i.e.,* body keypoints, has not been fully exploited in VI-ReID. Additionally, existing feature learning paradigms imposed constraints on either global features or partitioned feature stripes, which neglect the prediction consistency of global and part features. To address the above problems, we exploit Pose Estimation as an auxiliary learning task to assist the VI-ReID task in an end-to-end framework. By jointly training these two tasks in a mutually beneficial manner, our model learns higher quality modality-shared and ID-related features. On top of it, the learnings of global features and local features are seamlessly synchronized by Hierarchical Feature Constraint (HFC), where the former supervises the latter using the knowledge distillation strategy. Experimental results on two benchmark VI-ReID datasets show that the proposed method consistently improves state-of-the-art methods by significant margins. Specifically, our method achieves nearly 20% mAP improvements against the state-of-the-art method on the RegDB dataset. Our intriguing findings highlight the usage of auxiliary task learning in VI-ReID. Our source code is available at https://github.com/yoqim/Pose_VIReID.**

***Index Terms*—visible-infrared person re-identification, auxiliary learning task.**

I. Introduction

Person re-identification (ReID) aims at retrieving the same identity across multiple disjoint cameras, which has gained much attention from the recent computer vision community [1]–[5]. Most existing ReID methods focus on the matching between visible images, which are generally collected under good illumination conditions [6]–[10]. However, those systems seem impractical because visible images cannot provide sufficient discriminatory information in poor lighting environments, *e.g.,* at night. To this end, Visible-Infrared person Re-identification (VI-ReID) emerges as an alternative to performing the retrieval between visible (RGB) images and infrared (IR) counterparts, thus enabling the day-to-night person re-identification.

Y. Miao and X.Ma are with the Warwick Manufacturing Group (WMG), University of Warwick, Coventry, CV4 7AL, United Kingdom (e-mail: Yunqi.Miao.1@warwick.ac.uk; X.Ma@warwick.ac.uk).

N. Huang and Q.Zhang are with the School of Mechano-Electronic Engineering, Xidian University, Xi'an, 710071, China (email: nchuang@stu.xidian.edu.cn; qzhang@xidian.edu.cn).

J. Han is with the Department of Computer Science, Aberystwyth University, Wales, SY23 3FL, United Kingdom (e-mail: jungonghan77@gmail.com).

However, VI-ReID is a challenging problem due to large intra-class variations and modality discrepancies across different cameras. The former refers to identity's appearance differences within a modality caused by poses, clothes, viewpoints, *etc*. While, the latter denotes intrinsic differences between visible and infrared images caused by the spectrum of cameras. To reduce both discrepancies, the central research question in this field has always been seeking better ways to extract discriminative features for identity retrieval, which are ID-related and modality-invariant [2], [11]–[13].

Despite the vigorous development of this field, we realize that most algorithms extract identity features in a heuristics manner with NO common view of what sort of features are specifically helpful for VI-ReID. To tackle this problem, we visualize the features extracted by several representative methods in Fig. 1, aiming to investigate how visual feature extractions evolved over the years to improve the performance of VI-ReID systems. Concretely, ZeroPad [11], as the first work in VI-ReID, extracts features from random regions in an image. Additionally, for a specific identity, the features derived from two modalities share NO commonality. As a result, the mean Average Precision (mAP) score is far from satisfactory. Later, TSLFN [12] horizontally partitions the backbone feature maps (global) into several stripes (local) and employs local-level constraints on each of them. Clearly, features derived by TSLFN cover more parts of the human body, compared to the global-level constraint-based method, *e.g.,* ZeroPad. That might be the reason why its performance is significantly superior to that of ZeroPad. In the meantime, a baseline for VI-ReID (AGW [2]) inserts non-local attention blocks during the feature extraction, which enforces the features to be extracted from identity's body instead of backgrounds. It reveals that *attention-aware* features help to increase performance. Recently, a state-of-the-art approach (DDAG [13]) integrates both local-level and global-level constraints into an end-to-end framework. Compared to previous works, the features derived by DDAG are more fine-grained, which are not only shared by two modalities but distinguishable for different identities. Based on the above observations, a conclusion can be drawn: *as more features from the human region (modality-shared) and attentive body part (ID-related) features are extracted, the retrieval performance improves consistently.* Benefiting from this conclusion, our algorithm extracts more features from body skeleton joints, which are not only ID-related but immune to modality changes. Therefore, we achieve over 10% mAP improvements against DDAG on the challenging SYSU-

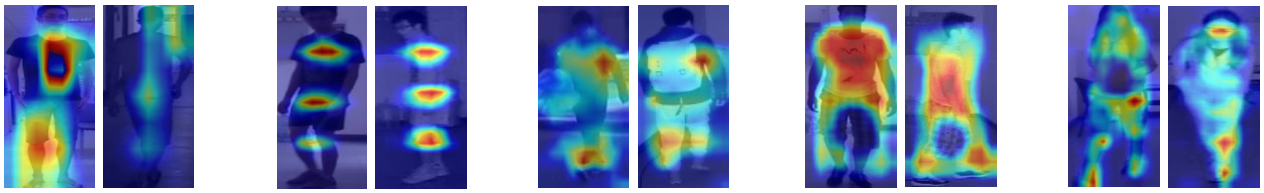


Fig. 1. Visualization of features derived by ZeroPad [11], TSLFN [12], AGW [2], DDAG [13], and our method on SYSU-MM01 dataset. For each method, the RGB image of an identity is shown on the left and the IR image is on the right. The publication year and mAP score (%) at *All-search* setting of each method are also reported. For a fair comparison, the mAP scores of TSLFN, AGW and our method are given by models trained merely with vanilla identity loss and triplet loss.

MM01 dataset at *All-search* setting.

Having depicted the visual features that are conducive to VI-ReID, the next question is: how can we extract them effectively. It is noted that body skeleton points are explicit modality-shared cues and the features describing certain skeleton points are ID-related. In light of this, in the paper, we aim to facilitate the extraction of discriminative features for identity retrieval with the aid of the pose estimation task. However, making effective use of the pose information for cross-modality ReID does not seem easy, though it has recently appeared to be exploited in some single-modality ReID works [14]–[19], where only visible images are involved. Earlier methods [14]–[16] utilize detected body joints to segment [14] or align [15], [16] body regions in order to cope with human pose changes. After the calibration of body parts, different body parts need to be stitched, which usually yields unrealistic transformed visual features. If we move to the cross-modality setting, the transformed errors would be further magnified due to the huge discrepancy between the two modalities. Alternatively, another group of methods employ the body keypoints information to refine ID-related feature maps either by means of highlighting discriminative body regions [17], [19] or complementing human appearance features [18]. Although pose-assisted features are proved to improve the feature discriminability under the single-modality setting, they are rigidly based on the outputs of off-the-shelf pose estimators. However, such a blind trust in the pre-trained pose estimators will lead to a poor re-identification performance if the gap between the source domain and the target domain of the pose estimator is huge. Therefore, employing pose information in the VI-ReID task is extremely challenging due to the massive gap between visible images (source domain) and infrared images (target domain). To highlight the problem, we extend two state-of-the-art pose-guided single-modality person ReID methods [18], [19] to the VI-ReID task and the results turn out that the best performance in the mAP score [18] only reaches 42.19%, which is far from satisfaction due to the inadequate usage of the pose information (more results and comparisons are provided in Section IV-C).

