Article:

Huang, N., Xing, B., Zhang, Q. et al. (2 more authors) (2024) Co-segmentation assisted cross-modality person re-identification. Information Fusion, 104. 102194. ISSN 1566-2535

https://doi.org/10.1016/j.inffus.2023.102194

Co-segmentation Assisted Cross-modality Person Re-identification

Nianchang Huang[a,b], Baichao Xing[a,b], Qiang Zhang[a,b,*], Jungong Han[c], Jin Huang[a]

[a]*Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an, 710071, Shaanxi, China*
[b]*Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an, 710071, Shaanxi, China*
[c]*Department of Computer Science, University of Sheffield, U.K*

Abstract

We present a deep learning-based method for Visible-Infrared person Re-Identification (VI-ReID). The major contribution lies in the incorporation of co-segmentation into a multi-task learning framework for VI-ReID, where the co-segmentation concept aids in making the distributions of RGB images and IR images the same for the same identity but diverse for different identities. Accordingly, a novel multi-task learning based model, *i.e.,* co-segmentation assisted VI-ReID (CSVI), is proposed in this paper. Specifically, the co-segmentation network first takes as the inputs the modality-shared features extracted from a set of RGB and IR images by using the VI-ReID model. Then, it exploits their semantic similarities for predicting the person masks of the common identities within the input RGB and IR images by using a cross-modality center based weight generation module and a segmentation

*Corresponding authors: Qiang Zhang.
Email addresses: `nchuang@stu.xidian.edu.cn` (Nianchang Huang), `baichxing@163.com` (Baichao Xing), `qzhang@xidian.edu.cn` (Qiang Zhang), `jungonghan77@gmail.com` (Jungong Han), `jhuang@mail.xidian.edu.cn` (Jin Huang)

decoder. Doing so enables the VI-ReID model to extract more additional modality-shared shape features for boosting performance. Meanwhile, the co-segmentation network implicitly establishes the interactions among the set of RGB and IR images, thus further bridging the large modality discrepancies. Our model's effectiveness and superiority are verified through experimental comparisons with state-of-the-art algorithms on several benchmark datasets.
Keywords: cross-modality person re-identification, co-segmentation, multi-task learning.

---

1. Introduction

Person Re-IDentification (PReID) is a crucial technology for intelligent video surveillance, aimed at identifying individuals across non-overlapping cameras. Recently, re-identifying persons from visible images (VV-PReID) has shown impressive performance and found applications in real-life scenarios[1, 2]. Although such progress has been made, researchers find that the applications of VV-RReID models in many realistic scenarios have been hindered, since visible cameras cannot capture informative images in case of inadequate illuminations (*e.g.,* at night). Motivated by such challenges, cross-modality PReID, *i.e.,* associating RGB and infrared (IR) pedestrian images for cross-modality image retrieval, has drawn increasing attention[3, 4], since Infrared (IR) cameras excel at capturing more information in challenging illuminations, particularly in low-light conditions [5]. Additionally, many surveillance cameras can automatically switch between RGB and IR modes, making the integration of cross-modality approaches feasible and practical.

Generally speaking, VI-ReID faces two major challenges, *i.e.,* cross-modality

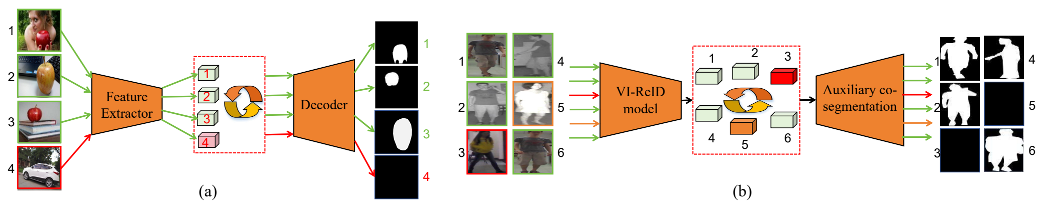


Figure 1: Frameworks of existing co-segmentation models. (a) Architecture of the co-segmentation task. (b) Framework of our proposed model. The images with the same color boxes belong to the same objects (a) or identities (b).

variations and intra-modality variations. The cross-modality variations result from the inherent differences between visible and infrared images. The intra-modality variations are caused by differences in viewpoints, poses, and exposures of individuals. Most existing models try to capture those discriminative features co-existing in the two modalities (*i.e.,* modality-shared features) for simultaneously tackling these challenges. However, the cross-modality variations not only lead to different feature distributions between visible features and infrared features but also cause much person-discriminative information within one modality to be interfered[3]. For instance,VV-PReID heavily relies on color information as a crucial appearance cue, while it is hardly used in VI-ReID. This often makes such modality-shared feature learning difficult.

Employing person masks as auxiliary information has proved to be an effective way for facilitating VI-ReID, since the person masks contain abundant and accurate modality-invariant person shape information. The most common way of exploiting person masks for VI-ReID is directly employing the person masks for selection [6, 7], *i.e.,* selecting persons from backgrounds or selecting features from person regions. However, this only helps VI-ReID models to eliminate the interference information within the backgrounds from their extracted person features, but cannot enable VI-ReID models themself to extract more accurate semantics from the input images, since those person masks do not directly provide any gradients for training (see Section III for more details), thus leading to unsatisfactory results.

Alternatively, Huang *et. al* [8] proposed a multi-task learning based VI-ReID model to facilitate their VI-ReID network extracting more modality-shared person shape information for VI-ReID by exploring the relations between person segmentation and VI-ReID. Specifically, in [8], two sub-networks are employed on top of a shared feature extractor for person segmentation and VI-ReID, respectively. By doing so, the person segmentation sub-network can facilitate the shared feature extractor directly extracting abundant person-related semantics for VI-ReID by predicting those person masks. Meanwhile, those person semantics extracted by the person segmentation sub-network can also be introduced into the VI-ReID sub-network for further boosting performance, thus achieving large performance improvements. *However, this model mainly focuses on exploring the relations between different tasks but ignores the relations between the features across modalities, thus also easily leading to sub-optimal results. Besides, it also introduces some extra computational costs, since an extra segmentation sub-network is required.*

In this paper, we gain inspiration from the task of co-segmentation and eventually utilize it to address the above issues. Specifically, co-segmentation aims to detect the common objects or regions in a set of relevant images, *e.g.,* the apples in Fig. 1(a). One of the main ideas of such a task is to utilize semantic similarity for segmenting objects with the same semantic class but with different appearances and backgrounds. In deep learning based models, the semantic similarity usually means that the distributions of high-level features, which are also called semantic features, are the same for those images with the same classes, but different for those images with different classes, *e.g.,* the semantic features' distributions of apples *vs* those of cars in Fig. 1 [9, 10]. Accordingly, those deep learning based models can achieve co-segmention by interacting those semantic features from different input images. For instance, a correlation layer can help to segment the objects with the same class across two input images, which can be implemented by either computing the correlations of the semantic features [11] or employing the shared weights to select a set of features from a set of input images for segmenting their co-existing objects [12].

Moreover, as depicted in Fig. 1(b), the concept of semantic similarity appears reasonable for VI-ReID. It typically aims to enable the distributions of RGB images and IR images to be the same for the same identity, and vice versa. Therefore, in this paper, we use the co-segmentation to facilitate the VI-ReID by transferring the VI-ReID task to a task that detects the same identities from a set of RGB and IR images. If incorporating the co-segmentation task into the framework of VI-ReID models, the co-segmention network will exploit the semantic similarity across a set of RGB and IR images with the same identities for predicting the masks of this identity, which will enhance the VI-ReID network to extract more additional discriminative modality-shared person shape information for VI-ReID. Moreover, during the co-segmentation, the features from the set of RGB and IR images will also be implicitly interacted with each other. This will further help the VI-ReID network to reduce the large modality discrepancies. Accordingly, those issues in VI-ReID mentioned above will be well addressed. To this end, a novel multi-task learning based VI-ReID model, *i.e.,* co-segmentation assisted VI-ReID (CSVI), is presented in this paper.

Concretely, during the training stage, a VI-ReID network will be first utilized to extract modality-shared RGB and IR features from a given set of input RGB and IR images. Then, an auxiliary co-segmentation model will be designed and cascaded after the VI-ReID network to perform co-segmentation on those input images for assisting the VI-ReID network. Especially, for those input images, the co-segmentation model will interact their cross-modality features and further segment their common identities by exploring their semantic similarity via a Cross-modality Center based Weight Generation (CCWG) module and a segmentation decoder. While, in the testing stage, the auxiliary co-segmentation model will be removed and only the VI-ReID model is employed for VI-ReID, thus without introducing any more parameters and computational costs.

Furthermore, we will theoretically prove that, compared with the ways of taking person masks as selection maps, our proposed model can learn to extract more person shape information from the input images with the aid of those person masks. Meanwhile, compared to the multi-task learning based models that explore the relations between person segmentation and VI-ReID, our proposed model will not only capture more discriminative person-related features by exploring relations between segmentation and VI-ReID, but also effectively reduce the large cross-modality variations by establishing the interactions between modality-shared RGB and IR features via the co-segmentation model, thus obtaining better VI-ReID results.

