ABSTRACT  Visible-Infrared person re-identification (VI Re-ID) is a challenging but practical task that aims at matching pedestrian images between the visible (daytime) modality and the infrared (nighttime) modality, playing an important role in criminal investigation and intelligent video surveillance applications. Numerous previous studies focused on alleviating the modality discrepancy and obtaining discriminating features by devising complex networks for VI Re-ID, but a cumbersome network structure is not suitable for practical industrial applications. In this paper, we propose a novel fusion method of global and part features to extract distinguishing features and alleviate cross-modality differences, named Global and Part Feature Fusion network (GPFF), which has not been well studied in the current literature. Specifically, we first adopt a dual-stream ResNet50 as a backbone network to alleviate the modality discrepancy. Then, we explore how to fuse global and local features to obtain discriminative features. Finally, we apply a heterogeneous center triplet loss (hetero-center triplet loss) instead of traditional triplet loss to guide sample center learning. Our proposed approach is simple but effective, and can remarkably boost the performance of VI Re-ID. The results of experiments on two public datasets (SYSU-MM01 and RegDB) demonstrate that our approach is superior to the state-of-the-art methods. Through experiments, we find that the effective fusion of global features and local features plays an important role in extracting discriminative features.

INDEX TERMS  Visible-infrared, Re-ID, feature fusion, cross-modality.

I. INTRODUCTION
Person re-identification is the work of searching images of a specific person that appears in a query set from a large gallery set. Many related studies have focused on metric learning [1], [2] and feature learning [3], [4] on the visible modal. Recently, these approaches achieved high performance [5]. However, most methods consider people’s images taken by visible spectrum camcorders during the daytime and are therefore not applicable to nighttime applications. To solve this problem, Alternative types of vision sensors, such as thermal camcorders, are now widely employed to supplement the data from visible cameras. Consider the example of looking for a person who ran away from home during the night. In this case, the query images could be captured by a visible camera during the daytime, and gallery images could be obtained from the infrared camera at night, and vice versa, as shown in Figure 1. However, matching visible-spectrum images to infrared images is a significant challenge [6].

There are two main problems with the visible-infrared cross-modality person identification (VI Re-ID) task: (1) The different reflectance spectra and induced emissivity of the visible camera and the infrared camera produce significant cross-modality differences, and (2) Similar to the visible-visible (VV) Re-ID task, there is a huge intra-modality variation due to factors such as viewpoint changes and different human postures [7], [48]. To mitigate the additional cross-modal variation in VI Re-ID, an obvious and intuitive approach is to take the cross-modal images into a shared feature space to compute the similarity metric. Consequently, we employ a dual-stream framework consisting of two modal-specific blocks with separate parameters for extracting features and a parameter-shared network for embedding features, which projects modal-specific features...
output by the dual-stream backbone network. Subsequently, we train every part-level feature and global feature with identification loss and hetero-center triplet loss, respectively. Then, we obtain discriminative features by adding global features to part-aggregated features. Moreover, we experimentally study the effect of changing the location of global features in the final feature (the fusion feature of global features and part-level features). To our knowledge, it is the first time to use the method for VI Re-ID. We apply hetero-center triplet loss and identify loss to train our network. Our method is efficient because we obtain rich and discriminative features by concatenating part-level features with global features in the channel dimension.

The main achievements of this research can be split into the following points.

- We propose GPFF to enhance feature representation through general operation and fusing the global feature and part-level features.
- We apply hetero-center triplet loss instead of traditional triplet loss to guide our network to learn modality-invariant features.
- We achieve superior performance on two authoritative datasets (SYSU-MM01 and RegDB), which can enhance high-quality research in the future.

II. RELATED WORK

This subsection reviews the current VI Re-ID methods. Compared with the conventional VV Re-ID, VI Re-ID should also process additional cross-modal discrepancies, except for intra-modal variations. To address the challenge, many scholars have focused on transforming heterogeneous cross-modal person images into shared spaces for similarity measurement. Previous works include the following three main components: network design, metric learning, and image translation.

A. NETWORK DESIGNING

Network design is an important step in VI Re-ID. Many researchers have focused on network design to alleviate modal differences between visible and infrared images. Ye et al. [15], [16], and [17] proposed to employ dual-stream networks to extract two different modality features separately. Subsequently, feature embedding was performed to map these features into a joint feature space, whose parameters are shared by the fully connected layer. Based on the above study, Liu et al. [18] incorporate intermediate features to strengthen the features of modal sharing with greater distinguishability. Furthermore, Liu et al. [7] explored the parameter-shared issue in a dual-stream network and analyzed the effect of the number of shared parameters on cross-modal feature learning. Zhang et al. [19] introduced a two-path inter-modal framework to obtain discriminative features, consisting of a two-path space structure retaining a common spatial network and the comparative relevant network that retains the inherent space structure and takes into account the variance of the input inter-modal image pairs. In order to investigate the potentiality of modal-sharing information and modal-specific

FIGURE 1. Task of VI Re-ID. Visible → Infrared means matching a specific person in a query set collected by a visible camera in the daytime to multiple persons images in a large gallery set provided by infrared cameras during the nighttime, and Infrared → Visible is the opposite. The green box denotes the positive sample, and the red box indicates the negative sample.
features in improving the performance of re-identification, Lu et al. [20] presented to model the affinity of different modality samples based on shared features. Subsequently, both shared and specific features were transferred across modalities. Following these methods, We apply a two-stream network to extract modality-specific and modality-shared features, respectively. Differently, we design the global and local feature fusion module from a completely new perspective.

