Graph convolution-based feature disentanglement for visible-infrared person re-identification

Ren Lou[a], Muyu Wang[*b], Yihao Shen[b], Sanyuan Zhao[b], Xinyuan Wang[a], Yueqi Zhou[a], Fangfang Li[c], Qiangqiang Xiang[a]
[a]Zhejiang Scientific Research Institute of Transport, Hangzhou, Zhejiang, China; [b]School of Computer Science, Beijing Institute of Technology, Beijing, China; [c]Enterprise Institute of Zhejiang Communications Investment Expressway Operation Management Co., Ltd, Hangzhou, Zhejiang, China

ABSTRACT

We propose a graph convolution-based disentanglement algorithm that is well-performed in the task of cross-modal person re-identification between visible and infrared images. Given the image of an individual in one modality, the problem to be addressed is whether the same person also appears in images from another modality. To tackle this issue, the main idea of our proposed method is to disentangle image features into modality-related and modality-invariant features, thereby alleviating feature discrepancies across different modal images. Unlike traditional disentanglement methods, our proposed graph convolution-based approach abandons the use of generative adversarial networks and employs attention mechanisms for initial disentanglement, followed by optimization of disentangled features using graph convolution. Comprehensive experimental results on the RegDB dataset and SYSU MM01 dataset demonstrate the superiority of our method in terms of effectiveness and efficiency.

Keywords: Visible-infrared person re-identification, disentanglement, graph convolution, cross-modal

1. INTRODUCTION

The objective of the visual-infrared cross-modal pedestrian re-identification task is to match pedestrians appearing in both modalities: RGB images captured by visible light cameras and infrared images captured by infrared cameras. In the task of visual-infrared cross-modal pedestrian re-identification, there exist not only modal intra-variations similar to traditional single-modal pedestrian re-identification tasks, such as issues related to low resolution, changes in viewpoint, and occlusion, but also more intricate cross-modal differences. The latter arises due to the inherent disparities between the reflectance in the visible spectrum and the emissivity in the thermal spectrum. The intertwining of modality-specific differential information, such as lighting and texture, with modality-agnostic differential information, such as posture and shape, poses significant challenges to pedestrian re-identification tasks.

Recent studies have explored the use of image-level constraints to disentangle inter-modal and intra-modal differences. Popular disentanglement methods employ generative adversarial networks to transform images from one modality to another, thereby eliminating modality differences. However, the process of transforming between two modalities with GAN is complex, demanding extensive experimentation to determine suitable parameters and substantial computational resources. Moreover, disentanglement methods relying on Generative Adversarial Networks (GANs) often yield poor results when the quality of the training datasets cannot be guaranteed to be high.

To address the aforementioned drawbacks of previous disentanglement methods, we propose a novel network architecture based on graph convolution for inter-modal and intra-modal feature disentanglement as shown in Figure 1. We discard the generator and discriminator structures typically employed in previous disentanglement approaches and directly utilize a single end-to-end network for disentanglement. We apply instance normalization to acquire fundamental modality-independent features. Subsequently, attention mechanisms are employed to combine the original features with features obtained through instance normalization for a coarse disentanglement. Afterward, graph convolution is applied to the coarse disentangled modality-specific and modality-agnostic features for further disentanglement. Our method is a feature-based disentanglement rather than image-based disentanglement like previous GAN-based methods, thus to a certain extent

*179787711@qq.com

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 1339543 · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3049214

addressing issues caused by poor image quality and limited samples in datasets.