To summarize, the main contributions of this paper are as follows:

(1) This paper takes the initiative to use the co-segmentation to assist the VI-ReID in a multi-task learning framework. Benefiting from their common idea, *i.e.,* semantic similarity, our co-segmentation model significantly enhances our VI-PeID model's abilities in the extraction of discriminative person-related features and the reduction of cross-modality variation.

(2) An auxiliary co-segmentation model is designed to segment the same identity from a set of input images with different modalities by establishing the interactions across those modality-shared features with different modalities. This will help the VI-ReID model to extract more accurate modality-shared features from the input images, thus significantly boosting the performance of VI-ReID.

(3) The theoretical comparisons between our proposed model and existing models are provided, which further verify our proposed model's effectiveness in theory.

In the following sections, we will discuss the relevant previous works on ReID and VI-ReID in Section 2. Subsequently, we will present the details of our proposed method in Section 3. Section 4 will showcase the experimental results used to validate our approach. Finally, a concise conclusion will be provided in Section 5.

2. Related Work

2.1. VV-PReID

VV-PReID has been extensively explored in the literature. Conventional models mainly focus on designing discriminative hand-crafted descriptors, such as colors, textures and some regular patterns [13]. Recently, CNN-based VV-PReID models have pushed performance to a new level. Specifically, some of these models primarily concentrate on representation learning, with the objective of capturing some person-related features to distinguish different individuals. For example, Jia *et al.* [14] proposed a transformer framework, termed by DRL-Net. This framework employs a novel alignment process, enabling the implicit disentanglement of person representations in a supervision-free manner. Consequently, this model can effectively extract those person-related information for occluded VV-PReID, thus achieving state-of-the-art performance.

Alternatively, others mainly dedicate themselves to metric learning, *i.e.,* aiming at learning an embedding representation that increases the feature similarity among the same identities and reduces the feature similarity among different identities in the embedding space. For example, Yang *et al.* [15] developed a structural metric learning objective for VV-PReID. In their approach, each positive pair was allowed to compete against all negative pairs in a minibatch. Moreover, they dynamically assigned a hardness-aware weight to each positive pair, adjusting their contributions based on the level of difficulty, leading to improved performance.

2.2. VI-ReID

Recently, a considerable number of VI-ReID models have been extensively studied, and providing a comprehensive summary of all these models is beyond the scope of this paper. For interested readers, we recommend referring to [3] for recent surveys on this subject.

Existing VI-ReID models can be broadly classified into two groups: modality-shared feature learning and modality-specific feature compensation.Similar to other cross-modality matching task [16], the former aims to extract discriminative features that are common across multiple modalities. For instance, Wei *et al.* [17] introduced the Adaptive Body Partition model, which employs separate sub-networks to extract single-modality information from RGB and IR images. A shared sub-network is then utilized to detect and segment body parts, enabling the extraction of more discriminative local information. Chen *et al.* [18] presented a structure-aware positional transformer, which leverages structural and positional information to explore semantically aware, shareable modality features. They introduce an ASR module to explicitly explore structure-related features from each modality, thereby reducing complex background noise. Furthermore, a TPI module is designed to model contextual and positional relations. Similar to Huang *et al.* [8], Miao [19] proposed a two-stream framework based VI-ReID model which explores the pose estimation as the auxiliary learning task to help the ReID task in VI-ReID. To facilitate transferring the pose information from pose estimation to VI-ReID stream, they proposed a Hierarchical Feature Constraint (HFC) to ensure the discriminability consistency of global features and local ones via the knowledge distillation strategy.

In contrast, modality-specific feature compensation-based models typically start by generating the missing modality information from the available modality, thsu addressing cross-modality variations. Subsequently, they utilize both the original and generated information to handle intra-modality variations. Liu *et al.*[20] proposed a new two-stage GAN based model, which first optimizes the image generator's structures and objective functions in the first stage. Then, it improves the ReID network by employing the feature-level fusion rather than the image-level fusion for their original and generated information in the second stage. Differently, Lai *et al.*[21] introduced a feature-level compensation approach for VI-ReID. They first disentangled the single-modality features into modality-specific features and modality-shared features. Subsequently, they generated the missing modality-specific features from the disentangled modality-shared features. Finally, they fused the original modality-specific features, the generated modality-specific features, and the disentangled modality-shared features to perform VI-ReID. This feature-level compensation strategy allowed for better handling of cross-modality variations and intra-modality variations in their model.

3. Proposed Model

Figure 2 illustrates the structure of our proposed model, which comprises a VI-ReID network and an auxiliary co-segmentation network. It is worth noting that the auxiliary co-segmentation network is only used during the training stage and will be excluded during the testing stage, without introducing any extra parameters. In the subsequent sections, we will delve into the specifics of each component in detail.

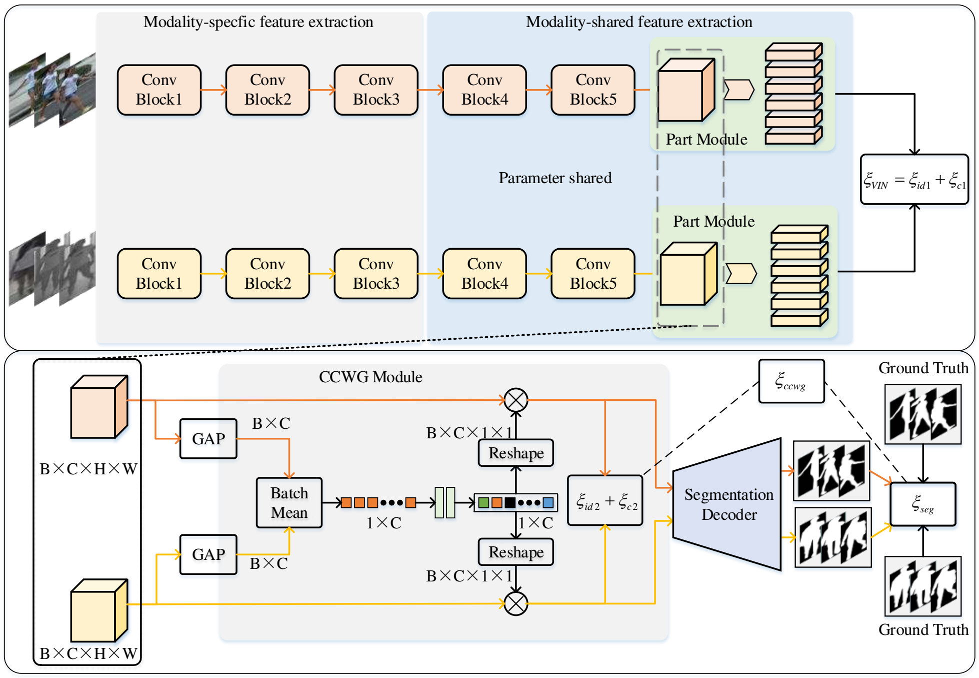


Figure 2: Illustration of the proposed model. In our proposed model, VI-ReID network first employs two independent subnetworks to extract those single-modality RGB and IR features from the input RGB and IR images, respectively, and further uses two parameter-shared subnetworks to extract those modality-shared features. Meanwhile, some of those extracted modality-shared features are fed into the auxiliary co-segmentation network to assist VI-ReID network. Here, a CCWG module is first employed to generate some feature weights for a set of RGB and IR images with the same identities and a segmentation decoder is further employed to segment their common identities.

3.1. VI-ReID Network

The VI-ReID network is a modification of ResNet-50. Initially, two separate sub-networks are employed to extract single-modality features from the input RGB and IR images, respectively. Subsequently, the extracted RGB features and IR features are projected into a shared feature space, where their modality-shared features are extracted using two sub-networks with shared parameters. Finally, a part module is applied to the extracted modality-shared features to obtain the final person part features for VI-ReID. Suppose that there are N identities and each identity has K RGB images and K IR images with the size of $W \times H$, *i.e.*, $X_R = \{x_R^i \in R^{H \times W}\}_i^J$ and $X_I = \{x_I^i \in R^{H \times W}\}_i^J$. Here, $J = NK$.

3.1.1. Modality-specific Feature Extraction

Typically, the low-level features obtained from RGB and IR images, respectively, exhibit significant discrepancies due to their capture in different spectrums. Therefore, as shown in Fig. 2, two shallower sub-networks are employed in our proposed VI-ReID network to extract those single-modality features from the input RGB images and IR images, respectively. Here, the sub-networks have the same structure but extract different modality-specific information using distinct parameters. Specifically, the two shallower sub-networks are constructed by using the same structures as the first three convolutional blocks in ResNet50 [22]. As a result, we obtain the RGB features $\mathbf{F}_R$ and the IR features $\mathbf{F}_I \in R^{J \times C_1 \times \frac{W}{8} \times \frac{H}{8}}$. Mathematically, this process is expressed by

$$\mathbf{F}_m = \mathrm{ConvB}(X_m, \alpha_m), \tag{1}$$

where $\mathrm{ConvB}(*, \alpha_m)$ denotes a sub-network with its corresponding parameters α_m.

