



This is a repository copy of *Co-segmentation assisted cross-modality person re-identification*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/206632/>

Version: Accepted Version

---

**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>

---

© 2023 The Authors. Except as otherwise noted, this author-accepted version of a journal article published in *Information Fusion* is made available via the University of Sheffield Research Publications and Copyright Policy under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.



[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>

# Co-segmentation Assisted Cross-modality Person Re-identification

Nianchang Huang<sup>a,b</sup>, Baichao Xing<sup>a,b</sup>, Qiang Zhang<sup>a,b,\*</sup>, Jungong Han<sup>c</sup>, Jin Huang<sup>a</sup>

<sup>a</sup>*Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an, 710071, Shaanxi, China*

<sup>b</sup>*Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an, 710071, Shaanxi, China*

<sup>c</sup>*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),  
baichaoxing@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

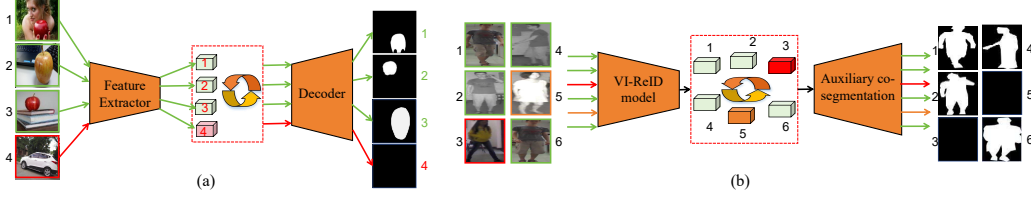


Figure 1: Frameworks of existing co-segmentation models. (a) Architecture of the co-segmentation task. (b) Framework of our proposed model. The images with the same color boxes belong to the same objects (a) or identities (b).

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

35 their extracted person features, but cannot enable VI-ReID models themself  
 36 to extract more accurate semantics from the input images, since those person  
 37 masks do not directly provide any gradients for training (see Section III for  
 38 more details), thus leading to unsatisfactory results.

39 Alternatively, Huang *et. al* [8] proposed a multi-task learning based VI-  
 40 ReID model to facilitate their VI-ReID network extracting more modality-  
 41 shared person shape information for VI-ReID by exploring the relations  
 42 between person segmentation and VI-ReID. Specifically, in [8], two sub-  
 43 networks are employed on top of a shared feature extractor for person seg-  
 44 mentation and VI-ReID, respectively. By doing so, the person segmentation  
 45 sub-network can facilitate the shared feature extractor directly extracting  
 46 abundant person-related semantics for VI-ReID by predicting those person  
 47 masks. Meanwhile, those person semantics extracted by the person segmen-  
 48 tation sub-network can also be introduced into the VI-ReID sub-network  
 49 for further boosting performance, thus achieving large performance improve-  
 50 ments. *However, this model mainly focuses on exploring the relations between*  
 51 *different tasks but ignores the relations between the features across modali-*  
 52 *ties, thus also easily leading to sub-optimal results. Besides, it also introduces*  
 53 *some extra computational costs, since an extra segmentation sub-network is*  
 54 *required.*

55 In this paper, we gain inspiration from the task of co-segmentation and  
 56 eventually utilize it to address the above issues. Specifically, co-segmentation  
 57 aims to detect the common objects or regions in a set of relevant images, *e.g.*,  
 58 the apples in Fig. 1(a). One of the main ideas of such a task is to utilize  
 59 semantic similarity for segmenting objects with the same semantic class but

60 with different appearances and backgrounds. In deep learning based models,  
 61 the semantic similarity usually means that the distributions of high-level  
 62 features, which are also called semantic features, are the same for those  
 63 images with the same classes, but different for those images with different  
 64 classes, *e.g.*, the semantic features' distributions of apples *vs* those of cars  
 65 in Fig. 1 [9, 10]. Accordingly, those deep learning based models can achieve  
 66 co-segmentation by interacting those semantic features from different input  
 67 images. For instance, a correlation layer can help to segment the objects  
 68 with the same class across two input images, which can be implemented by  
 69 either computing the correlations of the semantic features [11] or employing  
 70 the shared weights to select a set of features from a set of input images for  
 71 segmenting their co-existing objects [12].

72 Moreover, as depicted in Fig. 1(b), the concept of semantic similarity  
 73 appears reasonable for VI-ReID. It typically aims to enable the distributions  
 74 of RGB images and IR images to be the same for the same identity, and  
 75 vice versa. Therefore, in this paper, we use the co-segmentation to facilitate  
 76 the VI-ReID by transferring the VI-ReID task to a task that detects the  
 77 same identities from a set of RGB and IR images. If incorporating the co-  
 78 segmentation task into the framework of VI-ReID models, the co-segmentation  
 79 network will exploit the semantic similarity across a set of RGB and IR im-  
 80 ages with the same identities for predicting the masks of this identity, which  
 81 will enhance the VI-ReID network to extract more additional discriminative  
 82 modality-shared person shape information for VI-ReID. Moreover, during the  
 83 co-segmentation, the features from the set of RGB and IR images will also be  
 84 implicitly interacted with each other. This will further help the VI-ReID net-

work 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-

110 segmentation model, thus obtaining better VI-ReID results.