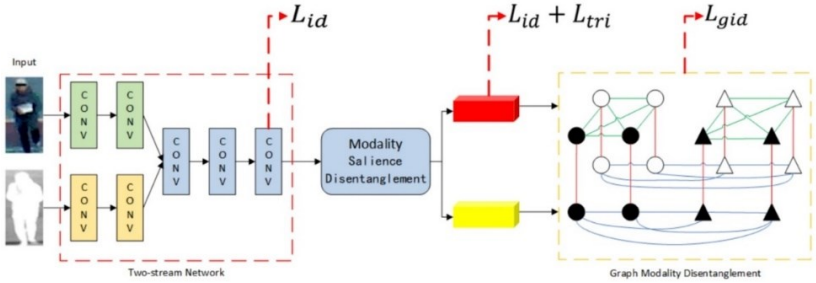

Figure 1. Network structure.

A two-stream network takes RGB images and infrared images as inputs and extracts the entangled features. The features are roughly disentangled in the Modality Salience Disentanglement, and then the Graph Modality Disentanglement is used for further refinement.

The main contributions of our work are summarized as follows:

- We propose a novel disentanglement method that is well performed in the task of cross-modal person reidentification. In comparison to previous disentanglement methods, the proposed method has a simpler network structure and a more efficient training process.
- To the best of our knowledge, the proposed graph-based disentanglement method is the first work to use graph convolution for feature disentanglement.
- Extensive experimental results demonstrate that the proposed method achieves or surpasses the state-of-the-art approaches.

2. APPROACH

2.1 Problem definition and network structure

2.2.1 Problem definition. For the visible image $X_v \in \mathbb{R}^{H\times W\times 3}$ and the infrared image $X_t \in \mathbb{R}^{H\times W\times 3}$, each corresponds to an identity label $\mathrm{l} \in \{1, 2, \ldots, \mathrm{N}\}$, where N is the number of pedestrians. During the training phase, we train a dual-stream input feature extraction network $\phi(\cdot)$ on a cross-modal image dataset. During the testing phase, given a query image from one modality, we calculate the features of all images in the query set from the other modality. We then compute the Euclidean distances for the corresponding features and use these distances as the sorting key. After passing through the feature extraction network $\phi(\cdot)$, we obtain an output feature denoted as X. X can be regarded as a combination of two distinct features entangled together: one representing the modality-independent feature x_p and the other representing the modality-specific feature x_m. Modality-independent features encompass information such as the form and posture of individuals in the image, which is independent of the imaging modality. Modality-specific features, on the other hand, arise due to the different imaging principles of the two modalities

2.2.2 Network structure. The network we proposed is designed to disentangle modality-independent features from modality-specific features, ensuring that the process of pedestrian re-identification is not influenced by modality differences. As shown in Figure 1, the first part is the dual-stream feature extraction network to extract entangled features. The second part is the Modality Salience Disentanglement, which disentangles the features extracted and outputs coarse modality-specific features and modality-independent features. The third part is Graph Convolution Disentanglement, which is used for further fine disentanglement. For modality-independent features (orbiculars), as shown in Figure 1, graphs are established that are relevant to identity categories (white) and modality-agnostic aspects (black) for fine-tuning features. For modality-specific features (triangles), graphs are established that are relevant to modality aspects (white) and independent of identity categories (black) for feature fine-tuning. Additionally, to facilitate information propagation and aggregation, connections are established between modality-specific and modality-agnostic features through graphs.

2.2 Modality salience disentanglement

The Modality Salience Disentanglement module is the first step in modality disentanglement and is dedicated to separating the features extracted by the dual-stream network into modality-agnostic features and modality-specific features. The architecture of the Modality Salience Disentanglement is shown in Figure 2. Due to the significant advantages of instance normalization in cross-modal applications and reducing sample diversity, we first apply instance normalization to the features X extracted by the dual-stream network and achieve the normalized feature X':

$$X_k' = \frac{X_k - E[X_k]}{\sqrt{Var[Z_k] + \varepsilon}} \tag{1}$$

where k denotes the k-th dimension of the corresponding feature. The addition of ε is to avoid division by 0, and the mean $E[\cdot]$ and standard deviation $Var[\cdot]$ need to be calculated dimension-wise. Next, we extract channel-wise attention information. We apply average pooling and max pooling to reduce the dimension of X, obtaining two distinct low-dimensional features, denoted as x_1 and x_2, respectively. Then, a multi-layer perceptron (MLP) is employed to fuse these features obtained from the two pooling methods, resulting in channel attention a:

$$x_1 = avgpool(X), x_2 = max\,pool(X) \tag{2}$$

$$a = \sigma(W_2(\delta(W_{11}x_1) + \delta(W_{12}x_2))) \tag{3}$$

where σ represents the sigmoid activation function and δ represents the ReLU activation function, W is weights in MLP. We remove the residual module X' in X to obtain an intermediate feature R, i.e. $R = X - X'$. In this way, we achieve the modality-independent feature X^+ as:

$$X^+ = X' + a \cdot R = a \cdot X + (1-a) \cdot X' \tag{4}$$

Similarly, the modality-specific feature X^- is:

$$X^- = X' + (1-a) \cdot R = a \cdot X' + (1-a) \cdot X \tag{5}$$

With an average pooling, we obtain the output modality-independent feature x_p and modality-related feature x_m:

$$x_p = avgpool(X^+), x_m = avgpool(X^-) \tag{6}$$

In the Modality Salience Disentanglement, we employ identity loss L_{id} and triplet loss L_{tri} for supervision. Fundamentally, the mechanism applied is a self-attention mechanism, lacking a reliable supervised method for effective feature dis-entanglement. Therefore, in the subsequent Graph Convolution Disentanglement module, the disentanglement is further optimized.

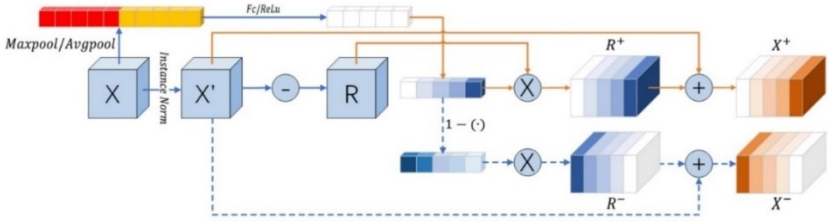

Figure 2. Diagram of the modality saliency disentanglement.

2.3 Graph convolution disentanglement

During training, we sample P identity categories. For each identity category K visible images and K infrared images are randomly selected, and there will be a total of $2PK$ images. After the modality salience disentanglement, we obtain the modality-specific feature $x_p \in \mathbb{R}^{2PK \times C}$ and the modality-agnostic features $x_m \in \mathbb{R}^{2PK \times C}$. Considering feature x_p and x_m as vertices, there will be $4PK$ vertices. Vertices are divided into two classes based on modality specificity, denoted as V^+ and V^-:

$$\begin{cases} V^- = \{X_{\mathrm{m}}^l\}, i \in \{1,...,2PK\} \\ V^+ = \{X_p^i\}, i \in \{1,...,2PK\} \end{cases} \tag{7}$$

We construct a graph with these vertices. The edges E_{ij} between V^+ and V^- in the graph can be classified into three categories: *modality-independent edges* E_{ij}^+ (both vertices at the ends are modality-independent vertices); *modality-specific edges* E_{ij}^- (both vertices at the ends are modality-specific vertices); and *disentanglement edges* E_{ij}^D (connecting a modality-specific vertex and a modality-agnostic vertex).

The graph convolution disentanglement part shown on the right side of Figure 1 visually illustrates the various connections between vertices and edges. Different shapes represent different identity categories, and different colors represent vertices from different modalities. Taking the graph in Figure 1 as an example, there are two identities, i.e. *P*=2 and for each identity, *K*=2 images for both modalities are selected. For each identity, the upper four (two black and two white) vertices are modality-independent vertices and the four vertices at the bottom are modality-specific vertices. The corresponding modality-specific and modality-independent vertices are connected vertically. In the upper part, vertices with the same shape (i.e., from the same identity category) are connected, while in the lower part, vertices with the same color (i.e., from the same modality) are connected.