3.1.2. Modality-shared Feature Extraction

As shown in Fig. 2, the extracted single-modality RGB and IR features are fed into two sub-networks, both with shared parameters, to obtain their corresponding modality-shared features. Both sub-networks follow the same structures, which are based on the last two convolutional blocks of ResNet-50 [22]. Notably, the strides of the convolutional layers in the last block are set to 1 to preserve more spatial details during the feature extraction process. Accordingly, the modality-shared features $\mathbf{F}_{s,m} \in R^{J \times C_1 \times \frac{W}{16} \times \frac{H}{16}}$ are obtained by

$$\mathbf{F}_{s,m} = \mathrm{ConvB}(\mathbf{F}_m, \beta), \tag{2}$$

where, $m \in \{R, I\}$ denotes RGB or IR modality. $\mathrm{ConvB}(*, \beta)$ denotes the convolutional blocks with the parameters β. C_1 denotes the number of the feature channels.

3.1.3. Part Module

After obtaining the modality-shared features $\mathbf{F}s, m$, a part module is utilized to further extract those discriminative modality-shared person features from different person parts. Specifically, following [23], those modality-shared features $\mathbf{F}_{s,m}$ are initially divided into six strips along their vertical direction. This division helps to focus on individual person parts and enables the extraction of more specific and discriminative information from each part. Then, a global average pooling is performed on each strip, thus obtaining six parts of features $\hat{\mathbf{F}}^p_{s,m} \in R^{J \times C_1}$. Here, $p = 1, 2, 3, ..., 6$ denotes different parts. After that, a fully connected layer is employed for the features of each part to embed them into a metric space, obtaining the final person part features $\mathbf{F}^p_m \in R^{J \times C_2}$. Finally, a fully connected layer based classifier is employed to predict their corresponding person identities, obtaining their corresponding scores cls_m^p, p=1,2,...,6. Mathematically, this process is expressed by

$$\hat{\mathbf{F}}_{s,m}^1, ..., \hat{\mathbf{F}}_{s,m}^6 = \text{GAP}(\text{Sep}(\mathbf{F}_{s,m})), \mathbf{F}_m^p = \text{FC}(\hat{\mathbf{F}}_{s,m}^p, \gamma_p), cls_m^p = \text{FC}(\mathbf{F}_m^p, \gamma_s), \quad (3)$$

where Sep($*$) denotes the separation operation. GAP($*$) denotes the global average pooling. FC($*, \gamma_p$) and FC($*, \gamma_s$) denote the fully connected layers with their corresponding parameters γ_p and γ_s, respectively. In this paper, $C_1 = 2048$ and $C_2 = 512$.

3.1.4. Loss Function

two loss functions, including an identity loss ξ_{id1} and a center loss ξ_{c1}, are employed to train our proposed VI-ReID network. The identity loss ξ_{id1} is performed on the classification scores cls_m^p to make the model extract those identity-related information, which is expressed by

$$\xi_{id1} = \sum_{p=1}^{6} \xi_{ce}(cls_R^p, cls_g) + \sum_{p=1}^{6} \xi_{ce}(cls_I^p, cls_g), \quad (4)$$

where cls_g are the corresponding ground-truth labels for those input images. ξ_{ce} is the cross-entropy loss, which is expressed by

$$\xi_{ce}(p_t, p_g) = -\frac{1}{L} \sum_{l=1}^{L} p_g^l \log(p_t^l), \quad (5)$$

where L is the class number and is equal to N, *i.e.,* the total identities in the dataset, in this paper. p_g^l are the ground-truth labels for the l-th class and p_t^l are their corresponding predicted values for the probability of being the l-th class.

The center loss ξ_{c1} is performed on the person part features $\mathbf{F}_m^p$, aiming to make the features from the same identity to be compact. It is expressed by

$$\xi_{c1} = \sum_{p=1}^{6} \xi_{hc}(\mathbf{F}_R^p, \mathbf{F}_I^p), \tag{6}$$

where ξ_{hc} is the hetero-center loss [24] and is expressed by

$$\xi_{hc}(\mathbf{F}_R^p, \mathbf{F}_I^p) = \sum_{n=1}^{N} \left\| \mathbf{F}_{CR}^{n,p} - \mathbf{F}_{CI}^{n,p} \right\|_2. \tag{7}$$

Here, $\mathbf{F}_{CR}^{n,p}$ and $\mathbf{F}_{CI}^{n,p}$ are the feature centers of the p-th part features for the n-th identity, respectively, which are computed by

$$\mathbf{F}_{CR}^{n,p} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{F}_R^{n,k,p}, \mathbf{F}_{CI}^{n,p} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{F}_I^{n,k,p}, \tag{8}$$

where $\mathbf{F}_{CR}^{n,k,p}$ and $\mathbf{F}_{CI}^{n,k,p}$ denote the features of the p-th part from the k-th RGB and IR images of the n-th identity. Therefore, the total loss for training the VI-ReID network is expressed by

$$\xi_{VIN} = \xi_{id1} + \xi_{c1}. \tag{9}$$

3.2. Co-segmentation Network

The modality-shared features $\mathbf{F}_{s,m} \in R^{J \times C_1 \times \frac{W}{16} \times \frac{H}{16}}$ obtained by using Eq. (2) will be further fed into an auxiliary co-segmentation network to predict their person masks. It should be noted that, different from person segmentation, which may directly predict the person mask from each modality image to facilitate the extraction of person semantics for VI-ReID, the co-segmentation network aims to detect the person masks of one certain identity from a set of images across different modalities.

For that, as shown in Fig. 2, the co-segmentation network first employs a cross-modality center based weight generation (CCWG) module, which will explore the relations across the modality-shared RGB and IR features extracted from the K RGB images and K IR images of one identity, and accordingly generate a set of shared co-segmentation weights for the K RGB images and K IR images. Then, by virtue of the generated weights, the co-segmentation network will select those unique features about this identity from such modality-shared features for each RGB or IR image. Subsequently, for each RGB or IR image of this identity, the co-segmentation network will further employ a parameter-shared decoder to predict a person mask of this identity by using those selected modality-shared features from this image. Through this approach, our proposed model can effectively explore the relationships between person segmentation and VI-ReID, resulting in the extraction of more discriminative person-related features. Simultaneously, the model also considers the relationships between the features of the two modalities, which helps in reducing cross-modality variations. Consequenceally, our model achieves enhanced performance in handling both intra-modality and cross-modality challenges for VI-ReID tasks. In the following content, we will take the n-th identity as an example for the introduction.

3.2.1. Cross-modality Center based Weight Generation (CCWG) Module

Theoretically, if a set of images contains the same semantics in the co-segmentation task, the features extracted from the set of images should have the same subset of highly activated features, thus enabling the co-segmentation network to segment the same semantics from the set of images. Considering that, the CCWG module is designed to generate the co-segmentation weights for selecting the subset of features according to their semantics.

Specifically, for the n-th identity, the VI-ReID network can extract its modality-shared RGB features $\mathbf{F}^n_{s,R} \in R^{K\times C_1\times\frac{W}{16}\times\frac{H}{16}}$ from K RGB images and IR features $\mathbf{F}^n_{s,I} \in R^{K\times C_1\times\frac{W}{16}\times\frac{H}{16}}$ from K IR images, respectively. Considering that the feature centers of different identities should be separated from each other in the VI-ReID task, the feature center of one identity can well represent the unique characteristics of this identity. Therefore, the feature centers of different identities can be used to segment the same identities across the images of different modalities. For that, the CCWG module first computes the cross-modality feature center $\mathbf{F}^n_{Center} \in R^{C_1\times 1\times 1}$ for the n-th identity by

$$\mathbf{F}^n_{Center} = \frac{1}{2K}\sum_{k=1}^{K}(\mathbf{F}^{n,k}_{g,R} + \mathbf{F}^{n,k}_{g,I}), \tag{10}$$

where the features $\mathbf{F}^{n,k}_{g,m} \in R^{C_1\times 1\times 1}$ denotes the global features of the modality-shared features $\mathbf{F}^{n,k}_{s,m}$. Here, $m \in R, I$ denotes the RGB or IR modality. They are computed by using a global average pooling layer, *i.e.,*

$$\mathbf{F}^{n,k}_{g,m} = \mathrm{GAP}(\mathbf{F}^{n,k}_{s,m}). \tag{11}$$

Accordingly, a fully connected layer is performed on the cross-modality feature center to generate their corresponding co-segmentation weights $\mathbf{w}^n \in R^{C_1\times 1\times 1}$ for selecting those unique features about the n-th identity, *i.e,*

$$\mathbf{w}^n = \mathrm{FC}(\mathbf{F}^n_{Center}, \theta), \tag{12}$$

where $\mathrm{FC}(*, \theta)$ denotes a fully connected layer with its corresponding parameters θ. Here, the weights w^n are shared for all the RGB or IR images of the n-th identity, since the single-modality features $\mathbf{F}^{n,k}_{sel,m}$ for different images of the n-th identity should have the same distributions. Accordingly, the subset of unique features $\mathbf{F}_{sel,m}^{n,k}$ for the k-th image related to the n-th identity are selected by

$$\mathbf{F}_{sel,R}^{n,k} = \mathbf{w}^n \otimes \mathbf{F}_{s,R}^{n,k}, \mathbf{F}_{sel,I}^{n,k} = \mathbf{w}^n \otimes \mathbf{F}_{s,I}^{n,k}, \tag{13}$$

where $\otimes$ denotes the channel-wise multiplication. The features $\mathbf{F}_{sel,R}^{n,k}$ and $\mathbf{F}_{sel,I}^{n,k} \in R^{K\times C_1 \times \frac{W}{16} \times \frac{H}{16}}$ are the selected features for the corresponding RGB and IR images, respectively.