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

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

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

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

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



## 131 2. Related Work

### 132 2.1. VV-PReID

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

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

## 154 2.2. VI-ReID

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

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

179 In contrast, modality-specific feature compensation-based models typi-  
 180 cally start by generating the missing modality information from the avail-  
 181 able modality, thus addressing cross-modality variations. Subsequently, they  
 182 utilize both the original and generated information to handle intra-modality  
 183 variations. Liu *et al.*[20] proposed a new two-stage GAN based model, which  
 184 first optimizes the image generator’s structures and objective functions in the  
 185 first stage. Then, it improves the ReID network by employing the feature-  
 186 level fusion rather than the image-level fusion for their original and gener-  
 187 ated information in the second stage. Differently, Lai *et al.*[21] introduced a  
 188 feature-level compensation approach for VI-ReID. They first disentangled the  
 189 single-modality features into modality-specific features and modality-shared  
 190 features. Subsequently, they generated the missing modality-specific features  
 191 from the disentangled modality-shared features. Finally, they fused the orig-  
 192 inal modality-specific features, the generated modality-specific features, and  
 193 the disentangled modality-shared features to perform VI-ReID. This feature-  
 194 level compensation strategy allowed for better handling of cross-modality  
 195 variations and intra-modality variations in their model.

### 196 **3. Proposed Model**

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

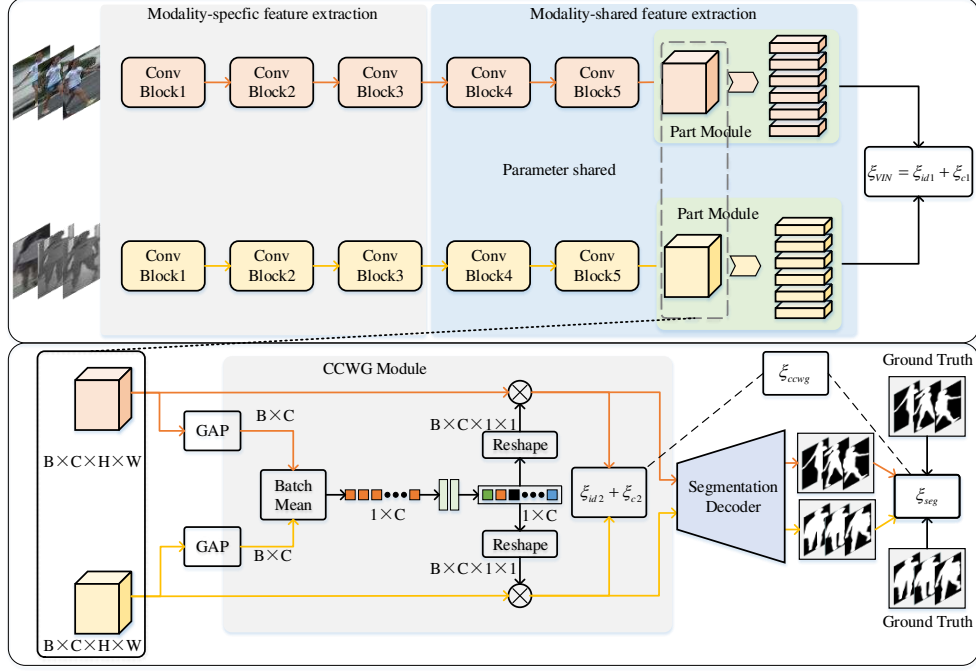


Figure 2: Illustration of the proposed model. In our proposed model, VI-ReID network first employs two independent subnetworks to extract those single-modality RGB and IR features from the input RGB and IR images, respectively, and further uses two parameter-shared subnetworks to extract those modality-shared 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,

208 where their modality-shared features are extracted using two sub-networks  
 209 with shared parameters. Finally, a part module is applied to the extracted  
 210 modality-shared features to obtain the final person part features for VI-ReID.  
 211 Suppose that there are  $N$  identities and each identity has  $K$  RGB images  
 212 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  
 213  $X_I = \{x_I^i \in R^{H \times W}\}_i^J$ . Here,  $J = NK$ .