To solve the above problems, we propose a two-stream VI-ReID framework, where modality-shared and ID-related features for identity retrieval are extracted by means of learning an auxiliary task (pose estimation) and the main task (person ReID) jointly. Unlike previous works, which rely dramatically on the off-the-shelf pose estimators, pose features are adaptively adjusted to facilitate the ReID task in our work. Additionally, apart from ID-related constraints, an extra constraint is imposed on the pose estimation branch, ensuring that not only body skeleton points are precisely estimated but also the ID-related information are fully embedded in feature maps, *i.e.,* at both local and global levels. Despite the significant improvements obtained by the horizontal-divided feature constraints [20] in the VI-ReID task [12], [13], the learning of individual striped features is generally independent and its discriminability consistency with global backbone features is neglected. To this end, we propose a Hierarchical Feature Constraint (HFC), in the paper, which bonds the learnings of global features and local ones via the knowledge distillation strategy. Concretely, predictions of backbone features serve as "soft-target" to supervise the learning of partitioned feature stripes, hence preserving the discriminability consistency of global features and local ones.

In summary, the contributions made in our work are mainly three-fold:

- A novel two-stream framework for VI-ReID is proposed, where the pose estimation, for the first time, acts as an auxiliary learning task to help the ReID task in VI-ReID. To learn fine-grained pose features embedded with ID-related information, both pose and ReID constraints are imposed on the pose estimation branch.
- Instead of imposing feature constraints on local feature stripes only, Hierarchical Feature Constraint (HFC) is proposed to ensure the discriminability consistency of global features and local ones via the knowledge distillation strategy.
- The proposed method performs far better than the state-of-the-art methods on two benchmark datasets: SYSU-MM01 [11] and RegDB [21].

II. Related work

A. Auxiliary tasks in person ReID

Research has shown that semantic information such as body parts, human pose *etc*., can significantly facilitate the person ReID task [1], [2]. Therefore, recent works utilize auxiliary tasks to improve the performance of ReID models, which can be mainly categorized as attribute-guided methods

[22], [23], segmentation-guided methods [24]–[26], and pose-guided methods [14]–[18], [27]–[29].

Attribute information provides complementary details of identities for person ReID. Ling *et al*. [22] propose a multi-task learning framework, which attempts to improve the discriminability of identity features by embedding attribute information. Tay *et al*. [23] enhance the identity features with attribute attention maps, where class-sensitive activation regions of various attributes, *i.e.,* clothing color, hair, gender *etc*. are emphasized.

Segmentation-guided methods are based on the pixel-level body parts segmentation, which improve the discriminability of identity features by masking out backgrounds [25] or leverage fine-grained local features from discriminative body regions [24], [26]. However, both attribute- and segmentation-assisted methods require extensive additional annotations, which are too expensive to obtain in real-world applications. Additionally, since color information is not available in infrared images, the usage of color-related attributes is restricted in VI-ReID, thereby limiting the performance improvement.

The pose estimation is employed to facilitate person ReID from mainly two aspects: 1) generating person images with various poses to augment training data [27]–[29], and 2) aligning body parts [14]–[19]. With the help of Generative Adversarial Networks (GAN), Qian *et al*. [29] synthesize eight new images for an identity with a set of target canonical poses. In addition to generating visually preferable samples, Liu *et al*. [28] propose a guider module to ensure that the generated samples have discriminative power for ReID. Apart from image generation, pose information is employed to handle the problem of feature misalignment in ReID [14]–[16]. Zhao *et al*. [14] segment the human body into seven regions by human landmark information and combine the representations over them for identity retrieval. Zheng *et al*. [16] align pedestrians to a standard pose by stitching the segmented body regions with affine transformations. However, such unnatural stitch destroys the authenticity of human and requires an elaborately designed following fusion method to fuse the local features. Instead of using detected body regions rigidly, Zhao *et al*. [17] and Miao *et al*. [19] employ body joint maps to refine image feature maps, where discriminative body parts for person ReID are emphasised. However, both approaches are rigidly based on the outputs of off-the-shelf pose estimators, which may generate unreliable information because of the gap between the source domain and the target domain. To avoid the blind trust in the pre-trained pose estimator, Suh *et al*. [18] take an on-the-fly pose estimator as an individual branch to derive pose features, which are then aggregated with the appearance features from the other branch via a bilinear pooling layer. Such combined training encourages the pose branch to learn features that are beneficial to the person ReID task. While pose features are learned to adapt to the ReID task in the work, only ID-related constraints are imposed on the fused features, which ignores the quality of pose features from the individual branch. Given the above concerns, the auxiliary pose estimation task designed for single-modality ReID cannot perform well in the VI-ReID task. In other words, how to enable the pose information to facilitate the VI-ReID task has not been elaborately investigated.

Therefore, in the paper, we propose a novel pose estimation assisted framework for VI-ReID, where body skeleton points cues are learned under both pose and identity guidance, which are then employed to enhance the discriminative identity feature extraction for the cross-modality person re-identification.

B. Feature constraints in VI-ReID

To improve the discriminability of learned features, most existing works impose either global-level feature constraints on backbone convolutional features or local-level feature constraints on partitioned feature stripes.

As a representative of global-level constraints, Ye *et al*. [30] propose a two-stream network, which jointly optimizes modality-specific and modality-shared metrics. Based on the idea, a bi-directional top-ranking loss is then introduced in [31] to incorporate the above two constraints. Alternatively, AGW [2] presents a weighted regularized triplet loss to embed the neighboring relationship of images from two modalities in a common feature space.

Inspired by the competitive performance of the Part-based Convolutional Baseline (PCB) model [20] in single-modality ReID, recent VI-ReID studies start imposing local-level feature constraints on feature stripes obtained by partitioning backbone convolutional features. a kickoff work [12] presents a center-based loss that pulls the centers of RGB features and IR features of a given identity closer. On top of it, Liu *et al*. [32] present a hetero-center triplet loss, where the feature centers of different identities are pushed away. Recently, the newly-proposed DDAG [13] simultaneously mines both cross-modality global-level and intra-modality part-level contextual cues, which achieves the state-of-the-art performance. Although both global-level and local-level constraints are considered in DDAG, those constraints are imposed independently without taking the discriminability consistency of them into account.

To this end, we propose a Hierarchical Feature Constraint (HFC) to bond the global feature learning with the local ones, where predictions of global features supervise the learning of part features via the knowledge distillation strategy.

III. Proposed method

We propose a novel pose estimation assisted framework for VI-ReID, which aims to learn modality-shared and ID-related features for identity retrieval. The framework of the proposed method is shown in Fig. 2. As can be seen, our model mainly consists of four components: Modality-specific module, Modality-shared module, Pose Estimation branch, and ReID branch. Details about these components will be discussed in the following subsections.