3.2.2. Segmentation Decoder

As shown in Fig. 2, the selected features $\mathbf{F}_{sel,R}^{n}$ and $\mathbf{F}_{sel,I}^{n}$ will be fed into a segmentation decoder to predict their person masks. Specifically, the selected features are fed into two stacked deconvolutional blocks to predict the final person masks $\mathbf{M}_R^n$ and $\mathbf{M}_I^n \in R^{K\times N\times \frac{W}{4}\times\frac{H}{4}}$, respectively, which are expressed by

$$\mathbf{M}_m^n = \text{DConv}(\text{DConv}(\mathbf{F}_{sel,m}^n, \lambda_1), \lambda_2), \tag{14}$$

where $\text{DConv}(*, \lambda_1)$ and $\text{DConv}(*, \lambda_2)$ denote two deconvolutional blocks with their parameters λ_1 and λ_1, respectively. Furthermore, for each deconvolutional block, a deconvolutional layer is first employed to up-sample the selected features and then two standard convolutional layers are employed to capture more features. Therefore, the sizes of $\mathbf{M}_R^n$ and $\mathbf{M}_I^n$ become $\frac{W}{4} \times \frac{H}{4}$. Similar operations are also performed on other identities. Accordingly, we can obtain the selected features $\mathbf{F}_{sel,R}$ and $\mathbf{F}_{sel,I} \in R^{J\times C_1\times\frac{W}{16}\times\frac{H}{16}}$ as well as their predicted masks $\mathbf{M}_R$ and $\mathbf{M}_I \in R^{J\times N\times\frac{W}{4}\times\frac{H}{4}}$.

During the training stage, the co-segmentation network only utilizes those modality-shared features extracted from the VI-ReID network to predict the masks of one certain identity from a set of images with different modalities. This will enhance the VI-ReID network's ability of extracting more discriminative modality-shared features, especially for those modality-invariant features related to person shapes. Meanwhile, the single-modality features from a set of RGB and IR images with the same identity will implicitly interact with each other in the CCWG module via the gradient backpropagation. This will also help the VI-ReID network to reduce the large modality discrepancies, thus further boosting performance.

Moreover, to keep the selected features corresponding to different identities rather than some general person features, as shown in Fig.2, an extra parameter-shared part module is performed on the selected features $\mathbf{F}_{sel,R}$ and $\mathbf{F}_{sel,I}$ for predicting their identities. Accordingly, the person part features $\mathbf{F}^{p}_{sel,m} \in R^{J \times C_2}$ and their scores $cls^{p}_{sel,m}$ are also obtained. Here, $p = 1, 2, ..., 6$ denotes different person parts, and $m \in \{R, I\}$ denotes the RGB or IR modality.

3.2.3. Loss Functions

Three loss functions, including an identity loss ξ_{id2}, a center loss ξ_{c2} and a segmentation loss ξ_{seg}, are employed for training the proposed co-segmentation network.

Similar to Eq.(4), the identity loss ξ_{id2} is performed on the classification scores $cls^{p}_{sel,m}$ to ensure that the co-segmentation network can extract those identity-related information, which is expressed by

$$\xi_{id2} = \sum_{p=1}^{6} \xi_{ce}(cls^{p}_{sel,R}, cls_g) + \sum_{p=1}^{6} \xi_{ce}(cls^{p}_{sel,I}, cls_g), \tag{15}$$

where cls_g denotes their corresponding ground truth labels. Similar to Eq.(6), the center loss ξ_{c2} is performed on the person part features $\mathbf{F}^p_{sel,m}$, aiming to make the features from the same identity to be compact, *i.e.,*

$$\xi_{c2} = \sum_{p=1}^{6} \xi_{hc}(\mathbf{F}^p_{sel,R}, \mathbf{F}^p_{sel,I}), \tag{16}$$

The segmentation loss ξ_{seg} is used to make the co-segmentation model learn to extract more person shape information, *i.e.,*

$$\xi_{seg} = \xi_{ce}(\mathbf{M}_R, \mathbf{M}_{Rg}) + \xi_{ce}(\mathbf{M}_I, \mathbf{M}_{Ig}), \tag{17}$$

where $\mathbf{M}_{Rg}$ and $\mathbf{M}_{Ig}$ denote their corresponding ground-truth person masks. It should be noted that all the ground-truth person masks are obtained from the paper [8]. Accordingly, the total loss for training CCWG is

$$\xi_{ccwg} = \xi_{id2} + \xi_{c2} + \xi_{seg}. \tag{18}$$

Furthermore, our proposed model is trained in an end-to-end manner. Accorrdingly, the total loss function is

$$\xi_{total} = \xi_{VIN} + \xi_{ccwg}. \tag{19}$$

3.3. Theoretical analysis

In this section, we will theoretically analyze different ways of using those person maps in the task of VI-ReID. As shown in Fig. 3, there are three ways of using those person maps in the task of VI-ReID. The most widely used way is shown in Fig. 3(a), which simply takes those person maps as the weight maps for selecting those person-related features. While, as shown in Fig. 3(b), some works try to explore the relations between the tasks of VI-ReID and person segmentation via a multi-task learning framework.

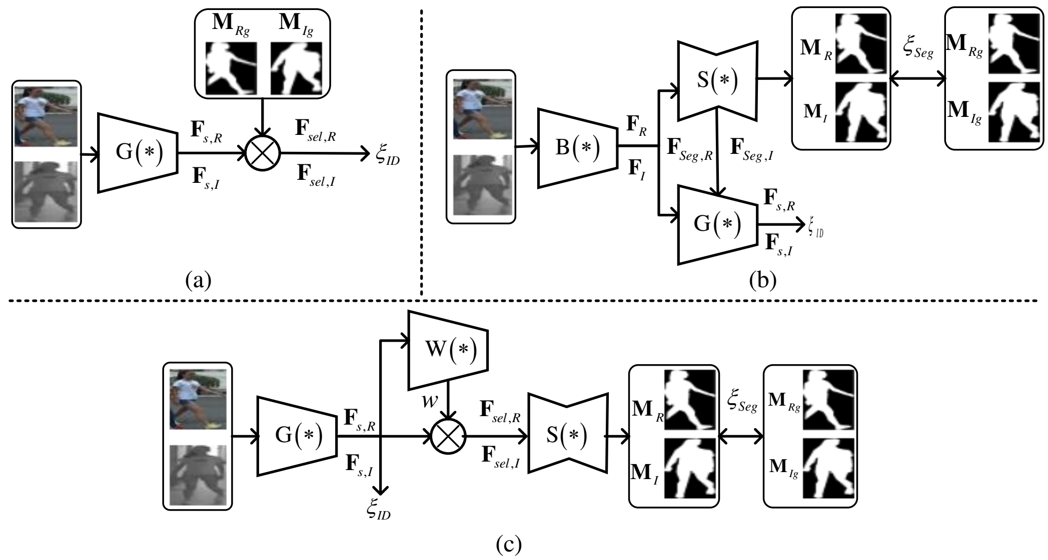


Figure 3: llustration of different ways of using those person maps. (a) Simply using person maps for feature selection, *i.e.,* [6] and [7]. (b) Exploring person maps via existing multi-task learning frameworks, *i.e.,* [8]. (c) Our proposed model.

Differently, as shown in Fig. 3(c), we propose a novel multi-task learning framework, which explores the relations between the tasks of VI-ReID and co-segmentation. We will first simplify their structures and then theoretically analyze the three ways of using person maps in the following contents.

Feature selection: The simplified structures of the feature selection based models are shown in Fig. 3(a). The input RGB/IR images are directly fed into the VI-ReID network $\mathrm{G}(*, \epsilon_G)$ for extracting their corresponding modality-shared features $\mathbf{F}_{s,R}/\mathbf{F}_{s,I}$. Here, ϵ_G denotes the VI-ReID network's parameters. Then, the person masks $\mathbf{M}_{Rg}$ and $\mathbf{M}_{Ig}$ are used for selecting those person-related features. Finally, the selected features are employed for computing the ID loss in the training stage. Accordingly, the gradients from the ID loss to the VI-ReID network in the backpropagation are computed by

$$\begin{aligned}\frac{\partial \xi_{ID}}{\partial \epsilon_G} &= \frac{\partial \xi_{ID}}{\partial \mathbf{F}_{sel,R}}\frac{\partial \mathbf{F}_{sel,R}}{\partial \mathbf{F}_{s,R}}\frac{\partial \mathbf{F}_{s,R}}{\partial \epsilon_G} + \frac{\partial \xi_{ID}}{\partial \mathbf{F}_{sel,I}}\frac{\partial \mathbf{F}_{sel,I}}{\partial \mathbf{F}_{s,I}}\frac{\partial \mathbf{F}_{s,I}}{\partial \epsilon_G} \\ &= \mathbf{M}_{Rg}\frac{\partial \xi_{ID}}{\partial \mathbf{F}_{sel,R}}\frac{\partial \mathbf{F}_{s,R}}{\partial \epsilon_G} + \mathbf{M}_{Ig}\frac{\partial \xi_{ID}}{\partial \mathbf{F}_{sel,I}}\frac{\partial \mathbf{F}_{s,I}}{\partial \epsilon_G}.\end{aligned} \tag{20}$$

Here,

$$\mathbf{F}_{sel,R} = \mathbf{M}_{Rg}\mathbf{F}_{s,R}, \mathbf{F}_{sel,I} = \mathbf{M}_{Ig}\mathbf{F}_{s,I}, \tag{21}$$

where $\mathbf{M}_R$ and $\mathbf{M}_I$ can be seen as the constant values.