### 214 3.1.1. Modality-specific Feature Extraction

215 Typically, the low-level features obtained from RGB and IR images, re-  
 216 spectively, exhibit significant discrepancies due to their capture in different  
 217 spectrums. Therefore, as shown in Fig. 2, two shallower sub-networks are  
 218 employed in our proposed VI-ReID network to extract those single-modality  
 219 features from the input RGB images and IR images, respectively. Here, the  
 220 sub-networks have the same structure but extract different modality-specific  
 221 information using distinct parameters. Specifically, the two shallower sub-  
 222 networks are constructed by using the same structures as the first three con-  
 223 volutional blocks in ResNet50 [22]. As a result, we obtain the RGB features  
 224  $\mathbf{F}_R$  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  
 225 expressed by

$$\mathbf{F}_m = \text{ConvB}(X_m, \alpha_m), \quad (1)$$

226 where  $\text{ConvB}(*, \alpha_m)$  denotes a sub-network with its corresponding parame-  
 227 ters  $\alpha_m$ .

### 228 3.1.2. Modality-shared Feature Extraction

229 As shown in Fig. 2, the extracted single-modality RGB and IR features  
 230 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

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

where,  $m \in \{R, I\}$  denotes RGB or IR modality.  $\text{ConvB}(*, \beta)$  denotes the convolutional blocks with the parameters  $\beta$ .  $C_1$  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  $\hat{\mathbf{F}}_{s,m}^p \in R^{J \times C_1}$ . Here,  $p = 1, 2, 3, \dots, 6$  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  $\mathbf{F}_m^p \in R^{J \times C_2}$ . Finally, a fully connected layer based classifier is employed to predict their corresponding person identities, obtaining their corresponding

253 scores  $cls_m^p, p=1,2,\dots,6$ . Mathematically, this process is expressed by

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

254 where  $\text{Sep}(\cdot)$  denotes the separation operation.  $\text{GAP}(\cdot)$  denotes the global  
 255 average pooling.  $\text{FC}(\cdot, \gamma_p)$  and  $\text{FC}(\cdot, \gamma_s)$  denote the fully connected layers  
 256 with their corresponding parameters  $\gamma_p$  and  $\gamma_s$ , respectively. In this paper,  
 257  $C_1 = 2048$  and  $C_2 = 512$ .

#### 258 3.1.4. Loss Function

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

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

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

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

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

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

271 by

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

272 where  $\xi_{hc}$  is the hetero-center loss [24] and is expressed by

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

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

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

275 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  
276 and IR images of the  $n$ -th identity. Therefore, the total loss for training the  
277 VI-ReID network is expressed by

$$\xi_{VIN} = \xi_{id1} + \xi_{c1}. \quad (9)$$

### 278 3.2. Co-segmentation Network

279 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.  
280 (2) will be further fed into an auxiliary co-segmentation network to predict  
281 their person masks. It should be noted that, different from person segmenta-  
282 tion, which may directly predict the person mask from each modality image to  
283 facilitate the extraction of person semantics for VI-ReID, the co-segmentation  
284 network aims to detect the person masks of one certain identity from a set  
285 of images across different modalities.

286 For that, as shown in Fig. 2, the co-segmentation network first employs  
287 a cross-modality center based weight generation (CCWG) module, which



288 will explore the relations across the modality-shared RGB and IR features  
 289 extracted from the  $K$  RGB images and  $K$  IR images of one identity, and  
 290 accordingly generate a set of shared co-segmentation weights for the  $K$  RGB  
 291 images and  $K$  IR images. Then, by virtue of the generated weights, the  
 292 co-segmentation network will select those unique features about this identity  
 293 from such modality-shared features for each RGB or IR image. Subsequently,  
 294 for each RGB or IR image of this identity, the co-segmentation network will  
 295 further employ a parameter-shared decoder to predict a person mask of this  
 296 identity by using those selected modality-shared features from this image.  
 297 Through this approach, our proposed model can effectively explore the re-  
 298 lationships between person segmentation and VI-ReID, resulting in the ex-  
 299 traction of more discriminative person-related features. Simultaneously, the  
 300 model also considers the relationships between the features of the two modal-  
 301 ities, which helps in reducing cross-modality variations. Consequenceally, our  
 302 model achieves enhanced performance in handling both intra-modality and  
 303 cross-modality challenges for VI-ReID tasks. In the following content, we  
 304 will take the  $n$ -th identity as an example for the introduction.

### 305 *3.2.1. Cross-modality Center based Weight Generation (CCWG) Module*

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

Specifically, for the  $n$ -th identity, the VI-ReID network can extract its modality-shared RGB features  $\mathbf{F}_{s,R}^n \in R^{K \times C_1 \times \frac{W}{16} \times \frac{H}{16}}$  from  $K$  RGB images and IR features  $\mathbf{F}_{s,I}^n \in R^{K \times C_1 \times \frac{W}{16} \times \frac{H}{16}}$  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  $\mathbf{F}_{Center}^n \in R^{C_1 \times 1 \times 1}$  for the  $n$ -th identity by