A. Modality-specific module and Modality-shared module

Following previous works [12], [31], [32], ResNet50 [33] is exploited as a backbone feature extractor to provide discriminative features for both pose estimation and ReID tasks. Specifically, Modality-specific module consists of two blocks

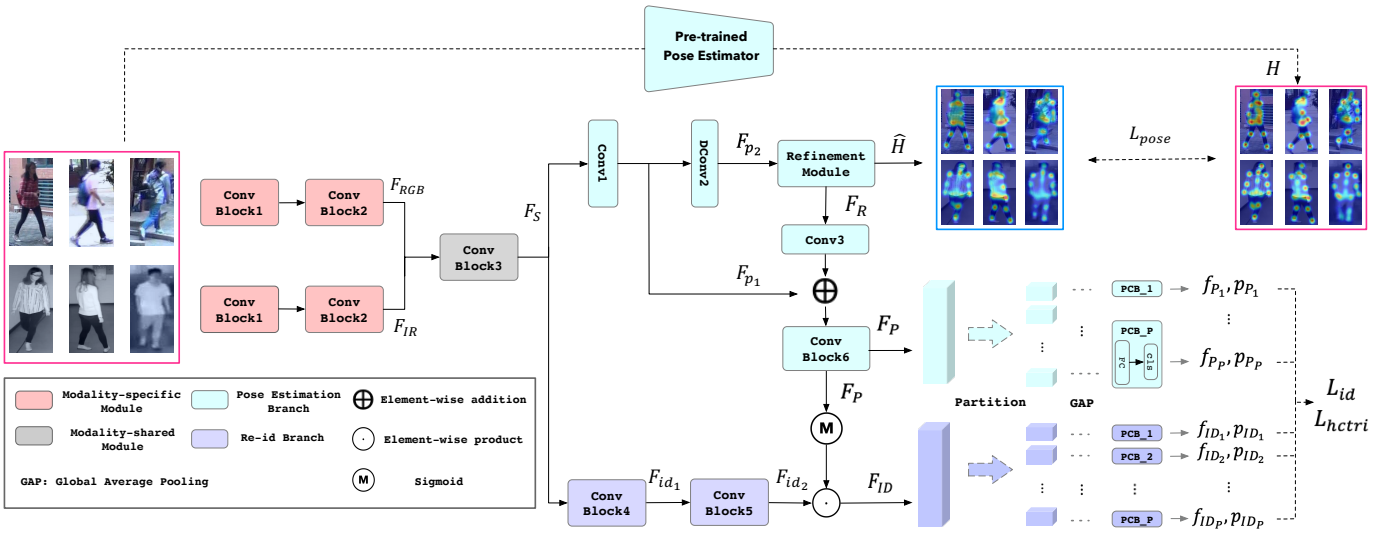


Fig. 2. The framework of the proposed method.

(*"Conv Block1~2"*), which adopt the structures of shallow convolution block (*layer0*) and the first res-convolution block (*layer1*) of ResNet50, respectively. Note that the parameters of Modality-specific module for RGB and IR modalities are separately updated. Then, modality-specific features of two modalities are projected into a shared feature space by the Modality-shared module (*"Conv Block3"*), which adopts the structure of the second res-convolution block (*layer2*) of ResNet50.

Mathematically, given a RGB image I_{RGB} and an IR image I_{IR}, modality-specific features $\mathbf{F}_m$ and modality-shared features $\mathbf{F}_S$ can be obtained by,

$$\begin{aligned}\mathbf{F}_m &= ConvB2(ConvB1(I_m, \theta_m^1), \theta_m^2), \quad m \in \{RGB, IR\},\\ \mathbf{F}_S &= ConvB3([\mathbf{F}_{RGB}, \mathbf{F}_{IR}]_b, \theta^3),\end{aligned} \tag{1}$$

where $[\cdot,\cdot]_b$ represents the feature concatenation along the data dimension. $ConvB1(*, \theta_m^1)$, $ConvB2(*, \theta_m^2)$, and $ConvB3(*, \theta^3)$ denote the convolution blocks of Modality-specific module and Modality-shared module with corresponding parameters θ_m^1, θ_m^2 and θ^3, respectively.

B. Pose Estimation branch

1) Body keypoint features extraction: Given the observation that modality-shared features are beneficial to the VI-ReID task, a Pose Estimation branch is integrated as an auxiliary to extract modality-shared features. The structure of our Pose Estimation branch is shown in Fig. 2. Specifically, a convolutional layer (*"Conv1"*) and a deconvolutional layer (*"DConv2"*) are employed to extract high-level features, *i.e.,* $\mathbf{F}_{p_1}$ and $\mathbf{F}_{p_2}$, and restore the resolution of feature maps to that of ground-truth body keypoint heatmaps, which are denoted as follows,

$$\mathbf{F}_{p_1} = Conv1(\mathbf{F}_S, \theta_P^1), \quad \mathbf{F}_{p_2} = DConv2(\mathbf{F}_{p_1}, \theta_P^2), \tag{2}$$

where $Conv1(*, \theta_P^1)$ and $DConv2(*, \theta_P^2)$ represents a 3×3 convolutional layer with parameters θ_P^1 and θ_P^2, respectively. Note that both layers are followed by a ReLU activation function, which are omitted in equations for simplicity.

Subsequently, a Refinement Module [34] is used to extract refined body keypoint features and predict body keypoint heatmaps. Specifically, the Refinement Module consists of a U-Shaped Block, three Refine Blocks and two convolutional layers. The U-Shaped Block and Refine Blocks are employed to extract refined features. On top of it, the convolutional layers are applied for the heatmap estimation. Mathematically, given high-level features $\mathbf{F}_{p_2}$, refined keypoint features $\mathbf{F}_R$ and predicted heatmaps $\hat{\mathbf{H}}$ can be respectively obtained by,

$$\mathbf{F}_R = RM_F(\mathbf{F}_{p_2}, \theta_{RM}^F), \tag{3}$$

$$\hat{\mathbf{H}} = RM_H(\mathbf{F}_R, \theta_{RM}^H), \tag{4}$$

where $RM_F(*, \theta_{RM}^F)$ and $RM_H(*, \theta_{RM}^H)$ denote the refined feature extraction stage with parameters θ_{RM}^F, and the heatmap estimation stage with parameters θ_{RM}^H in the Refinement Module, respectively.

2) Body keypoint features transferring: To exploit body keypoint features derived by Pose Estimation branch in the ReID branch, a convolutional layer (*"Conv3"*) and a convolutional block (*"Conv Block6"*) are employed to deal with the mismatch in terms of the resolution and the channel number of feature maps.

Specifically, the refined keypoint features $\mathbf{F}_R$ are firstly downsampled by *"Conv3"* so that the derived $\mathbf{F}'_R$ has the identical resolution as $\mathbf{F}_{p_1}$, which is denoted as,

$$\mathbf{F}_{R'} = Conv3(\mathbf{F}_R, \theta_P^3), \tag{5}$$

where $Conv3(*, \theta_P^3)$ represents the 3×3 convolutional layer with a stride of 2 and parameters θ_P^3. On top of that, *"Conv Block6"* aligns the channel number of feature maps from the Pose Estimation branch with that from the ReID branch. Therefore, the final body keypoint features $\mathbf{F}_P$ can be derived as follows,

$$\mathbf{F}_P = ConvB6(\mathbf{F}_{R'} + \mathbf{F}_{p_1}, \theta_P^4), \tag{6}$$

where $ConvB6(*, \theta_P^4)$ denotes the convolutional block with the parameters θ_P^4, consisting of a 3×3 convolutional layer with a stride of 2 and a 1×1 convolutional layer. Note that all convolutional layers are followed by a ReLU activation function.

3) Body keypoint features integration: To highlight the body keypoint regions in the features output by the ReID branch, the final keypoint features $\mathbf{F}_P$ are employed to generate the body keypoint masks $\mathbf{M}$, *i.e.,*

$$\mathbf{M} = sigmoid(\mathbf{F}_P). \tag{7}$$

The values of the masks are regularized by the sigmoid function to [0, 1], which serve as soft attention maps to refine the identity features from the ReID branch.