It can be seen that the person masks in Eq.(20) are taken as the constant values for filtering out those background information. While, they do not directly provide any gradients for training the VI-ReID model. Accordingly, the VI-ReID network cannot learn to extract more modality-invariant shape information, thus leading to suboptimal results in VI-ReID tasks.

Multi-task learning framework based on segmentation and VI-ReID: In this VI-ReID model, the input RGB/IR images are first fed into a task-shared sub-network B$(*, \epsilon_B)$ for extracting their single-modality features $\mathbf{F}_R$ and $\mathbf{F}_I$. Then, the task-shared features are fed into a sub-network S$(*, \epsilon_S)$ for segmentation and a sub-network G$(*, \epsilon_G)$ for VI-ReID, respectively. Here, ϵ_B, ϵ_S and ϵ_G denote the parameters of their corresponding networks. Besides, the features $\mathbf{F}_{Seg,R}$ and $\mathbf{F}_{Seg,I}$ extracted by the segmentation sub-network are also introduced into the sub-network G$(*, \epsilon_G)$ for boosting the performance. The total loss function of this process is computed by summing the ID loss (ξ_{ID}) and the segmentation loss (ξ_{Seg}). Accordingly, the gradients from the total loss to the VI-ReID network are computed by

$$\frac{\partial \xi_{total}}{\partial \epsilon_G} = \frac{\partial \xi_{ID}}{\partial \epsilon_G} + \frac{\partial \xi_{Seg}}{\partial \epsilon_G} = \frac{\partial \xi_{ID}}{\partial \mathbf{F}_{s,R}}\frac{\partial \mathbf{F}_{s,R}}{\partial \epsilon_G} + \frac{\partial \xi_{ID}}{\partial \mathbf{F}_{s,I}}\frac{\partial \mathbf{F}_{s,I}}{\partial \epsilon_G}. \tag{22}$$

And,

$$\frac{\partial \xi_{total}}{\partial \epsilon_B} = \frac{\partial \xi_{ID}}{\partial \epsilon_B} + \frac{\partial \xi_{Seg}}{\partial \epsilon_B} = \boxed{\frac{\partial \xi_{ID}}{\partial \mathbf{F}_{s,R}} \frac{\partial \mathbf{F}_{s,R}}{\partial \epsilon_G} (\frac{\partial \epsilon_G}{\partial \mathbf{F}_R} + \frac{\partial \epsilon_G}{\partial \mathbf{F}_{Seg,R}} \frac{\partial \mathbf{F}_{Seg,R}}{\partial \epsilon_S} \frac{\partial \epsilon_S}{\partial \mathbf{F}_R}) \frac{\partial \mathbf{F}_R}{\partial \epsilon_B}}$$
$$+ \boxed{\frac{\partial \xi_{ID}}{\partial \mathbf{F}_{s,I}} \frac{\partial \mathbf{F}_{s,I}}{\partial \epsilon_G} (\frac{\partial \epsilon_G}{\partial \mathbf{F}_I} + \frac{\partial \epsilon_G}{\partial \mathbf{F}_{Seg,I}} \frac{\partial \mathbf{F}_{Seg,I}}{\partial \epsilon_S} \frac{\partial \epsilon_S}{\partial \mathbf{F}_I}) \frac{\partial \mathbf{F}_I}{\partial \epsilon_B}} + \boxed{\frac{\partial \xi_{Seg}}{\partial \mathbf{M}_R} \frac{\partial \mathbf{M}_R}{\partial \epsilon_S} \frac{\partial \epsilon_S}{\partial \mathbf{F}_R} \frac{\partial \mathbf{F}_R}{\partial \epsilon_B}}$$
$$+ \boxed{\frac{\partial \xi_{Seg}}{\partial \mathbf{M}_I} \frac{\partial \mathbf{M}_I}{\partial \epsilon_S} \frac{\partial \epsilon_S}{\partial \mathbf{F}_I} \frac{\partial \mathbf{F}_I}{\partial \epsilon_B}}. \tag{23}$$

Similar to the ID loss in Eq. (20), the ID loss in Eq. (22) and Eq. (23) can facilitate the VI-ReID network to extract more person-related and ID-discriminative information for identifying different persons. Differently, the last two items of Eq. (23) (marked by the green boxes) indicate that the person masks can directly provide gradients to train the VI-ReID network, thus enabling the VI-ReID network to learn the ability of extracting more accurate and modality-invariant person semantics from the person masks for VI-ReID. Accordingly, this framework can explore the relations between the tasks of VI-ReID and person segmentation, thus achieving better results. It can be also seen that the last two items of Eq. (23) are independent for each other. This means that, in this framework, the modality-shared RGB features and the modality-shared IR features are not interacted with each other, which cannot well reduce the modality differences, thus leading to sub-optimal results.

Multi-task learning framework based on co-segmentation and VI-ReID (our model): The simplified structure of our proposed model is shown in Fig.3(c). It first employs a VI-ReID sub-network $\mathrm{G}(*, \epsilon_G)$ to extract those modality-shared features from the input images. Then, a weight generation sub-network $\mathrm{W}(*, \epsilon_W)$ is employed to predict the weights $\mathbf{w}$ for selecting a set of unique features of one identity. Here, ϵ_S also denotes the parameters of the weight generation sub-network. Finally, the selected features will be fed into a co-segmentation sub-network $\mathrm{S}(*, \epsilon_S)$ to segment those objects co-existing within the input images. Accordingly, the gradients from the total loss, including the ID loss (ξ_{ID}) and the segmentation loss (ξ_{Seg}), to the VI-ReID network are computed by

$$
\begin{aligned}
\frac{\partial \xi_{total}}{\partial \epsilon_G} &= \frac{\partial \xi_{ID}}{\partial \epsilon_G} + \frac{\partial \xi_{Seg}}{\partial \epsilon_G} = \frac{\partial \xi_{ID}}{\partial \mathbf{F}_{s,R}} \frac{\partial \mathbf{F}_{s,R}}{\partial \epsilon_G} + \frac{\partial \xi_{ID}}{\partial \mathbf{F}_{s,I}} \frac{\partial \mathbf{F}_{s,I}}{\partial \epsilon_G} + \frac{\partial \xi_{Seg}}{\partial \mathbf{M}_R} \frac{\partial \mathbf{M}_R}{\partial \epsilon_S} \frac{\partial \epsilon_S}{\partial \mathbf{F}_{sel,R}} \\
&(\mathbf{w} + \mathbf{F}_{s,R} \frac{\partial \mathbf{w}}{\partial \mathbf{F}_{s,R}}) \frac{\partial \mathbf{F}_R}{\partial \epsilon_G} + \frac{\partial \xi_{Seg}}{\partial \mathbf{M}_I} \frac{\partial \mathbf{M}_I}{\partial \epsilon_S} \frac{\partial \epsilon_S}{\partial \mathbf{F}_{sel,I}} (\mathbf{w} + \mathbf{F}_{s,I} \frac{\partial \mathbf{w}}{\partial \mathbf{F}_{s,I}}) \frac{\partial \mathbf{F}_I}{\partial \epsilon_G} \\
&= \boxed{\frac{\partial \xi_{ID}}{\partial \mathbf{F}_{s,R}} \frac{\partial \mathbf{F}_{s,R}}{\partial \epsilon_G}} + \boxed{\frac{\partial \xi_{ID}}{\partial \mathbf{F}_{s,I}} \frac{\partial \mathbf{F}_{s,I}}{\partial \epsilon_G}} + \boxed{\mathbf{F}_{s,R} \frac{\partial \xi_{Seg}}{\partial \mathbf{M}_R} \frac{\partial \mathbf{M}_R}{\partial \epsilon_S} \frac{\partial \epsilon_S}{\partial \mathbf{F}_{sel,R}} \frac{\partial \mathbf{w}}{\partial \mathbf{F}_{s,R}} \frac{\partial \mathbf{F}_R}{\partial \epsilon_G}} \\
&+ \boxed{\mathbf{F}_{s,I} \frac{\partial \xi_{Seg}}{\partial \mathbf{M}_I} \frac{\partial \mathbf{M}_I}{\partial \epsilon_S} \frac{\partial \epsilon_S}{\partial \mathbf{F}_{sel,I}} \frac{\partial \mathbf{w}}{\partial \mathbf{F}_{s,I}} \frac{\partial \mathbf{F}_I}{\partial \epsilon_G}} + \boxed{\mathbf{w}(\frac{\partial \xi_{Seg}}{\partial \mathbf{M}_R} \frac{\partial \mathbf{M}_R}{\partial \epsilon_S} \frac{\partial \epsilon_S}{\partial \mathbf{F}_{sel,R}} \frac{\partial \mathbf{F}_R}{\partial \epsilon_G}} \\
&+ \boxed{\frac{\partial \xi_{Seg}}{\partial \mathbf{M}_I} \frac{\partial \mathbf{M}_I}{\partial \epsilon_S} \frac{\partial \epsilon_S}{\partial \mathbf{F}_{sel,I}} \frac{\partial \mathbf{F}_I}{\partial \epsilon_G})}.
\end{aligned}
\tag{24}
$$

Similar to that in Eq.(24), the proposed model can also effectively explore the relations between the tasks of VI-ReID and person segmentation via the two items marked by the green boxes in Eq.(24). Accordingly, the VI-ReID network can also learn the ability of extracting those accurate and modality-invariant person semantics from the person masks for VI-ReID. Moreover, as shown in the last item of Eq. (24) (marked by the red box), the modality-shared RGB features and the modality-shared IR features will be interacted with each other with the aid of those generated weights, thus benefiting to reduce their modality discrepancies. Accordingly, the proposed model can effectively explore the relationships between person segmentation and VI-ReID, leading to the extraction of more discriminative person-related features. Moreover, it also considers the relationships between the features of the two modalities, which helps in reducing cross-modality variations. As a result, the model achieves improved performance by addressing both intra-modality and cross-modality challenges for VI-ReID tasks. Besides, the co-segmentation network only appears in the training stage, which does not introduce any more parameters in the testing stage.