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

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

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

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.*,

$$\mathbf{w}^n = \text{FC}(\mathbf{F}_{Center}^n, \theta), \quad (12)$$

where  $\text{FC}(*, \theta)$  denotes a fully connected layer with its corresponding parameters  $\theta$ . Here, the weights  $w^n$  are shared for all the RGB or IR images of the  $n$ -th identity, since the single-modality features  $\mathbf{F}_{sel,m}^{n,k}$  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

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

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

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

where  $\text{DConv}(*, \lambda_1)$  and  $\text{DConv}(*, \lambda_2)$  denote two deconvolutional blocks with their parameters  $\lambda_1$  and  $\lambda_2$ , 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

353 masks of one certain identity from a set of images with different modalities.  
 354 This will enhance the VI-ReID network’s ability of extracting more discriminative  
 355 modality-shared features, especially for those modality-invariant features related to  
 356 person shapes. Meanwhile, the single-modality features from a set of RGB and IR  
 357 images with the same identity will implicitly interact with each other in the CCWG  
 358 module via the gradient backpropagation. This will also help the VI-ReID network to  
 359 reduce the large modality discrepancies, thus further boosting performance.

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

### 368 3.2.3. Loss Functions

369 Three loss functions, including an identity loss  $\xi_{id2}$ , a center loss  $\xi_{c2}$   
 370 and a segmentation loss  $\xi_{seg}$ , are employed for training the proposed co-segmentation  
 371 network.

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

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

375 where  $cls_g$  denotes their corresponding ground truth labels. Similar to Eq.(6),

the center loss  $\xi_{c2}$  is performed on the person part features  $\mathbf{F}_{sel,m}^p$ , aiming to make the features from the same identity to be compact, *i.e.*,

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

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

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

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

$$\xi_{ccwg} = \xi_{id2} + \xi_{c2} + \xi_{seg}. \quad (18)$$

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

$$\xi_{total} = \xi_{VIN} + \xi_{ccwg}. \quad (19)$$

### 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.

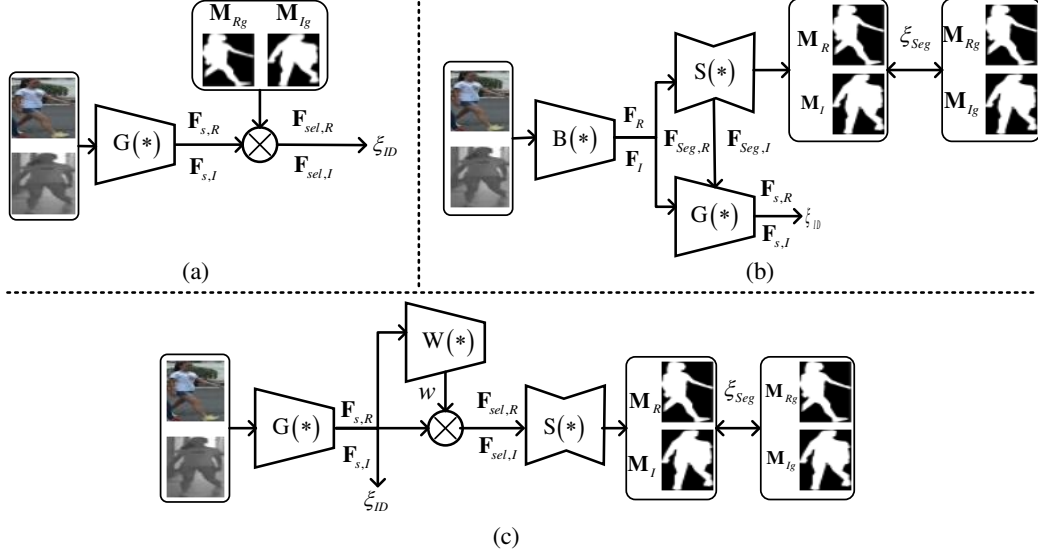


Figure 3: Illustration of different ways of using those person maps. (a) Simply using person maps for feature selection, *i.e.*, [6] and [7]. (b) Exploring person maps via existing multi-task learning frameworks, *i.e.*, [8]. (c) Our proposed model.

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

397 **Feature selection:** The simplified structures of the feature selection  
 398 based models are shown in Fig. 3(a). The input RGB/IR images are directly  
 399 fed into the VI-ReID network  $G(*, \epsilon_G)$  for extracting their corresponding  
 400 modality-shared features  $F_{s,R}/F_{s,I}$ . Here,  $\epsilon_G$  denotes the VI-ReID network's  
 401 parameters. Then, the person masks  $M_{Rg}$  and  $M_{Ig}$  are used for selecting  
 402 those person-related features. Finally, the selected features are employed for  
 403 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

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

Here,

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

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,  $\epsilon_B$ ,  $\epsilon_S$  and  $\epsilon_G$  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 ( $\xi_{ID}$ ) and the segmentation loss ( $\xi_{Seg}$ ). Accordingly, the gradients from the total loss to the VI-ReID network are computed by