We use Graph Attention Networks (GAT) for graph convolution. The graph convolution network is denoted as $\phi(\cdot)$, in this way the convolved feature $(x^g, x^g) = \phi(x_p, x_m)$. When performing node classification, for x_p^g vertices, classification can be directly done based on the identity category. For x_m^g vertices, they can be divided into two classes: visible background class and infrared background class. In this stage, since the number of samples in the visible background class and infrared background classes is much larger than the number of samples in the identity categories, the focal loss is used to calculate the loss function:

$$L_{gid} = \begin{cases} -\alpha(1-y')^{\gamma}\log(y'), y \in S_P \\ -(1-\alpha)y'^{\gamma}\log(1-y'), y \in S_N \end{cases} \tag{8}$$

where *a* adjusts the balance between positive and negative samples, and γ adjusts the weight between easy and hard samples. S_P represents the positive samples and S_n represents the negative samples.

3. EXPERIMENTS

3.1 Datasets and metrics

In the visual-infrared pedestrian re-identification task, commonly used public datasets include SYSU-MM01[1] and RegDB[2] datasets. The images in the SYSU-MM01 dataset are collected by 4 visible cameras and 2 infrared cameras on the campus of Sun Yat-sen University. These cameras capture both indoor and outdoor scenes. The training set of SYSU-MM01 contains 22,258 visible images and 11,909 infrared images, featuring 395 pedestrians. The test set includes 96 individuals. The RegDB dataset consists of images captured by one visible camera and one infrared camera. The dataset includes images of 412 pedestrians, with each individual having 10 visible and 10 infrared images. statistically stable results and the average is recorded.

For the RegDB dataset, we conduct experiments under two settings: the experiments, where known infrared modality images are used to search for visible modality images, are referred to as experiments under the infrared-visual setting; the experiments where known visible modality images are used to search for infrared modality images are referred to as experiments under the visual-infrared setting. For the SYSU-MM01 dataset, we conduct experiments under a global-single-shot setting, which means performing pedestrian recognition with a given picture of arbitrary modality.

To evaluate the performance of the pedestrian re-identification system, we use metrics including Cumulative Matching Characteristics (CMC) and Mean Average Precision (mAP), as well as the newly proposed evaluation method Mean Inverse Negative Penalty (mINP).

3.2 Experiment settings

In the preprocessing stage of the images, all images are first resized to the resolution of 288×144. The backbone network, ResNet-50, is pre-trained on the ImageNet dataset, and the stride of the last residual block is changed from 2 to 1. For data augmentation, horizontal flipping and random erasing are applied during preprocessing. In the RegDB dataset, is set to 8, K is set to 4, and each batch contains 64 images. In the SYSU MM01 dataset, P is set to 6, K is set to 8, and each batch contains 96 images. We use an SGD optimizer with a momentum parameter set to 0.9, and the initial learning rate is set to 0.1. A warm-up learning strategy is applied during the training process

3.3 Ablation study

To demonstrate the effectiveness of each module of our method, we conduct ablation experiments on the RegDB dataset. Firstly, we introduce the baseline network of the disentanglement network. The baseline network, as well as the feature extraction network of the proposed disentanglement network, is based on the ResNet50 architecture. In the disentanglement part, we use a naive channel attention technique for feature disentanglement. We let X be the features extracted by the dual-stream network, and the learned channel attention be α. The disentanglement of features is calculated as follows:

$$X_p = a \cdot X, X_m = (1-a) \cdot X \tag{9}$$

The same loss function as the modality salience disentanglement module is used. The first experiment of the ablation study is set to use only the baseline network (baseline), the second experiment is set to use the modality salience disentanglement without the graph convolution disentanglement module (w/MSD), and the third experiment is set to include all the proposed modules in this paper (total). The results of the ablation experiments under the setting of visual-infrared and infrared-visual are respectively shown in Tables 1 and 2.