Fig. 4 shows the person masks of two identities obtained from our proposed model. The person masks in the second row are predicted by taking a set of RGB and IR images of one of the two identities as the inputs. While, the person masks in the third row are obtained by simultaneously taking as the inputs all of the RGB and IR images of the two identities, where the images of the first identity are far more than those of the second identity. It can be seen that our proposed model can rightly predict the person masks across a set of RGB and IR images if the input images only contain one identity. Meanwhile, if the images of two identities are mixed together as the inputs, the results of our proposed model are degraded. Nonetheless, it can still well predict the person masks of the first identity. While, for the second identity, it can only detect a small person region. This indicates that our proposed co-segmentation sub-network can select those id-related features from the inputs by generating those image-shared weights from their feature centers of different modalities.

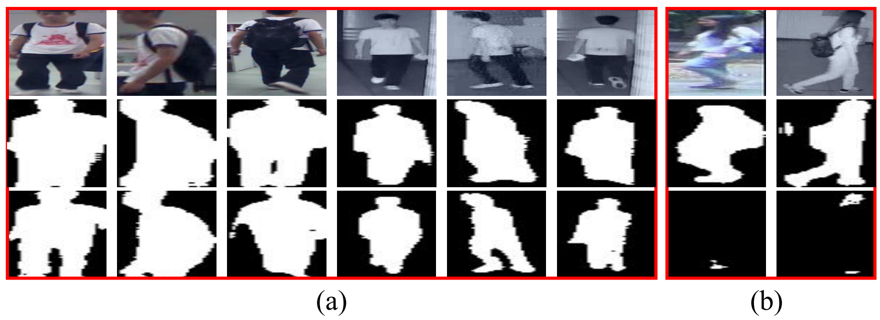


Figure 4: Person masks detected under different settings. (a) and (b) show two identities. The person masks in the second row are predicted by separately taking the images in (a) and (b) as the inputs. The person masks in the third row are obtained by simultaneously taking the images in (a) and (b) as the inputs.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets: Our proposed model is trained and evaluated on two publicly available datasets, *i.e.,* SYSU-MM01 [25] and RegDB [26]. SYSU-MM01 [25] a large-scale VI-ReID dataset, comprising RGB images and IR images from both indoor and outdoor scenes. It uses four visible cameras and two infrared cameras for data collection. The dataset includes two test modes: indoor-search and all-search, each with single-shot and multi-shot settings. RegDB [26] contains 8240 images from 412 person identities captured using several dual-mode cameras. It divides the images into a training set of 206 identities and a testing set of the remaining 206 identities. The dataset also includes two test modes: RGB-to-IR mode and IR-to-RGB mode.

Evaluation metrics: As in existing works [24, 27, 28, 17], the performance of our model is evaluated with the standard metrics (*i.e.,* Cumulated Matching Characteristics (CMC) and mean Average Precision (mAP)) in the ReID task. CMC evaluates the recognition accuracy of a model in the top-K matches, *i.e.,* R1, R10 and R20 in this paper. mAP is the ratio of the numbers of correctly matched pedestrians to the total number of matched pedestrians, which considers each pedestrian in the query and averages the AP (Average Precision) for each pedestrian.

4.2. Online Batch Sampling Strategy

In the training phase, we first sample N person identities for each batch from the dataset. For each selected identity, we randomly choose K RGB images and K IR images. Consequently, each batch contains a total of $2 \times N \times K$ images. In this paper, we set $N = 8$ and $K = 4$ for our training process.

In the testing stage, we extract person features from all query images and gallery images. Subsequently, we calculate the similarities between each query image and all gallery images using the Euclidean distance metric. Finally, we generate the ranking list for each query image by sorting the computed similarities in descending order.

4.3. Implementation details

We implement our proposed model using PyTorch libraries [29] and conduct its training and testing on an NVIDIA 2080Ti GPU. We first use a pre-trained ResNet50 to initialize the parameters of the feature extractor. After initializing some parameters using the Xavier algorithm [30], we optimize the model using the SGD (Stochastic Gradient Descent) algorithm with an initial learning rate of 0.01 and a weight decay of 0.0005. To prevent overfitting and ensure better convergence, we reduce the learning rates by

Table 1: Comparisons with some state-of-the-art models on SYSU-MM01 dataset.

-	All-Search								Indoor-Search							
-	Single-shot				Multi-shot				Single-shot				Multi-shot			
Methods	R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP
eBDTR [27]	27.8	67.3	81.3	28.4	-	-	-	-	32.4	77.4	89.6	42.4	-	-	-	-
AlignGAN[31]	42.4	85.0	93.7	40.7	51.5	89.4	95.7	33.9	45.9	87.6	94.4	54.3	57.1	92.7	97.4	45.3
ABP [17]	51.56	75.65	81.69	32.50	-	-	-	-	-	-	-	-	-	-	-	-
HATML [32]	55.29	92.41	97.36	53.89	-	-	-	-	62.10	95.75	99.20	69.37	-	-	-	-
DG-VAE [33]	59.49	93.77	-	58.46	-	-	-	-	-	-	-	-	-	-	-	-
BDF [34]	51.05	87.85	94.43	49.63	-	-	-	-	55.93	91.55	96.95	63.38	-	-	-	-
GECNet [35]	53.37	89.86	95.66	51.83	-	-	-	-	60.60	94.29	98.10	62.89	-	-	-	-
NFS [9]	56.91	91.34	96.52	55.45	63.51	94.42	97.81	48.56	62.79	96.53	99.07	69.79	70.03	97.70	99.51	61.45
FMI [10]	60.02	94.18	98.14	58.80	-	-	-	-	66.05	96.59	99.38	72.98	-	-	-	-
PSE [36]	61.68	93.10	97.17	57.51	-	-	-	-	63.41	91.69	95.28	68.17	-	-	-	-
DTRM[37]	63.03	93.82	97.56	58.63	-	-	-	-	66.35	95.58	98.80	71.76	-	-	-	-
SPOT[38]	65.34	92.73	97.04	62.25	-	-	-	-	69.42	96.22	99.12	74.63	-	-	-	-
ML [8]	67.25	95.38	98.46	64.29	72.95	96.94	99.27	57.62	69.58	96.66	99.03	74.37	80.39	98.80	99.83	68.60
OUR	70.13	96.15	98.79	65.32	77.06	97.87	99.28	59.23	71.00	96.96	98.99	75.21	83.22	98.99	99.78	70.20

a factor of 0.1 every 8 epochs. Furthermore, data augmentation techniques, such as random flipping, cropping, and erasing, are employed during training to enhance the model's generalization ability.

4.4. Comparison with SOTA models

In this subsection, the following SOTA VI-ReID methods: BDTR [40], DGD_MSR[41], AlignGAN[31], eBDTR [27], Hi-CMD[42], EDFL [43], BEAT [44], CMPG [45], HPILN[46], ABP [17], HATML [32], HC [24], DG-VAE [33], cm-SSFT [47], FBP-AL [48], DDSN [49], AMBT [39], BDF [34], GECNet [35], NFS [9], FMI [10], SPOT[38], DTRM[37] and PSE [36], are compared with our proposed VI-ReID model.

As shown in Table 1, our proposed model outperforms SOTA models in most metrics. Particularly, in the all-search mode with single-shot/multi-shot settings, our model achieves the best performance across all metrics.

Table 2: Comparisons with some state-of-the-art models on RegDB dataset.