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

423 And,

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

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

438 **Multi-task learning framework based on co-segmentation and**  
439 **VI-ReID (our model):** The simplified structure of our proposed model is  
440 shown in Fig.3(c). It first employs a VI-ReID sub-network  $G(*, \epsilon_G)$  to ex-  
441 tract those modality-shared features from the input images. Then, a weight



442 generation sub-network  $W(*, \epsilon_W)$  is employed to predict the weights  $\mathbf{w}$  for  
 443 selecting a set of unique features of one identity. Here,  $\epsilon_S$  also denotes the  
 444 parameters of the weight generation sub-network. Finally, the selected fea-  
 445 tures will be fed into a co-segmentation sub-network  $S(*, \epsilon_S)$  to segment those  
 446 objects co-existing within the input images. Accordingly, the gradients from  
 447 the total loss, including the ID loss ( $\xi_{ID}$ ) and the segmentation loss ( $\xi_{Seg}$ ),  
 448 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} \left( \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} \right.} \\
 &+ \left. \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} \right).
 \end{aligned} \tag{24}$$

449 Similar to that in Eq.(24), the proposed model can also effectively explore  
 450 the relations between the tasks of VI-ReID and person segmentation via  
 451 the two items marked by the green boxes in Eq.(24). Accordingly, the VI-  
 452 ReID network can also learn the ability of extracting those accurate and  
 453 modality-invariant person semantics from the person masks for VI-ReID.  
 454 Moreover, as shown in the last item of Eq. (24) (marked by the red box),  
 455 the modality-shared RGB features and the modality-shared IR features will  
 456 be interacted with each other with the aid of those generated weights, thus  
 457 benefiting to reduce their modality discrepancies. Accordingly, the proposed

458 model can effectively explore the relationships between person segmentation  
 459 and VI-ReID, leading to the extraction of more discriminative person-related  
 460 features. Moreover, it also considers the relationships between the features  
 461 of the two modalities, which helps in reducing cross-modality variations.  
 462 As a result, the model achieves improved performance by addressing both  
 463 intra-modality and cross-modality challenges for VI-ReID tasks. Besides,  
 464 the co-segmentation network only appears in the training stage, which does  
 465 not introduce any more parameters in the testing stage.

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

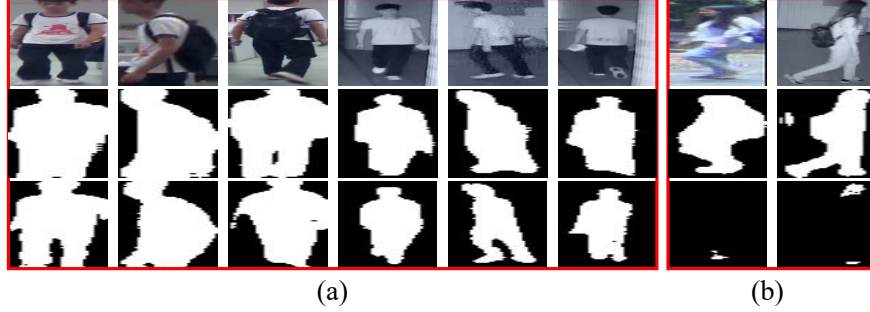


Figure 4: Person masks detected under different settings. (a) and (b) show two identities. The person masks in the second row are predicted by separately taking the images in (a) and (b) as the inputs. The person masks in the third row are obtained by simultaneously taking the images in (a) and (b) as 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	96.57	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

533 Additionally, in the indoor-search mode with single-shot/multi-shot settings,  
534 our model achieves the best results in Rank-1, top-10 accuracies of CMC, and

535 mAP. Moreover, it also achieves competitive results compared to the ML [8]  
 536 method. These results indicate that our proposed model, with the aid of  
 537 person masks and by exploring the relations between co-segmentation and  
 538 VI-ReID, effectively extracts more discriminative modality-shared features  
 539 from the input RGB and IR images for VI-ReID tasks.

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

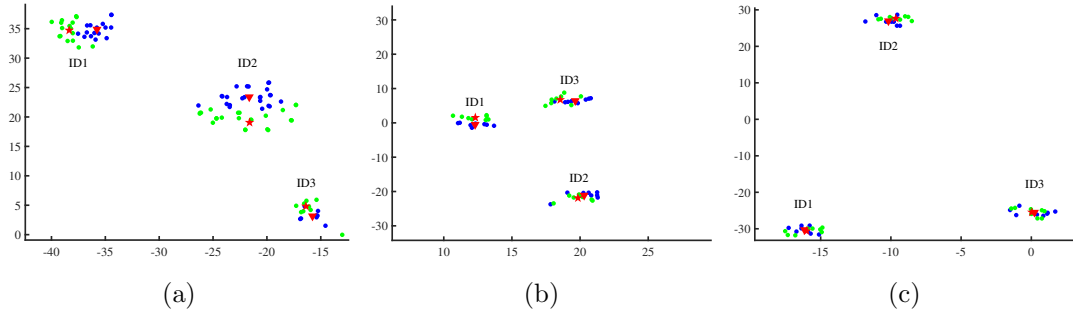


Figure 5: Distributions of the features extracted by different models. (a) ‘Baseline’. (b) ‘Baseline+Seg\_ReID’. (c) ‘Baseline+CoSeg\_ReID’. The green dots and the blue dots denote the RGB features and the IR features of different identities, respectively. Accordingly, the red pentagrams and the blue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