Table 1. Results of the ablation experiments under the visual-infrared setting on the RegDB dataset.

Methods	**r=1**	**r=10**	**r=20**	**mAP**	**mINP**
Baseline	88.47	95.81	97.50	86.36	79.50
w/MSD	89.49	96.22	97.83	87.67	81.57
Total	90.64	96.79	98.21	88.41	81.91

Table 2. Results of the full ablation experiments under the infrared-visual setting on the RegDB dataset

Methods	**r=1**	**r=10**	**r=20**	**mAP**	**mINP**
Baseline	88.56	95.67	97.45	86.50	79.50
w/ MSD	89.26	95.68	97.33	86.92	79.68
Total	89.69	96.67	98.32	87.86	81.09

The experimental results on the RegDB dataset show that compared to the baseline method, the Modality Salience Disentanglement and Graph Convolution Disentanglement we proposed both improve the results of the cross-modal pedestrian re-identification task in the two experimental settings. Here, we present retrieval results under different settings of our method. Figure 3 shows the retrieval results under the infrared-visual setting. Figure 4 illustrates the retrieval results under the visual-infrared setting. In these figures, the first column shows the images to be queried, and the other columns show the correct query results. The green rectangles display the correct retrieval results, while the red rectangles show the incorrect ones.

Figure 3. Retrieval result illustrations for infrared-visible setting.

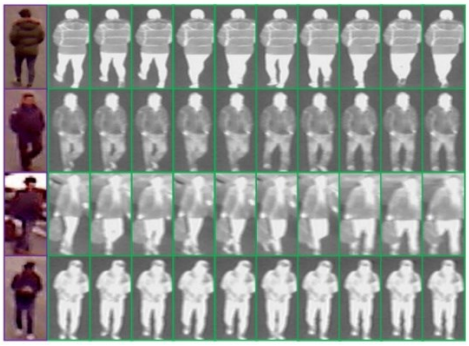

Figure 4. Retrieval result illustrations for visible-infrared setting.

3.4 Comparison with other methods

We compare our method with the state-of-the-art methods on the RegDB dataset and SYSU MM01 dataset. The experimental results on the RegDB dataset of w are presented in Tables 3 and 4. The experiments on the SYSU MM01 dataset under the Golbal-Single-Shot experimental condition are shown in Table 5.

Experiments on the RegDB dataset (Tables 3 and 4) demonstrate that our method achieves the best performance in both query settings. For the visual-infrared query setting, our method achieves a performance of CMC (rank1)/mAP/mINP at 90.64\%/88.41\%/81.91\%. For the infrared-visual query setting, it achieves CMC (rank1)/mAP/mINP at 89.69\%/87.86\%/81.09\%. These results prove that our method effectively extracts modality-independent features from visible and infrared images, avoiding interference from different modalities in feature extraction. Through the graph convolution disentanglement module, it extracts features related to identity categories, significantly improving the performance of visual-infrared cross-modal pedestrian re-identification tasks.

Table 3. Comparisons of our method with state-of-the-art techniques under the visual-infrared setting on the RegDB dataset.

Methods	r=1	r=10	r=20	mAP	mINP
MAC[3]	36.43	62.36	71.63	37.03	-
AliGAN[4]	57.90	-	-	53.60	-
eBDTR[5]	34.62	58.96	68.72	33.46	-
EDFL[6]	52.58	72.10	81.47	52.98	-
expAT[7]	66.48	-	-	67.31	-
MPMN[8]	86.56	96.68	98.28	82.91	-
CMAlign[9]	74.17	-	-	67.64	-
NFS[10]	80.54	91.96	95.07	72.10	-
MPANet[11]	82.8	-	-	80.7	-
Variational distillation[12]	73.2	-	-	71.6	-
Ours (global feature)	90.64	96.79	98.21	88.41	81.91

Table 4. Comparisons of our method with state-of-the-art techniques under the infrared-visual setting on the RegDB dataset.