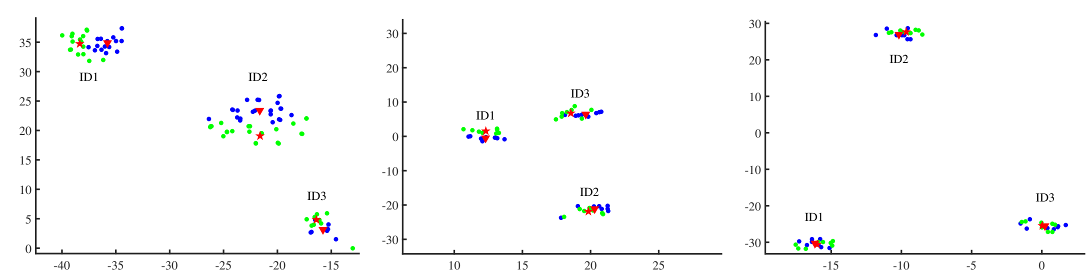
-	RGB-to-IR				IR-to-RGB			
Methods	R1	R10	R20	mAP	R1	R10	R20	mAP
eBDTR [27]	31.8	56.1	66.8	33.2	34.21	58.74	68.64	32.49
HATML [32]	71.83	87.16	92.16	67.56	70.02	86.45	91.61	66.30
DG-VAE [33]	72.97	86.89	-	71.78	-	-	-	-
AMBT [39]	71.10	-	-	68.10	-	-	-	-
GECNet [35]	82.33	92.72	95.49	78.45	78.93	91.99	95.44	75.58
NFS [9]	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
FMI [10]	73.2	-	-	71.6	71.8	-	-	70.1
SPOT[38]	80.35	93.48	96.44	72.46	79.37	92.79	96.01	72.26
DTRM[37]	79.09	92.25	95.66	70.09	78.02	91.75	95.19	69.56
PSE [36]	91.05	97.16	98.57	83.28	89.30	96.41	98.16	81.46
ML[8]	89.91	9 6.57	98.33	85.64	88.34	96.16	97.98	84.06
OUR	91.41	97.72	98.92	85.14	90.06	97.46	98.74	83.86

Table 3: Quantitative results of different ablation experiments.

Methods	r1	r10	MAP
Baseline	63.22	94.02	59.97
Baseline+Sel	64.05	93.96	60.32
Baseline+Seg_ReID	67.25	95.38	64.29
Baseline+Decoder	65.86	95.24	62.02
Baseline+CoSeg_ReID	70.13	98.79	65.32

Additionally, in the indoor-search mode with single-shot/multi-shot settings, our model achieves the best results in Rank-1, top-10 accuracies of CMC, and mAP. Moreover, it also achieves competitive results compared to the ML [8] method. These results indicate that our proposed model, with the aid of person masks and by exploring the relations between co-segmentation and VI-ReID, effectively extracts more discriminative modality-shared features from the input RGB and IR images for VI-ReID tasks.

Likewise, the results on the RegDB dataset, as presented in Table 2, further reinforce the effectiveness of our proposed model. Specifically, our model achieves competitive or superior results compared to most state-of-the-art models in both the RGB-to-IR and IR-to-RGB modes. Moreover, it achieves comparable results in the RGB-to-IR and IR-to-RGB modes. These findings serve as additional evidence of the effectiveness and robustness of our proposed model on the RegDB dataset.



(a) (b) (c)

Figure 5: Distributions of the features extracted by different models. (a) 'Baseline'. (b) 'Baseline+Seg_ReID'. (c) 'Baseline+CoSeg_ReID'. The green dots and the blue dots denote the RGB features and the IR features of different identities, respectively. Accordingly, the red pentagrams and the blue triangles denotes the centers of RGB features and IR features, respectively. These figures are visualized by using the T-SNE algorithm[50].

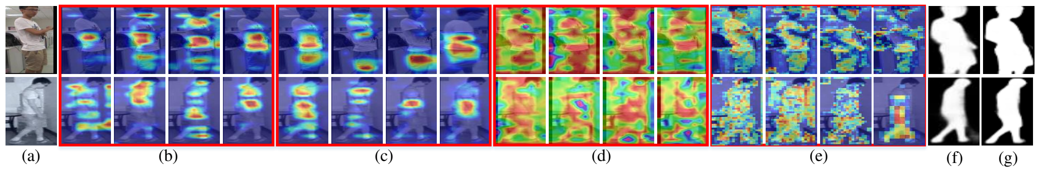


Figure 6: Illustration of the features extracted by different models. (a) RGB and IR images. (b) (c) (d) and (e) The features extracted by 'Baseline', 'Baseline+Sel', 'Baseline+Seg_ReID' and our proposed model, respectively. (f)Person masks predicted by our proposed model. (g) The pseudo ground truth maps generated by [8].

4.5. Ablation study

In this section, we conduct several ablation experiments on the SYSU-MM01 dataset to validate the effectiveness of each component in our proposed model.

4.5.1. Effectiveness of each component in our proposed model

We verify each component of our proposed model. As shown in Table 3, we first remove the auxiliary co-segmentation model from our proposed model. The model denoted as 'Baseline+Sel' uses person masks for feature selection. In other words, it employs the person masks (as shown in Fig. 3(a)) to select modality-shared features from the person regions. Subsequently, these selected features are fed into the part module for further processing. 'Baseline+Seg_ReID' denotes the model in Fig. 3(b), which performs multi-task learning with segmentation and VI-ReID. 'Baseline+Decoder' denotes the model that removes the CCWG module from our proposed model. It is also a multi-task learning based model, which stacks the segmentation sub-network after the VI-ReID sub-network rather than parallels them. While, 'Baseline+CoSeg_ReID' is our final model. The quantitative results of different models are shown in Table 3.

The results of ‘Baseline+Sel’ indicate that taking the person masks for feature selection may slightly improve the performance. This may be due to the fact that, although the VI-ReID model can reduce the interfering information within backgrounds to some extent via those person masks, the VI-ReID model cannot learn how to extract those person semantics by itself, since those person masks do not provide gradients for training in such a feature selection way. Besides, this model may also discard some personal information, since those person masks may be incomplete. The results of ‘Baseline+Seg_ReID’ indicate that the multi-task learning based VI-ReID model can obtain better results. This may owe to the fact that it can directly extract many person semantics from the person masks.

The results of ‘Baseline+Decoder’ indicate that directly taking segmentation as an auxiliary model and linking it after a VI-ReID model obtain sub-optimal results. This may result from the task difference between person segmentation and VI-ReID, *i.e.,* the person segmentation task aims to extract those person-related information without caring about their identities, while VI-ReID tries to extract those identity-related person information. Differently, compared with ‘Baseline+Decoder’, ‘Baseline+CoSeg_ReID’, *i.e.,* our final model, which employs the CCWG module for co-segmentation, significantly boosts the performance and becomes the best one. This indicates that our proposed CCWG module can address the task difference by segmenting the same identities across a set of input images, and can extract more person semantics from the input images for VI-ReID, thus obtaining better results.

4.5.2. Visualization of the feature distributions of different models

The distributions of features extracted by 'Baseline', 'Baseline+Seg_ReID' and 'Baseline+CoSeg_ReID' are shown in Fig.5, respectively. It can be seen that, compared with 'Baseline', 'Baseline+Seg_ReID' can better reduce the modality discrepancy, since it can effectively extract more discriminative modality-invariant features from the input RGB/IR images by exploring their inner relations between VI-ReID and segmentation. While, compared with 'Baseline+Seg_ReID', our proposed model 'Baseline+CoSeg_ReID' can further reduce the large modality discrepancy, due to the fact that our proposed model can simultaneously explore the relations between person segmentation and VI-ReID for extracting more discriminative person-related features, and the relations between the features of two modalities for reducing the cross-modality variation by using the co-segmentation as an auxiliary model.

4.5.3. Visualization of those person masks and features from different models

Fig. 6 shows the person masks and features extracted by different models, which is obtained by first normalizing the features extracted by our proposed model via min-max normalization and combining them with the inputs, thus generating those heatmaps. The visualized features are from the last feature extraction block of different models, which are taken as the heatmaps and projected into the input images.

Fig. 6(c) proves that the models, simply taking the person masks for feature selection, can eliminate those background information, but cannot learn to extract more accurate person-related semantics for VI-ReID. Fig. 6(d) and Fig. 6(e) show that, even without providing those person masks, such multi-task learning based models have already learned to extract more

Table 4: Number of parameters of different models.

Models	BDTR [40]	HC[24]	PSE [36]	ML [8]	OUR (training)	OUR (testing)
Parameters (M)	48.2	58.6	33.2	46.8	52.6	32.1

accurate person-related semantics from the input images for VI-ReID. Furthermore, they also reveal that our proposed model pays more attention on the persons than on the backgrounds. This may result from the fact that, by virtue of our proposed CCWG, our proposed model will interact a set of images and generate shared weights for segmenting their common objects, thus helping our proposed model to focus more on the foregrounds and less on the backgrounds. Consequently, the VI-ReID network can extract more modality-shared person-related features for further improving results. As shown in Fig. 6(f) and Fig. 6(g), our proposed model can well predict the person masks, which also proves that our proposed model can learn abundant person-related semantics for mask prediction.

4.5.4. Number of parameters

As shown in Table. 4, we further compare the number of parameters between our proposed model and some existing modality-shared feature learning based models. It should be noted that BDTR and HC employ two full ResNet50 for feature extraction. While, HC, ML and our proposed model share the feature extractors for modality-shared feature extraction. It can be seen that, if removing the segmentation model in the testing stage, our proposed model will reduce its parameters from 52.6M to 32.1M. As a result, this enables our proposed model to have competitive and even fewer parameters than others during the test stage.