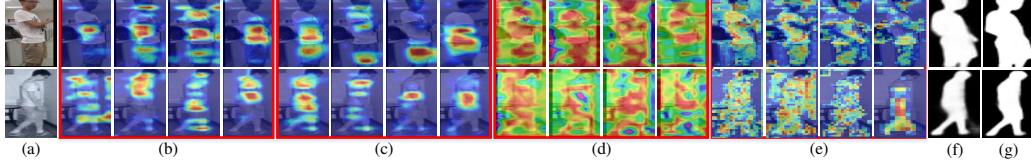


Figure 6: Illustration of the features extracted by different models. (a) RGB and IR images. (b) (c) (d) and (e) The features extracted by ‘Baseline’, ‘Baseline+Sel’, ‘Baseline+Seg\_ReID’ and our proposed model, respectively. (f) Person masks predicted by our proposed model. (g) The pseudo ground truth 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 dif-



564 ferent models are shown in Table 3.

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

576 The results of ‘Baseline+Decoder’ indicate that directly taking segmen-  
577 tation as an auxiliary model and linking it after a VI-ReID model obtain  
578 sub-optimal results. This may result from the task difference between person  
579 segmentation and VI-ReID, *i.e.*, the person segmentation task aims to extract  
580 those person-related information without caring about their identities, while  
581 VI-ReID tries to extract those identity-related person information. Differ-  
582 ently, compared with ‘Baseline+Decoder’, ‘Baseline+CoSeg\_ReID’, *i.e.*, our  
583 final model, which employs the CCWG module for co-segmentation, signifi-  
584 cantly boosts the performance and becomes the best one. This indicates that  
585 our proposed CCWG module can address the task difference by segmenting  
586 the same identities across a set of input images, and can extract more person  
587 semantics from the input images for VI-ReID, thus obtaining better results.

#### 588 4.5.2. Visualization of the feature distributions of different models

589 The distributions of features extracted by ‘Baseline’, ‘Baseline+Seg\_ReID’  
590 and ‘Baseline+CoSeg\_ReID’ are shown in Fig.5, respectively. It can be seen  
591 that, compared with ‘Baseline’, ‘Baseline+Seg\_ReID’ can better reduce the  
592 modality discrepancy, since it can effectively extract more discriminative  
593 modality-invariant features from the input RGB/IR images by exploring their  
594 inner relations between VI-ReID and segmentation. While, compared with  
595 ‘Baseline+Seg\_ReID’, our proposed model ‘Baseline+CoSeg\_ReID’ can fur-  
596 ther reduce the large modality discrepancy, due to the fact that our proposed  
597 model can simultaneously explore the relations between person segmentation  
598 and VI-ReID for extracting more discriminative person-related features, and  
599 the relations between the features of two modalities for reducing the cross-  
600 modality variation by using the co-segmentation as an auxiliary model.

#### 601 4.5.3. Visualization of those person masks and features from different models

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

608 Fig. 6(c) proves that the models, simply taking the person masks for  
609 feature selection, can eliminate those background information, but cannot  
610 learn to extract more accurate person-related semantics for VI-ReID. Fig.  
611 6(d) and Fig. 6(e) show that, even without providing those person masks,  
612 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

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

#### 624 4.5.4. Number of parameters

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

## 634 5. Conclusion

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

## 650 6. Acknowledgment

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

## 655 References

- 656 [1] F.-P. An, J. e Liu, Pedestrian re-identification algorithm based on vi-  
657 sual attention-positive sample generation network deep learning model,  
658 Information Fusion 86-87 (2022) 136–145.
- 659 [2] J. Suutala, J. Rönning, Methods for person identification on a pressure-  
660 sensitive floor: Experiments with multiple classifiers and reject option,  
661 Information Fusion 9 (1) (2008) 21–40.
- 662 [3] N. Huang, J. Liu, Y. Miao, Q. Zhang, J. Han, Deep learning for visible-  
663 infrared cross-modality person re-identification: A comprehensive re-  
664 view, Information Fusion 91 (2023) 396–411.
- 665 [4] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. H. Hoi, Deep learning  
666 for person re-identification: A survey and outlook, IEEE Transactions  
667 on Pattern Analysis and Machine Intelligence 44 (6) (2022) 2872–2893.
- 668 [5] Z. Chang, S. Yang, Z. Feng, Q. Gao, S. Wang, Y. Cui, Semantic-relation  
669 transformer for visible and infrared fused image quality assessment, In-  
670 formation Fusion 95 (2023) 454–470.
- 671 [6] M. Qi, S. Wang, G. Huang, J. Jiang, J. Wu, C. Chen, Mask-guided dual  
672 attention-aware network for visible-infrared person re-identification,  
673 Multimedia Tools and Applications 80 (12) (2021) 17645–17666.
- 674 [7] Z. Zhao, B. Liu, Q. Chu, Y. Lu, N. Yu, Joint color-irrelevant consistency  
675 learning and identity-aware modality adaptation for visible-infrared  
676 cross modality person re-identification, Proceedings of the AAAI Con-  
677 ference 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.