C. ReID branch

1) Global-level feature extraction: Apart from the Pose Estimation branch that aids modality-shared features extraction, a ReID branch is employed to extract ID-related features. As can be seen from Fig. 2, the ReID branch mainly consists of 2 convolutional blocks (*"Conv Block4~5"*). Following previous works [2], [12], [31], *"Conv Block4"*, *"Conv Block5"* follow the structures of the third and fourth res-convolution block (*layer3*, *layer4*) of ResNet50 [33], respectively. Mathematically, identity features $\mathbf{F}_{id_2}$ are extracted by,

$$\mathbf{F}_{id_2} = ConvB5(ConvB4(I_m, \theta_{ID}^1), \theta_{ID}^2), m \in \{RGB, IR\}, \tag{8}$$

where $ConvB4(*, \theta_{ID}^1)$ and $ConvB5(*, \theta_{ID}^2)$ denote a convolutional block with the parameters θ_{ID}^1 and θ_{ID}^2, respectively. Then, identity features $\mathbf{F}_{id_2}$ are refined by the body keypoint masks $\mathbf{M}$ derived by Pose Estimation branch by performing the element-wise product operation $\odot$ to obtain the final identity features $\mathbf{F}_{ID}$, *i.e.,*

$$\mathbf{F}_{ID} = \mathbf{F}_{id_2} \odot \mathbf{M}. \tag{9}$$

2) Local-level feature partition: Since part features can offer fine-grained information for identity identification, PCB models [20] are exploited in the proposed framework for local feature learning. Following [12], [13], convolutional features $\mathbf{F}_{ID}$ from ReID branch are firstly partitioned into P horizontal stripes and then transferred to feature vectors via Global Average Pooling (GAP) before being sent to the corresponding PCB model, which can be formulated as follow,

$$\mathbf{F}_{ID_1}, ..., \mathbf{F}_{ID_P} = GAP\big(Part(\mathbf{F}_{ID})\big), \tag{10}$$

where $Part(\cdot)$, $GAP(\cdot)$ denote the horizontal partition and GAP, respectively.

As can be seen from Fig. 2, a PCB model consists of a Fully-Connected (FC) layer and a classifier. The former reduces the dimensions of feature vectors from 2048-dim to 512-dim, and the latter is employed for identity prediction. For the i-th PCB model, the fine-grained part features f_{ID_i} are obtained by,

$$f_{ID_i} = FC_{ID_i}(\mathbf{F}_{ID_i}, \theta_{fc_{id}}^i), \quad i = \{1, ..., P\}, \tag{11}$$

where $FC_{ID_i}(*, \theta_{fc_{id}}^i)$ denotes the FC layer with the parameters $\theta_{fc_{id}}^i$. P is the number of PCB models, which is empirically set as 6 in the paper.

The local feature learning is also performed on convolutional features $\mathbf{F}_P$ from Pose Estimation branch to obtain the corresponding fine-grained part features f_{P_i}. During inference, fine-grained part features are concatenated for the identity retrieval, *i.e.,*

$$f_{ID} = [f_{ID_1}, ..., f_{ID_P}]_c, \quad f_P = [f_{P_1}, ..., f_{P_P}]_c, \tag{12}$$

$$f_{ALL} = [f_{ID}, f_P]_c, \tag{13}$$

where $[\cdot, ..., \cdot]_c$ represents the feature concatenation along the channel dimension.

D. Hierarchical Feature Constraint

To ensure the discriminability consistency of global and local features, Hierarchical Feature Constraint (HFC) is proposed to bond the learnings of global features and local ones. The structure of HFC is illustrated in Fig. 3, which is inspired by the Teacher-Student learning spirit in Knowledge Distillation (KD) [35]. As can be seen, instead of introducing an additional pre-trained teacher model, the predictions of convolutional features are employed as "soft-targets" to provide an *extra* supervision for *"Student"* models, *i.e.,* PCB models of Pose Estimation branch and ReID branch.

Specifically, convolutional features of Pose Estimation branch and ReID branch are firstly concatenated along channel dimension and then transferred to feature vectors by Global Average Pooling. Then, a PCB model (*"PCB_T"*) is employed to obtain "soft-targets", *i.e.,* $P_T = \{p_T^i\}_{i=1}^N$. N refers to the number of training images. Formally, given an image I_i with the identity label y_i, p_T^i can be obtained as follows,

$$p_T^i = p(y_T^i = y_i | I_i) = \frac{\exp(y_T^i)}{\sum_{k=1}^{N_{id}} \exp(y_T^k)}, \tag{14}$$

where $y_T^k = w_{t_k}^T f_T^i$, f_T^i is the fine-grained feature of the i-th image output by the FC layer in *"PCB_T"*. w_{t_k} indicates the parameter of the classifier in *"PCB_T"* for the k-th identity. N_{id} is the number of identities in the whole training set. For the j-th PCB model of Pose Estimation branch and ReID branch, the corresponding probability predictions, *i.e.,* $P_{P_j} = \{p_{P_j}^i\}_{i=1}^N$ and $P_{ID_j} = \{p_{ID_j}^i\}_{i=1}^N$, can be calculated in the same way, respectively.

In order to supervise the local feature learning with the global one, KD loss L_{KD} is employed to reduce the distance between two prediction distributions, *i.e.,* P_T and $P_{ID(P)_j}$. Given a mini-batch with M images, L_{KD} is formulated as follows,

$$L_{KD} = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{P} \big(KL(p_T^i, p_{ID_j}^i) + KL(p_T^i, p_{P_j}^i) \big), \tag{15}$$

where $KL(p, q)$ measures the Kullback-Leibler divergence between distribution p and distribution q. P denotes the number of PCB models of each branch.

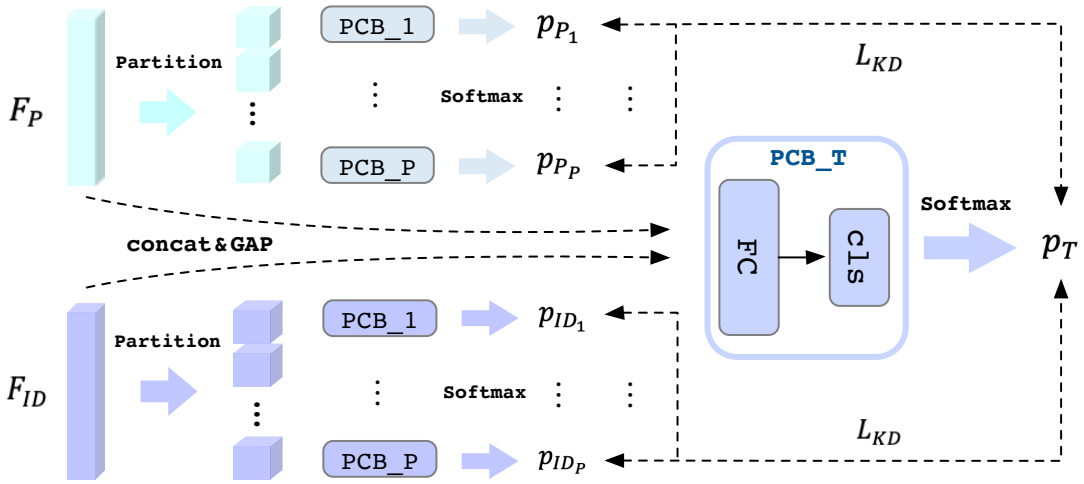


Fig. 3. Illustration of Hierarchical Feature Constraint (HFC). GAP represents Global Average Pooling.

E. Loss functions

Batch sampling method Following [12], [13], an online batch sampling strategy is adopted during training. Specifically, D identities are randomly selected at each iteration. For each identity, K RGB images and K IR images are then randomly selected to build the mini-batch. Therefore, the batch size $M = 2 * DK$. In the paper, we set $D = 8$, $K = 4$, and $M = 64$ during training.