Methods	r=1	r=10	r=20	mAP	mINP
MAC[3]	36.20	61.68	70.99	39.23	-
AliGAN[4]	56.30	-	-	53.40	-
eBDTR[5]	34.21	58.74	68.64	32.49	-

Methods	r=1	r=10	r=20	mAP	mINP
EDFL[6]	51.89	72.09	81.04	52.13	-
expAT[7]	67.45	-	-	66.51	-
MPMN[8]	84.62	95.51	97.33	79.49	-
CMAlign[9]	72.43	-	-	65.46	-
NFS[10]	77.95	90.45	93.62	69.79	-
MPANet[11]	83.7	-	-	80.9	-
Variational distillation[12]	71.8	-	-	70.1	-
Ours (global feature)	89.69	96.67	98.32	87.86	81.09

On the SYSU-MM01 dataset, our experimental results are shown in Table 5. Compared to current SOTA methods, our method outperforms most other methods and achieves similar performance and achieves CMC(rank1)/mAP/mINP at 60.98\%/57.81\%/43.65\%, proving the superiority of graph convolutional-based method for disentanglement. It's worth noting that, compared to GAN-based disentanglement methods, our proposed disentanglement method does not use generative adversarial networks for generating fake images, greatly reducing the complexity of the network and the difficulty of the training process. Comparative results also demonstrate that disentanglement in the feature space is more efficient and effective than disentanglement at the image scale.

Table 5. Comparisons of our method with state-of-the-art techniques on the SYSU-MM01 dataset.

Methods	r=1	r=10	r=20	mAP	mINP
HOG[21]	2.76	18.30	31.90	4.24	-
AliGAN[4]	42.40	85.00	93.70	40.70	-
eBDTR[5]	27.82	67.34	81.34	28.42	-
EDFL[6]	36.94	85.42	93.22	40.77	-
expAT[7]	38.57	76.64	86.39	38.61	-
XIV[13]	49.92	89.79	95.96	50.73	-
MSR[14]	37.35	83.40	93.34	38.11	-
JSIA[15]	38.10	80.70	89.90	36.90	-
CMSP[16]	43.56	86.25	-	44.98	-
Attri[17]	47.14	87.93	94.45	47.08	-
HAT[18]	55.29	92.14	97.36	53.89	-
HC[19]	56.96	91.50	96.82	54.95	-
Hi-CMD[20]	34.94	77.58	-	35.94	-
CMAlign[9]	55.41	-	-	54.14	-
NFS[10]	56.91	91.34	96.52	55.45	-
MPANet[11]	70.58	96.21	98.80	68.24	-
Variational distillation[12]	60.02	94.18	98.14	58.80	-
Ours	60.98	91.38	96.19	57.81	43.65

4. CONCLUSION

Deviating from the traditional network architecture involving Generative Adversarial Networks (GANs) in disentanglement networks, we propose a novel disentanglement network based on graph convolution. The proposed approach employs attention mechanisms for the preliminary disentanglement of modality-independent and modality-dependent features. Subsequently, a new graph structure is designed to refine feature disentanglement through graph convolution. Compared to other disentanglement methods, the network of the proposed approach is simpler yet more effective. Extensive experimental results demonstrate that the proposed method achieves or surpasses the current state-of-the-art methods.

ACKNOWLEDGEMENTS

This work was supported by a public welfare project of the Zhejiang Provincial Department of Science and Technology named Research on Vehicle Trajectory Data Quality Evaluation Technology based on radar-vision integrated equipment (No. LGC22E080003) and supported by a Science and Technology Planning Project of the Zhejiang Provincial Department of Transportation named Research on evaluation technology of traffic flow lighting vision fusion sensing system (No. 202209).