- 701 [15] X. Yang, P. Zhou, M. Wang, Person reidentification via structural deep  
702 metric learning, *IEEE Transactions on Neural Networks and Learning*  
703 *Systems* 30 (10) (2019) 2987–2998.
- 704 [16] C. Chen, M. Ye, M. Qi, B. Du, Sketch transformer: Asymmetrical dis-  
705 entanglement learning from dynamic synthesis, in: *Proceedings of the*  
706 *ACM International Conference on Multimedia*, 2022, pp. 4012–4020.
- 707 [17] Z. Wei, X. Yang, N. Wang, B. Song, X. Gao, ABP: Adaptive body par-  
708 tition model for visible infrared person re-identification, in: *Proceedings*  
709 *of the IEEE International Conference on Multimedia and Expo*, 2020,  
710 pp. 1–6.
- 711 [18] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, C.-W. Lin, Structure-aware  
712 positional transformer for visible-infrared person re-identification, *IEEE*  
713 *Transactions on Image Processing* 31 (2022) 2352–2364.
- 714 [19] Y. Miao, N. Huang, X. Ma, Q. Zhang, J. Han, On exploring pose es-  
715 timation as an auxiliary learning task for visible–infrared person re-  
716 identification, *Neurocomputing* 556 (2023) 126652.
- 717 [20] J. Liu, J. Wang, N. Huang, Q. Zhang, J. Han, Revisiting  
718 modality-specific feature compensation for visible-infrared person re-  
719 identification, *IEEE Transactions on Circuits and Systems for Video*  
720 *Technology* 32 (10) (2022) 7226–7240.
- 721 [21] Q. Zhang, C. Lai, J. Liu, N. Huang, J. Han, Fmcnet: Feature-level  
722 modality compensation for visible-infrared person re-identification, in:

- 723        Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-  
724        tern Recognition, 2022, pp. 7349–7358.
- 725    [22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recog-  
726        nition, in: Proceedings of the IEEE Conference on Computer Vision and  
727        Pattern Recognition, 2015, pp. 770–778.
- 728    [23] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models:  
729        Person retrieval with refined part pooling and a strong convolutional  
730        baseline, in: Proceedings of the European Conference on Computer Vi-  
731        sion, 2018, pp. 501–518.
- 732    [24] Y. Zhu, Z. Yang, L.-C. Wang, S. Zhao, X. Hu, D. Tao, Hetero-center loss  
733        for cross-modality person re-identification, *Neurocomputing* 386 (2020)  
734        97–109.
- 735    [25] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J.-H. Lai, RGB-Infrared cross-  
736        modality person re-identification, in: Proceedings of the IEEE Interna-  
737        tional Conference on Computer Vision, 2017, pp. 5390–5399.
- 738    [26] D. T. Nguyen, H. G. Hong, K. W. Kim, K. R. Park, Person recognition  
739        system based on a combination of body images from visible light and  
740        thermal cameras, *Sensors* 17 (3) (2017) 605.
- 741    [27] M. Ye, X. Lan, Z. Wang, P. C. Yuen, Bi-directional center-constrained  
742        top-ranking for visible thermal person re-identification, *IEEE Transac-*  
743        *tions on Information Forensics and Security* 15 (2020) 407–419.
- 744    [28] Y. Hao, J. Li, N. Wang, X. Gao, Modality adversarial neural network for



- 745 visible-thermal person re-identification, *Pattern Recognition* 107 (2020)  
746 107533.
- 747 [29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan,  
748 T. Killeen, Z. Lin, N. Gimeshein, L. Antiga, A. Desmaison, A. Köpf,  
749 E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner,  
750 L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-  
751 performance deep learning library, in: *Proceedings of the Neural In-*  
752 *formation Processing Systems*, 2019, pp. 8026–8037.
- 753 [30] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feed-  
754 forward neural networks, in: *Proceedings of the International Confer-*  
755 *ence on Artificial Intelligence and Statistics*, 2010, pp. 249–256.
- 756 [31] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, Z. Hou, RGB-Infrared  
757 cross-modality person re-identification via joint pixel and feature align-  
758 ment, in: *Proceedings of the IEEE International Conference on Com-*  
759 *puter Vision*, 2019, pp. 3622–3631.
- 760 [32] M. Ye, J. Shen, L. Shao, Visible-infrared person re-identification via  
761 homogeneous augmented tri-modal learning, *IEEE Transactions on In-*  
762 *formation Forensics and Security* 16 (2020) 728–739.
- 763 [33] N. Pu, W. Chen, Y. Liu, E. M. Bakker, M. S. Lew, Dual gaussian-  
764 based variational subspace disentanglement for visible-infrared person  
765 re-identification, in: *Proceedings of the ACM International Conference*  
766 *on Multimedia*, 2020, pp. 2149–2158.