Pose estimation loss To encourage Pose Estimation branch to learn modality-shared features, pose estimation loss L_{pose} is introduced to minimize pixel-wise Euclidean distances between ground-truth body keypoint heatmaps and the predicted ones. In the paper, ground-truth heatmaps are derived by a pose estimation model [34], which is pre-trained on the LIP dataset [36]. Formally, L_{pose} across a mini-batch is defined by,

$$L_{pose} = \frac{1}{M}\sum_{i=1}^{M}\sum_{x,y}(H_i(x,y) - \hat{H}_i(x,y))^2, \tag{16}$$

where $H_i(x,y)$, $\hat{H}_i(x,y)$ represent the pixel value at (x,y) position of the i-th ground truth and the predicted body keypoint heatmap, respectively.

Identity loss To extract ID-related features, identity loss is performed on each PCB model of both Pose Estimation branch and ReID branch. Given the probability predictions P_{P_j} and P_{ID_j} given by the j-th PCB model of Pose Estimation branch and ReID branch, respectively, identity loss L_{id} across a mini-batch is formulated as,

$$L_{id} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{P}(\log p^i_{P_j} + \log p^i_{ID_j}), \tag{17}$$

where P denotes the number of PCB models of each branch.

Hetero-center triplet loss To reduce both intra- and inter-modality discrepancies, hetero-center triplet (HC-tri) loss [32] is also employed for local feature learning. Similar to identity loss, HC-tri loss is also performed on each PCB model. Specifically, for each PCB model, the centers of fine-grained part features of the i-th identity in a given mini-batch from RGB modality $c^{RGB}_{i,j}$ and IR modality $c^{IR}_{i,j}$ can be computed as follows,

$$c^{RGB}_{i,j} = \frac{1}{K}\sum_{j=1}^{K} f^{RGB}_{i,j}, \quad c^{IR}_{i,j} = \frac{1}{K}\sum_{j=1}^{K} f^{IR}_{i,j}, \quad i = \{1, ..., D\}, \tag{18}$$

where $f^{RGB(IR)}_{i,j}$ represents the fine-grained part feature of the j-th RGB (IR) image in the given mini-batch. Therefore, HC-tri loss for the z-th PCB model of Pose Estimation branch is computed as follows,

$$\begin{aligned} L^{P_z}_{hctri} = &\sum_{i=1}^{D}[\rho + ||c^{RGB}_i - c^{IR}_i||_2 - \min_{\substack{m\in\{RGB,IR\}\\ j\neq i}} ||c^{RGB}_i - c^m_j||_2]_+ \\ &+ \sum_{i=1}^{D}[\rho + ||c^{IR}_i - c^{RGB}_i||_2 - \min_{\substack{m\in\{RGB,IR\}\\ j\neq i}} ||c^{IR}_i - c^m_j||_2]_+, \end{aligned} \tag{19}$$

where $[x]_+ = max(x, 0)$. ρ refers to the margin value, which is empirically set as 0.3. HC-tri loss of the $z-th$ PCB model of ReID branch, *i.e.*, $L^{ID_z}_{hctri}$, can be calculated in the same way. Therefore, the overall HC-tri loss across the given mini-batch can be derived as follows,

$$L_{hctri} = \sum_{j=1}^{P}(L^{P_j}_{hctri} + L^{ID_j}_{hctri}). \tag{20}$$

The overall objective for training is defined as,

$$L = L_{id} + \beta L_{hctri} + \lambda L_{pose} + \gamma L_{KD}, \tag{21}$$

where β, λ, and γ are the weighting factors to balance each loss term, which are empirically set as 0.1, 5, 1 in the paper, respectively.

IV. Experiments

A. Datasets, evaluation metrics and implementation details

Datasets Two benchmark datasets (SYSU-MM01 [11] and RegDB [21]) are employed to evaluate the performance of the proposed method.

SYSU-MM01 [11] consists of images captured by 6 cameras, including 2 IR cameras and 4 RGB ones (2 outdoors

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE **SYSU-MM01** DATASET. RANK-R ($r = 1, 10, 20$) ACCURACY(%) AND MAP(%) ARE REPORTED. $Ours_{ID}$ AND $Ours_{ALL}$ DENOTE THE FEATURES USED FOR EVALUATION ARE OBTAINED FROM REID BRANCH AND BOTH BRANCHES, RESPECTIVELY.

Method	Venue	All-search				Indoor-search			
		Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Zero-Pad [11]	ICCV2017	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
HCML [30]	AAAI2018	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
cmGAN [37]	IJCAI2018	26.97	67.51	80.56	31.49	31.63	77.23	89.18	42.19
eDBTR [31]	TIFS2019	27.82	67.34	81.34	28.42	32.46	77.42	89.62	42.46
HSME [38]	AAAI2019	20.68	32.74	77.95	23.12	-	-	-	-
D^2RL [39]	CVPR2019	28.90	70.60	82.40	29.20	-	-	-	-
MSR [40]	TIP2019	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
AlignGAN [41]	ICCV2019	42.40	85.00	93.70	40.70	45.90	87.60	94.40	54.30
TSLFN [12]	Neuro2020	56.96	91.50	96.82	54.95	59.74	92.07	96.22	64.91
AGW [2]	Arxiv2020	47.50	-	-	47.65	54.17	-	-	62.97
X-Modal [42]	AAAI2020	49.92	89.79	95.96	50.73	-	-	-	-
MACE [43]	TIP2020	51.64	87.25	94.44	50.11	57.35	93.02	97.47	64.79
DDAG [13]	ECCV2020	54.75	90.36	95.81	53.02	61.02	94.06	98.41	67.98
cm-SSFT [44]	CVPR2020	61.60	89.20	93.90	63.20	70.50	94.90	97.70	72.60
NFS [45]	CVPR2021	56.91	91.34	96.52	55.45	62.69	96.53	99.07	69.79
CICL [46]	AAAI2021	57.2	94.3	98.4	59.3	66.6	98.8	99.7	74.7
GLMC [47]	TNNLS2021	64.37	93.90	97.53	63.43	67.35	98.10	99.77	74.02
LbA [48]	ICCV2021	55.41	-	-	54.14	58.46	-	-	66.33
$Ours_{ID}$	-	65.82	94.53	98.23	64.52	71.74	94.57	97.60	74.54
$Ours_{ALL}$	-	**71.21**	**95.35**	**98.81**	**67.15**	**72.55**	**97.15**	**98.60**	**77.05**

and 2 indoors). The training set contains 395 persons, with 22,258 RGB images and 11,909 IR images. The test set contains 96 persons, with 3,803 IR images for query and 301 randomly selected RGB images as the gallery. Following [11], two evaluation modes are conducted: *All-search* and *Indoor-search*. For *Indoor-search* mode, images collected by indoor RGB cameras are exclusively selected to built the gallery set. For *All-search* mode, images are randomly selected from all RGB cameras to form the gallery set.

RegDB [21] contains 412 identities, with 206 for training and 206 for testing. Each identity has 10 RGB and 10 IR images. Two evaluation modes are employed: *Visible-Thermal* and *Thermal-Visible*. The former refers to searching for corresponding IR images with a RGB image and vice versa. The dataset is randomly split into 10 training/testing trials. The evaluation results are given by averaging the performances over the 10 trials.

Evaluation metrics Following the standard evaluation protocol given by [2], [12], [30], Cumulative Matching Characteristics (CMC) curve and mean Average Precision (mAP) are adopted as evaluation metrics. Here, CMC reports the probabilities of the targeted identity occurring at top-r in the ranking list, *i.e.,* "Rank-r" accuracy. mAP measures the overall retrieval performance when multiple matching cases occur in the gallery set.