REFERENCES

[1] Wu, A., Zheng, W. S., Yu, H. X., Gong, S. and Lai, J., "RGB-infrared cross-modality person reidentification," in Proceedings of the IEEE International Conference on Computer Vision, 5380-5389 (2017).
[2] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L. and Hoi, S. C., "Deep learning for person re-identification: A survey and outlook," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2872-2893 (2021).
[3] Ye, M., Lan, X. and Leng, Q., "Modality-aware collaborative learning for visible thermal person reidentification," in Proceedings of the 27th ACM International Conference on Multimedia, 347-355 (2019).
[4] Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y. and Hou, Z., "RGB-infrared cross-modality person reidentification via joint pixel and feature alignment," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 3623-3632 (2019).
[5] Ye, M., Lan, X., Wang, Z. and Yuen, P. C., "Bi-directional center- constrained top-ranking for visible thermal person re-identification," IEEE Transactions on Information Forensics and Security, 15, 407-419 (2019).
[6] Liu, H., Cheng, J., Wang, W., Su, Y. and Bai, H., "Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification," Neurocomputing, 398, 11-19 (2020).
[7] Ye, H., Liu, H., Meng, F. and Li, X., "Bi-directional exponential angular triplet loss for RGB-infrared person re-identification," IEEE Transactions on Image Processing, 30, 1583-1595 (2020).
[8] Wang, P., Zhao, Z., Su, F., Zhao, Y., Wang, H., Yang, L. and Li, Y., "Deep multi-patch matching network for visible thermal person re-identification," IEEE Transactions on Multimedia, 23, 1474-1488 (2020).
[9] Park, H., Lee, S., Lee, J. and Ham, B., "Learning by aligning: Visible- infrared person re-identification using cross-modal correspondences," Proceedings of the IEEE/CVF International Conference on Computer Vision, 12046-12055 (2021).
[10] Chen, Y., Wan, L., Li, Z., Jing, Q. and Sun, Z., "Neural feature search for RGB-infrared person re-identification," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 587- 597 (2021).
[11] Wu, Q., Dai, P., Chen, J. C., Lin, W., Wu, Y., Huang, F., Zhong, B. and Ji R., "Discover cross-modality nuances for visible-infrared person re-identification," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4330-4339 (2021).
[12] Tian, X., Zhang, Z., Lin, S., Qu, Y., Xie, Y. and Ma L., "Farewell to mutual information: Variational distillation for cross-modal person re-identification," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1522-1531 (2021).
[13] Li, D., Wei, X., Hong, X. and Gong, Y., "Infrared-visible cross-modal person re-identification with an x modality," in Proceedings of the AAAI conference on artificial intelligence, 4610-4617 (2020).
[14] Feng, Z., Lai, J. and Xie, X., "Learning modality-specific representations for visible-infrared person re-identification," IEEE Transactions on Image Processing, 29, 579-590 (2019).

[15] Wang, G. A., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X. and Hou, Z. G., "Cross-modality paired-images generation for RGB-infrared person re-identification," Proceedings of the AAAI Conference on Artificial Intelligence, 12144-12151 (2020).
[16] Wu, A., Zheng, W. S., Gong, S. and Lai, J., "RGB-ir person reidentification by cross-modality similarity preservation," International Journal of Computer Vision, 128, 1765-1785 (2020).
[17] Zhang, S., Chen, C., Song, W. and Gan, Z., "Deep feature learning with attributes for cross-modality person re-identification," Journal of Electronic Imaging, 29, 033017-033017 (2020).
[18] Ye, M., Shen, J. and Shao, L., "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," IEEE Transactions on Information Forensics and Security, 16, 728-739 (2020).
[19] Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X. and Tao, D., "Hetero-center loss for cross-modality person re-identification," Neurocomputing, 386, 97-109 (2020).
[20] Choi, S., Lee, S., Kim, Y., Kim, T. and Kim, C., "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10257-10266 (2020).
[21] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 886-893 (2005).