- 767 [34] H. Liu, Z. Miao, B. Yang, R. Ding, A base-derivative framework for  
768 cross-modality rgb-infrared person re-identification, in: Proceedings of  
769 the International Conference on Pattern Recognition, 2021, pp. 7640–  
770 7646.
- 771 [35] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia, C.-W. Lin, Grayscale  
772 enhancement colorization network for visible-infrared person re-  
773 identification, IEEE Transactions on Circuits and Systems for Video  
774 Technology (2021).
- 775 [36] H. Liu, X. Tan, X. Zhou, Parameter sharing exploration and hetero-  
776 center triplet loss for visible-thermal person re-identification, IEEE  
777 Transactions on Multimedia (2020).
- 778 [37] M. Ye, C. Chen, J. Shen, L. Shao, Dynamic tri-level relation mining with  
779 attentive graph for visible infrared re-identification, IEEE Transactions  
780 on Information Forensics and Security 17 (2022) 386–398.
- 781 [38] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, C.-W. Lin, Structure-aware  
782 positional transformer for visible-infrared person re-identification, IEEE  
783 Transactions on Image Processing 31 (2022) 2352–2364.
- 784 [39] Y. Huang, Q. Wu, J. Xu, Y. Zhong, P. Zhang, Z. Zhang, Alleviating  
785 modality bias training for infrared-visible person re-identification, IEEE  
786 Transactions on Multimedia (2021).
- 787 [40] M. Ye, Z. Wang, X. Lan, P. C. Yuen, Visible thermal person re-  
788 identification via dual-constrained top-ranking, in: Proceedings of the

- 789 International Joint Conference on Artificial Intelligence, 2018, pp. 1092–  
790 1099.
- 791 [41] Z. Feng, J. Lai, X. Xie, Learning modality-specific representations for  
792 visible-infrared person re-identification, *IEEE Transactions on Image*  
793 *Processing* 29 (2019) 579–590.
- 794 [42] S. Choi, S. Lee, Y. Kim, T. Kim, C. Kim, Hi-CMD: Hierarchical cross-  
795 modality disentanglement for visible-infrared person re-identification, in:  
796 *Proceedings of the IEEE Conference on Computer Vision and Pattern*  
797 *Recognition*, 2020, pp. 10257–10266.
- 798 [43] H. Liu, J. Cheng, W. Wang, Y. Su, H. Bai, Enhancing the discrim-  
799 inative feature learning for visible-thermal cross-modality person re-  
800 identification, *Neurocomputing* 398 (2020) 11–19.
- 801 [44] H. Ye, H. Liu, F. Meng, X. Li, Bi-directional exponential angular triplet  
802 loss for rgb-infrared person re-identification, *IEEE Transactions on Im-*  
803 *age Processing* 30 (2020) 1583–1595.
- 804 [45] Y. Yang, T. Zhang, J. Cheng, Z. Hou, P. Tiwari, H. M. Pandey, et al.,  
805 Cross-modality paired-images generation and augmentation for RGB-  
806 Infrared person re-identification, *Neural Networks* 128 (2020) 294–304.
- 807 [46] Y.-B. Zhao, J.-W. Lin, Q. Xuan, X. Xi, Hpiln: A feature learning frame-  
808 work for cross-modality person re-identification, *IET Image Processing*  
809 13 (14) (2019) 2897–2904.
- 810 [47] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, N. Yu, Cross-modality

- 811 person re-identification with shared-specific feature transfer, in: Pro-  
812 ceedings of the IEEE Conference on Computer Vision and Pattern  
813 Recognition, 2020, pp. 13379–13389.
- 814 [48] Z. Wei, X. Yang, N. Wang, X. Gao, Flexible body partition-based adver-  
815 sarial learning for visible infrared person re-identification, IEEE Trans-  
816 actions on Neural Networks and Learning Systems (2021).
- 817 [49] Y. Cheng, X. Li, G. Xiao, W. Ma, X. Gou, Dual-path deep supervision  
818 network with self-attention for visible-infrared person re-identification,  
819 in: Proceedings of the IEEE International Symposium on Circuits and  
820 Systems, 2021, pp. 1–5.
- 821 [50] L. van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of  
822 Machine Learning Research 9 (86) (2008) 2579–2605.