Implementation details The experiments are deployed on an NVIDIA GeForce 2080Ti GPU with Pytorch. Following most existing works [2], [12], [13], all input images are resized to 288×144. Random cropping, random erasing, and horizontal flipping are adopted for data augmentation. The parameters of Modality-specific module, Modality-shared module, and ReID branch are initialized by ResNet50 [33] pre-trained on ImageNet. Other parameters are initialized by Xavier initialization [49]. We adopt the SGD optimizer with a weight decay of 0.0005 for optimization. The learning rate is initialized as 0.01 and decays by 0.5 at every 20 epoch. The training process iterates for 100 epochs in total.

B. Comparison with state-of-the-arts

We extensively compare our algorithm with the current State-Of-The-Art (SOTA) methods on both SYSU-MM01 [11] and RegDB [21] datasets. The SOTA methods include pioneering ones (Zero-Pad [11] and HCML [30]), GAN-based ones (cmGAN [37], AlignGAN [41], and D^2RL [39]), middle modality based ones (X-Modal [42] and cm-SSFT [44]), feature constraints based ones (eBDTR [31], HSME [38], MSR [40], TSLFN [12], AGW [2], and GLMC [47]), dual-level feature alignment based ones (DDAG [13], MACE [43], CICL [46], and LbA [48]).

Evaluations on SYSU-MM01 Table I reports the performance of our model and State-Of-The-Art (SOTA) approaches on the SYSU-MM01 [11] dataset. "$Ours_{ID}$" and "$Ours_{ALL}$" refer to the retrieval performance with features f_{ID} (Equ. 12) and f_{ALL} (Equ. 13), respectively. It can be seen that both "$Ours_{ID}$" and "$Ours_{ALL}$" outperform SOTA approaches on ALL evaluation metrics in both *All-search* and *Indoor-search* evaluation modes. Compared to the second-best method (GLMC [47]), the model performance is improved by approximately 7% and 4% with f_{ALL} in terms of the rank-1 accuracy and the mAP score, respectively.

Evaluations on RegDB The evaluation results on RegDB [21] are shown in Table II. It can be observed that the proposed

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON **REGDB** DATASET. RANK-R ($r = 1, 10, 20$) ACCURACY(%) AND MAP(%) ARE REPORTED. $Ours_{ID}$ AND $Ours_{ALL}$ DENOTE THE FEATURES USED FOR EVALUATION ARE OBTAINED FROM REID BRANCH AND BOTH BRANCHES, RESPECTIVELY.

Method	Visible-Thermal				Thermal-Visible			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Zero-Pad [11]	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
HCML [30]	24.44	47.53	56.78	20.80	21.70	45.02	55.58	22.24
eDBTR [31]	34.62	58.96	68.72	33.46	34.21	58.74	68.64	32.49
HSME [38]	50.85	73.36	81.66	47.00	50.15	72.40	81.07	46.16
D^2RL [39]	43.40	66.10	76.30	44.10	-	-	-	-
MSR [40]	48.43	70.32	79.95	48.67	-	-	-	-
AlignGAN [41]	57.9	-	-	53.6	56.3	-	-	53.4
X-Modal [42]	62.21	83.13	91.72	60.18	-	-	-	-
DDAG [13]	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
AGW [2]	70.05	-	-	66.37	-	-	-	-
MACE [43]	72.37	88.40	93.59	69.09	72.12	88.07	93.07	68.57
cm-SSFT [44]	72.3	-	-	72.9	71.0	-	-	71.7
NFS [45]	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
CICL [46]	78.8	-	-	69.4	77.9	-	-	69.4
GLMC [47]	91.84	97.86	98.98	81.42	91.12	97.86	98.69	81.06
LbA [48]	74.17	-	-	67.64	72.43	-	-	65.46
$Ours_{ID}$	92.14	98.16	99.22	87.88	91.36	97.57	98.88	86.70
$Ours_{ALL}$	**93.35**	**98.61**	**99.42**	**88.98**	**92.72**	**98.79**	**99.36**	**87.83**

model obtains surprisingly good results in both *"Visible-Thermal"* and *"Thermal-Visible"* modes. Specifically, the performance of $Ours_{ID}$ exceeds GLMC [47] by 6.46% and 5.64% in terms of the mAP score in two evaluation modes, respectively. The improvements achieve 7.56% and 6.77% when $Ours_{ALL}$ is employed for identity retrieval. The comparison results also prove that, with the proposed Hierarchical Feature Constraint (HFC), each feature stripe is embedded with modality-shared and ID-related information. Such advantage leads to a satisfactory identity accuracy even when features with less dimensions are used for identity retrieval.

C. Comparison with pose-guided single-modality person ReID

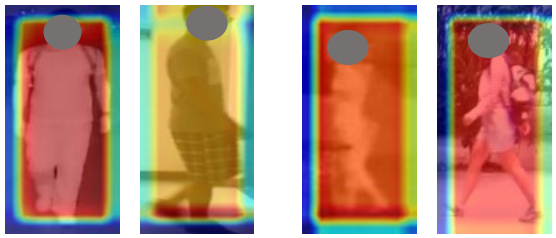
As aforementioned, employing pose information in the VI-ReID task is extremely challenging due to the massive gap between visible images (source domain) and infrared images (target domain). Although there are some pose-guided single-modality person ReID works, how to enable pose information to facilitate the VI-ReID task has been unrevealed. In the section, several experiments are conducted on the SYSU-MM01 [11] dataset to explore the potential of pose estimation in the VI-ReID task.

Firstly, the VI-ReID task is considered as the single-modality person ReID task with two different types of images. For each modality, a modality-specific pose-assisted feature extractor is trained. During the inference, features of query set (IR images) and gallery set (RGB images) are derived by the corresponding feature extractor to conduct the cross-modality person retrieval. Taking a state-of-the-art (SOTA) pose-guided single-modality person ReID work, *i.e.,* PGFA [19], as an example, where human landmarks obtained by an off-the-shelf pose estimator are rigidly used to generate attention maps to highlight body regions, the performance is reported in the first row of Table III. As can be seen, although the same pose estimator is used for two modalities, retrieving identities across modalities with modality-specific pose-assisted features can only achieve 11.45% in terms of the mAP score. The poor performance demonstrates that, the off-the-shelf pose estimators cannot handle the huge gap between the data distribution of two modalities. In this case, the pose information cannot serve as effective modality-shared cues in the VI-ReID task.

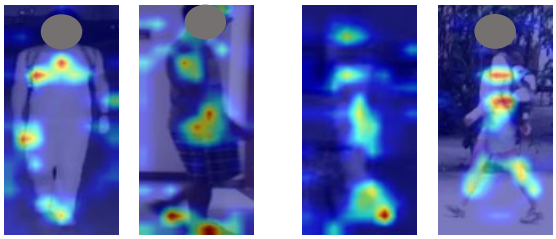
To make the best of the pose information in the VI-ReID task, we applied the two-stream network proposed by a single-modality person ReID approach, *i.e.,* PABR [18], where the ReID branch and the pose branch extract appearance and pose features, respectively. Then two kinds of features are fused for the identity retrieval. To adapt PABR to the VI-ReID task, the cross-modality hard triplet loss [50] is used to replace the single-modality one, where the hardest cross-modality triplets are also considered. The performance of PABR in the VI-ReID task is shown in the second row of Table III. Compared to PGFA, PABR achieves a higher Rank-1 accuracy (40.73%) and mAP score (42.19%). In addition to numerical results,

TABLE III
COMPARISON WITH SOTA POSE-GUIDED SINGLE-MODALITY PERSON REID METHODS ON THE SYSU-MM01 [11] DATASET IN TERMS OF RANK-1 ACCURACY(%) AND MAP(%) AT *All-search* SETTING. THE DIMENSION OF FEATURES USED FOR THE EVALUATION IS SET AS 2560 FOR ALL EXPERIMENTS. $Ours_{ID}$* MEANS P IN EQU. (11) IS SET AS 5. IMPL. MEANS EXPERIMENTS ARE IMPLEMENTED WITH THE OFFICIAL SOURCE CODE PROVIDED.