5. Conclusion

This paper presents a novel multi-task learning framework that uses the co-segmentation to assist the VI-ReID by bridging the two tasks via the exploitation of their common concepts, *i.e.,* semantic similarity. By doing so, the co-segmentation model can effectively enhance the VI-ReID network's feature extraction ability of extracting more person shape information via person mask prediction. Furthermore, the co-segmentation model can also help the VI-ReID network to interact those features across different modalities when segmenting the same objects from a set of multi-modality images, thus reducing their large cross-modality variations. Consequently, the VI-ReID network extracts more discriminative and modality-invariant modality-shared features for VI-ReID and achieves significant performance improvements. Moreover, the auxiliary co-segmentation model is only employed in the training stage and is removed in the testing stage, thus increasing no more parameters and computational costs. Theoretical analysis and experimental results both validate the superiorities of our model over existing ones.

6. Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No.61773301. It is also supported by the Shaanxi Innovation Team Project under Grant No.2018TD-012 and the China Postdoctoral Science Foundation under Grant No.2023M742745.

References

[1] F.-P. An, J. e Liu, Pedestrian re-identification algorithm based on visual attention-positive sample generation network deep learning model, Information Fusion 86-87 (2022) 136–145.

[2] J. Suutala, J. Röning, Methods for person identification on a pressure-sensitive floor: Experiments with multiple classifiers and reject option, Information Fusion 9 (1) (2008) 21–40.

[3] N. Huang, J. Liu, Y. Miao, Q. Zhang, J. Han, Deep learning for visible-infrared cross-modality person re-identification: A comprehensive review, Information Fusion 91 (2023) 396–411.

[4] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. H. Hoi, Deep learning for person re-identification: A survey and outlook, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (6) (2022) 2872–2893.

[5] Z. Chang, S. Yang, Z. Feng, Q. Gao, S. Wang, Y. Cui, Semantic-relation transformer for visible and infrared fused image quality assessment, Information Fusion 95 (2023) 454–470.

[6] M. Qi, S. Wang, G. Huang, J. Jiang, J. Wu, C. Chen, Mask-guided dual attention-aware network for visible-infrared person re-identification, Multimedia Tools and Applications 80 (12) (2021) 17645–17666.

[7] Z. Zhao, B. Liu, Q. Chu, Y. Lu, N. Yu, Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification, Proceedings of the AAAI Conference on Artificial Intelligence 35 (4) (2021) 3520–3528.

[8] N. Huang, K. Liu, Y. Liu, Q. Zhang, J. Han, Cross-modality person re-identification via multi-task learning, Pattern Recognition 128 (2022) 108653.

[9] Y. Chen, L. Wan, Z. Li, Q. Jing, 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.

[10] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, L. Ma, Farewell to mutual information: Variational distillation for cross-modal person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1522–1531.

[11] W. Li, O. H. Jafari, C. Rother, Deep object co-segmentation, in: Proceedings of the Asian Conference on Computer Vision, 2018, pp. 638–653.

[12] H. Chen, Y. Huang, H. Nakayama, Semantic aware attention based deep object co-segmentation, in: Proceedings of the Asian Conference on Computer Vision, 2018, pp. 435–450.

[13] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. H. Hoi, Deep learning for person re-identification: A survey and outlook, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (6) (2022) 2872–2893.

[14] M. Jia, X. Cheng, S. Lu, J. Zhang, Learning disentangled representation implicitly via transformer for occluded person re-identification, IEEE Transactions on Multimedia 25 (2023) 1294–1305.

[15] X. Yang, P. Zhou, M. Wang, Person reidentification via structural deep metric learning, IEEE Transactions on Neural Networks and Learning Systems 30 (10) (2019) 2987–2998.

[16] C. Chen, M. Ye, M. Qi, B. Du, Sketch transformer: Asymmetrical disentanglement learning from dynamic synthesis, in: Proceedings of the ACM International Conference on Multimedia, 2022, pp. 4012–4020.

[17] Z. Wei, X. Yang, N. Wang, B. Song, X. Gao, ABP: Adaptive body partition model for visible infrared person re-identification, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2020, pp. 1–6.

[18] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, C.-W. Lin, Structure-aware positional transformer for visible-infrared person re-identification, IEEE Transactions on Image Processing 31 (2022) 2352–2364.

[19] Y. Miao, N. Huang, X. Ma, Q. Zhang, J. Han, On exploring pose estimation as an auxiliary learning task for visible–infrared person re-identification, Neurocomputing 556 (2023) 126652.

[20] J. Liu, J. Wang, N. Huang, Q. Zhang, J. Han, Revisiting modality-specific feature compensation for visible-infrared person re-identification, IEEE Transactions on Circuits and Systems for Video Technology 32 (10) (2022) 7226–7240.

[21] Q. Zhang, C. Lai, J. Liu, N. Huang, J. Han, Fmcnet: Feature-level modality compensation for visible-infrared person re-identification, in:

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7349–7358.

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 770–778.

[23] Y. Sun, L. Zheng, Y. Yang, Q. Tian, 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, 2018, pp. 501–518.

[24] Y. Zhu, Z. Yang, L.-C. Wang, S. Zhao, X. Hu, D. Tao, Hetero-center loss for cross-modality person re-identification, Neurocomputing 386 (2020) 97–109.

[25] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J.-H. Lai, RGB-Infrared cross-modality person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5390–5399.

[26] D. T. Nguyen, H. G. Hong, K. W. Kim, K. R. Park, Person recognition system based on a combination of body images from visible light and thermal cameras, Sensors 17 (3) (2017) 605.

[27] M. Ye, X. Lan, Z. Wang, P. C. Yuen, Bi-directional center-constrained top-ranking for visible thermal person re-identification, IEEE Transactions on Information Forensics and Security 15 (2020) 407–419.

[28] Y. Hao, J. Li, N. Wang, X. Gao, Modality adversarial neural network for

visible-thermal person re-identification, Pattern Recognition 107 (2020) 107533.

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Proceedings of the Neural Information Processing Systems, 2019, pp. 8026–8037.

[30] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feed-forward neural networks, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.

[31] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, Z. Hou, RGB-Infrared cross-modality person re-identification via joint pixel and feature alignment, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3622–3631.

[32] M. Ye, J. Shen, L. Shao, Visible-infrared person re-identification via homogeneous augmented tri-modal learning, IEEE Transactions on Information Forensics and Security 16 (2020) 728–739.

[33] N. Pu, W. Chen, Y. Liu, E. M. Bakker, M. S. Lew, Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification, in: Proceedings of the ACM International Conference on Multimedia, 2020, pp. 2149–2158.

[34] H. Liu, Z. Miao, B. Yang, R. Ding, A base-derivative framework for cross-modality rgb-infrared person re-identification, in: Proceedings of the International Conference on Pattern Recognition, 2021, pp. 7640–7646.

[35] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia, C.-W. Lin, Grayscale enhancement colorization network for visible-infrared person re-identification, IEEE Transactions on Circuits and Systems for Video Technology (2021).

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

[37] M. Ye, C. Chen, J. Shen, L. Shao, Dynamic tri-level relation mining with attentive graph for visible infrared re-identification, IEEE Transactions on Information Forensics and Security 17 (2022) 386–398.

[38] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, C.-W. Lin, Structure-aware positional transformer for visible-infrared person re-identification, IEEE Transactions on Image Processing 31 (2022) 2352–2364.

[39] Y. Huang, Q. Wu, J. Xu, Y. Zhong, P. Zhang, Z. Zhang, Alleviating modality bias training for infrared-visible person re-identification, IEEE Transactions on Multimedia (2021).

[40] M. Ye, Z. Wang, X. Lan, P. C. Yuen, Visible thermal person re-identification via dual-constrained top-ranking, in: Proceedings of the

International Joint Conference on Artificial Intelligence, 2018, pp. 1092–1099.

[41] Z. Feng, J. Lai, X. Xie, Learning modality-specific representations for visible-infrared person re-identification, IEEE Transactions on Image Processing 29 (2019) 579–590.

[42] S. Choi, S. Lee, Y. Kim, T. Kim, C. Kim, Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10257–10266.

[43] H. Liu, J. Cheng, W. Wang, Y. Su, H. Bai, Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification, Neurocomputing 398 (2020) 11–19.

[44] H. Ye, H. Liu, F. Meng, X. Li, Bi-directional exponential angular triplet loss for rgb-infrared person re-identification, IEEE Transactions on Image Processing 30 (2020) 1583–1595.

[45] Y. Yang, T. Zhang, J. Cheng, Z. Hou, P. Tiwari, H. M. Pandey, et al., Cross-modality paired-images generation and augmentation for RGB-Infrared person re-identification, Neural Networks 128 (2020) 294–304.

[46] Y.-B. Zhao, J.-W. Lin, Q. Xuan, X. Xi, Hpiln: A feature learning framework for cross-modality person re-identification, IET Image Processing 13 (14) (2019) 2897–2904.

[47] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, N. Yu, Cross-modality

person re-identification with shared-specific feature transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 13379–13389.

[48] Z. Wei, X. Yang, N. Wang, X. Gao, Flexible body partition-based adversarial learning for visible infrared person re-identification, IEEE Transactions on Neural Networks and Learning Systems (2021).

[49] Y. Cheng, X. Li, G. Xiao, W. Ma, X. Gou, Dual-path deep supervision network with self-attention for visible-infrared person re-identification, in: Proceedings of the IEEE International Symposium on Circuits and Systems, 2021, pp. 1–5.

[50] L. van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (86) (2008) 2579–2605.