Methods	Rank-1	mAP
PGFA(Impl.) [19]	10.04	11.45
PABR(Impl.) [18]	40.73	42.19
$Ours_{ID}$*	62.82	61.90



(a) PABR



(b) $Ours_{ID}$*

Fig. 4. Visualization of attentive feature maps output by the pose branch of PABR [18] and the proposed method on the SYSU-MM01 dataset. For each person (group), the RGB image is shown on the left and the IR image is on the right.

we also visualize the attentive feature maps output by the pose branch of PABR [18] by means of Grad-CAM [51] in Fig. 4(a), to have a better understanding of where pose features are extracted. As can be seen, due to the lack of adequate guidance on the pose branch during training, the extracted pose features only locate the whole body coarsely.

To solve the problem, as shown in Fig. 2, the body key-points generated by a pre-trained pose estimator serve only as the guidance of Pose Estimation branch in our method. By applying the *pose estimation loss* during training, more fine-grained pose features can be obtained. The performance of the proposed method is shown in the third row of Table III. It can be seen that, with above improvements, our method outperforms PGFA and PABR by a large margin in terms of the Rank-1 accuracy (62.92%) and the mAP score (61.90%). Additionally, the visualization of the attentive feature maps output by our Pose Estimation branch is shown Fig. 4(b). In comparison to PABR, our method focuses on more body details, such as shoulders and feet, which can serve as distinctive cues for the VI-ReID task.

D. Ablation studies

An ablation analysis is conducted on SYSU-MM01 [11] and RegDB [21] to verify the effectiveness of the proposed components in our model. Firstly, a "Baseline" model is trained under the supervision of identity loss L_{id} and HC-tri loss L_{hctri}, which only consists of Modality-specific module, Modality-shared module, and ReID branch. The evaluation results are shown in the first row in Table IV. Based on "Baseline", Pose Estimation branch (PEB), pose estimation loss (L_{pose}), and Hierarchical Feature Constraint (HFC) are gradually applied, which results are illustrated in the following rows in Table IV. Note that the reported results are in the

TABLE IV
ABLATION STUDIES ON SYSU-MM01 [11] AND REGDB [21] DATASETS. "PEB", "HFC" REFER TO POSE ESTIMATION BRANCH, AND HIERARCHICAL FEATURE CONSTRAINT, RESPECTIVELY.

	Components			SYSU-MM01		RegDB	
	PEB	L_{pose}	HFC	Rank-1	mAP	Rank-1	mAP
Baseline				57.03	56.21	85.76	77.21
$Ours_{ID}$	✓			60.19	58.96	87.39	81.75
	✓	✓		63.90	62.79	89.25	84.16
	✓	✓	✓	**65.82**	**64.52**	**92.14**	**87.88**
$Ours_{ALL}$	✓			65.05	62.78	89.53	84.12
	✓	✓		68.84	65.94	91.96	87.08
	✓	✓	✓	**71.21**	**67.15**	**93.35**	**88.98**

All-search mode for the SYSU-MM01 dataset while in the *"Visible-Thermal"* mode for the RegDB dataset.

Effectiveness of Pose Estimation branch (PEB) As can be seen from the 2*nd* row in Table IV, by integrating the PEB branch that is pre-trained on a pose estimation dataset, the framework yields an increase of approximately 2% and 4% in terms of the mAP score on the SYSU-MM01 dataset and the RegDB dataset, respectively, when f_{ID} is applied for identity retrieval. Likewise, the improvements exceed 6.5% when employing f_{ALL}, which clearly demonstrates the effectiveness of our PEB branch.

Apart from L_{id} and L_{hctri}, L_{pose} is further employed on such a structure to ensure the body skeleton points are precisely estimated. The results for f_{ID} and f_{ALL} are shown in the 3*rd* and 6*th* rows, respectively. Specifically, an increase of 3.83% and 2.41% can be found in terms of the mAP score on two benchmark datasets for f_{ID}. Similar increase, *i.e.,* 3.16% and 2.96%, can be obtained if using f_{ALL}.

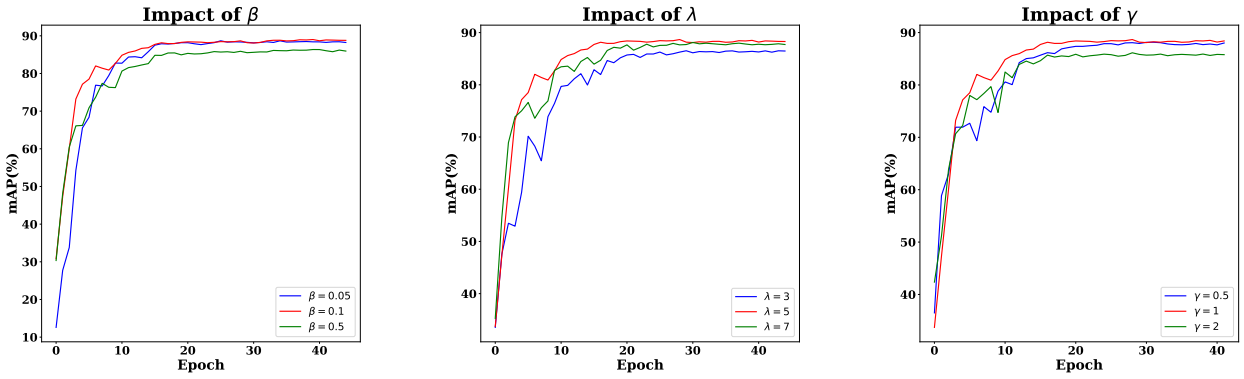
Effectiveness of Hierarchical Feature Constraint (HFC) It can be observed from the 4*th* row that HFC yields a satisfactory improvement for f_{ID} in terms of the mAP score, which are 1.73% and 3.72% on the SYSU-MM01 dataset and the RegDB dataset, respectively. A similar improvement can also be found for f_{ALL} (the 7*th* row), which are around 2% for both datasets. The boost of performance proves that the proposed HFC encourages the information flow between global and local features by introducing an extra supervision on part-level features.

E. Impact of Loss coefficient

In this section, we aim to investigate an optimal combination of weight factors to balance multiple loss terms. According to the paper, the overall objective L for training is defined as,

$$L = L_{id} + \beta L_{hctri} + \lambda L_{pose} + \gamma L_{KD}. \tag{22}$$

β, λ, and γ adjust the contribution of Hetero-center triplet loss, Pose estimation loss, and KD loss, respectively. For a fair comparison, we only change one factor at a time while keeping the rest fixed. Subsequently, models are trained by objectives with different combinations of weight factors. The corresponding mAP scores on the test set are depicted in Fig. 5.



(a) (b) (c)

Fig. 5. Parameter analysis on the RegDB dataset. **Best viewed in color.**

It can be observed that no matter which combination is employed, the proposed framework could achieve a stable performance within 50 epoch. Additionally, according to the experimental results, the optimal combination $\{\beta, \lambda, \gamma\} = \{0.1, 5, 1\}$ is chosen during the training.

F. Visualization

Apart from quantitative results, we also visualize the gradient feature maps output by ReID branch with Grad-CAM [51], in order to examine where the features are extracted.

Visualization results of "Baseline" and the proposed framework ("Ours") are illustrated in Fig. 6(a) and Fig. 6(b), respectively. As can be seen, instead of rigidly extracting features from several vertical regions, the proposed method extracts features from body skeleton joints, which are not only ID-related but highly immune to viewpoint changes and modality changes. The visualization results not only intuitively reveal the reason why our method performs better, but also show the potential of pose estimation task in the field of VI-ReID.

V. Conclusion

We have proposed a novel two-stream VI-ReID framework, where modality-shared and ID-related features for identity retrieval are extracted by means of learning an auxiliary task (pose estimation) and the main task (person ReID) simultaneously. By imposing pose estimation and ReID constraints on the Pose Estimation branch at the same time, both modality-shared and ID-related are fully embedded on each feature stripe. Apart from learning discriminative features at the local level, we also propose a Hierarchical Feature Constraint to bond the learning of global features with local ones by employing the knowledge distillation strategy to ensure the discriminability consistency. The proposed framework achieves a new state-of-the-art on VI-ReID benchmarks in terms of the rank-1 accuracy and the mAP score.

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*, vol. 30, no. 4, pp. 1092–1108, 2019.

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

[3] J. Si, H. Zhang, C.-G. Li, and J. Guo, "Spatial pyramid-based statistical features for person re-identification: A comprehensive evaluation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 7, pp. 1140–1154, 2017.

[4] A. Zheng, X. Zhang, B. Jiang, B. Luo, and C. Li, "A subspace learning approach to multishot person reidentification," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 1, pp. 149–158, 2018.

[5] L. Pang, Z. Cao, J. Yu, P. Guan, X. Rong, and H. Chai, "A visual leader-following approach with a t-d-r framework for quadruped robots," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 4, pp. 2342–2354, 2021.

[6] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and h. Li, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[7] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

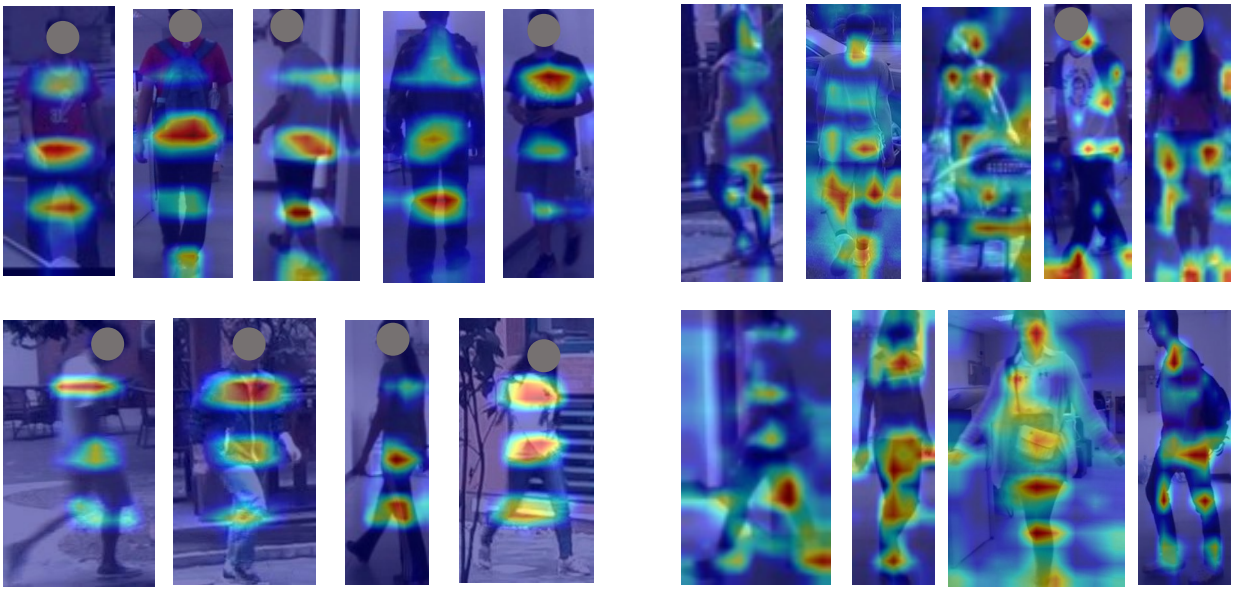
[8] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, "Beyond human parts: Dual part-aligned representations for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[9] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] Y. Lv, Y. Gu, and L. Xinggao, "The dilemma of trihard loss and an element-weighted trihard loss for person re-identification," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 17 391–17 402.

[11] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5380–5389.

[12] Y. Zhu, Z. Yang, L. Wang, S. Zhao, X. Hu, and D. Tao, "Hetero-



(a) Baseline (b) Ours

Fig. 6. Visualization of gradient feature maps output by ReID branch on the SYSU-MM01 dataset.

center loss for cross-modality person re-identification," *arXiv preprint arXiv:1910.09830*, 2019.

[13] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," pp. 229–247, 2020.

[14] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1077–1085.

[15] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3960–3969.

[16] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4500–4509, 2019.

[17] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3219–3228.

[18] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 402–419.

[19] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 542–551.

[20] 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 *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 480–496.

[21] 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.

[22] H. Ling, Z. Wang, P. Li, Y. Shi, J. Chen, and F. Zou, "Improving person re-identification by multi-task learning," *Neurocomputing*, vol. 347, pp. 109–118, 2019.

[23] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7134–7143.

[24] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1062–1071.

[25] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[26] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 346–363.

[27] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," *arXiv preprint arXiv:1810.02936*, 2018.

[28] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4099–4108.

[29] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 650–667.

[30] M. Ye, X. Lan, J. Li, and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[31] 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*, vol. 15, pp. 407–419, 2019.

[32] H. Liu, X. Tan, and X. Zhou, "Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification," *IEEE Transactions on Multimedia*, 2020.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[34] D. Osokin, "Global context for convolutional pose machines," *arXiv preprint arXiv:1906.04104*, 2019.

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

[36] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 871–885, 2018.

[37] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *IJCAI*, vol. 1, 2018, p. 2.

[38] 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*, vol. 33, no. 01, 2019, pp. 8385–8392.

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 618–626.

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

[41] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3623–3632.

[42] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4610–4617.

[43] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 9387–9399, 2020.

[44] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu, "Cross-modality person re-identification with shared-specific feature transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 379–13 389.

[45] Y. Chen, L. Wan, Z. Li, Q. Jing, and Z. Sun, "Neural feature search for rgb-infrared person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 587–597.

[46] Z. Zhao, B. Liu, Q. Chu, Y. Lu, and N. Yu, "Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3520–3528.

[47] L. Zhang, G. Du, F. Liu, H. Tu, and X. Shu, "Global-local multiple granularity learning for cross-modality visible-infrared person reidentification," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[48] H. Park, S. Lee, J. Lee, and B. Ham, "Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 046–12 055.

[49] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[50] Y.-B. Zhao, J.-W. Lin, Q. Xuan, and X. Xi, "Hpiln: a feature learning framework for cross-modality person re-identification," *IET Image Processing*, vol. 13, no. 14, pp. 2897–2904, 2019.

[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.