B. METRIC LEARNING
Metric learning is a key stage of similarity measurement for Re-ID. In the deep neural network, Re-ID achieved excellent performance only using the Euclidean distance metric [21]. As a result, metric learning is intrinsic to the DNN training loss function, which guides the training procedure and makes the extracted features more robust and discriminative. Zhu et al. [22] presented the heterogeneous center loss to decrease the variations of intra-class and inter-modal. Based on [22], Liu et al. [7] also proposed the heterogeneous center triplet loss to constrict the distance between different class centers both from a single and multiple modalities. Hao et al. [23] introduced an end-to-end dual-stream hypersphere streaming embedding network with classification and recognition losses that constrains the variation within and across modalities to this hypersphere. Wu et al. [24] formulated the shared knowledge learning issue of cross-modality matching as a cross-modality similarity preserving problem. They propose a similarity preserving loss for focal modality perception to exploit intra-modal similarity and guide similarity learning between modalities. Zhao et al. [25] proposed a hard pentaplet loss to have better performance in cross-modal re-identification. Inspired by [7], we introduce the hetero-center triplet loss to guide our network to learn modality-invariant features.

C. IMAGE TRANSLATION
The aforementioned studies resolve both inter-modal and intra-modal variation at the feature extraction level. In recent years, the approaches of image generation on the basis of generative adversarial networks (GANs) have attracted widespread attention in VI Re-ID, it reduces the domain discrepancy between two different modalities at the level of the image. Kniaz et al. [44] first proposed to use GAN to convert visible images into a multi-modal infrared image set with Re-ID in the infrared domain. Wang et al. [45] introduced a two-level disparity mitigation learning framework that is based on bidirectional recurrent GAN for reducing domain discrepancy from two levels: feature level and image level. Choi et al. [46] presented a Stratified cross-modal disentanglement (Hi-CMD) method, which automatically separates ID-discriminating as well as ID-excluding factors from both visible and infrared images.

However, one infrared image may correspond to multiple visible images of a person during image generation. For example, people can have clothes of different colors in the visible modality that look identical in the thermal modality. When generating images, it is difficult to find out which one is the right target, because the model does not have access to the images of the gallery that appear only in the reasoning phase. Approaches based on image generation have performance instability, high training requirements, and high-level complexity.

Different from the above methods, we design a simple but effective network through simple operation and fusing the global feature and part-level features for VI Re-ID.

III. METHODOLOGY
In this chapter, we detail the framework of our presented feature learning method for VI Re-ID as shown in Figure 2. The framework mainly is composed of three components: A. the backbone network; B. GPFF; C. loss function.

A. BACKBONE NETWORK
Based on the results of [7], we empirically set the dual-stream network as our backbone to process images across modalities, as shown in Figure 2. Following [26], [27], the ResNet50 model [28] is adopted to construct the backbone network. The three remaining res-convolution modules (blocks2, 3 and 4) are configured as modal-sharing sub-modules with shared parameters to learn multi-modal shared middle layer feature representations in a shared 3D feature space [8].

B. GLOBAL AND PART FEATURE FUSION MODULE (GPFF)
In VV Re-ID, advanced performance is realized by combining the full-body representation and local features [9], [29], [30]. The global feature map is output from the backbone network. For the part-level feature map, a simple and typical approach is splitting the body image into horizontal strips, refer to the part-level feature extraction methods in [9] and [22]. We also use a uniform partitioning technique to get rough part body features.

Our proposed GPFF includes two aspects: global and part-level feature learning and one global and several part-level feature vectors fusion.

1) GLOBAL AND PART-LEVEL FEATURE LEARNING
Given an image of a person (visible or thermal), the 3D feature map is obtained after going through the dual-stream backbone. On the basis of 3D feature maps, as shown in figure 2, the following three steps are used to extract the global feature vector and the part-level feature vectors.

a) On the one hand, the 3D feature map is directly used as a global feature map without any operations. On the other hand, they are evenly split into p strips in the horizontal direction to generate rough body part feature maps, as illustrated in Figure 2, where p is a hyper-parameter.

b) We employ a pooling layer named generalized-mean (GeM) to transform the both 3D global feature map and all 3D part-level feature maps into 1D feature vectors rather than using traditional adaptive maximum pooling and adaptive average pooling. Given a 3D feature map $X \in \mathbb{R}^{C \times H \times W}$, the
Two-stream backbone network

Visible

stage0    stage1

Modality-specific

stage2    stage3    stage4

Modality-shared

Feature maps

GPFF

Global feature

1x1Conv

Loss

L_{id} + \beta L_{hc}

L_{id} + \beta L_{hc}

L_{id} + \beta L_{hc}

L_{id} + \beta L_{hc}

L_{hc}

FIGURE 2. Framework of our proposed Global and Local Feature Fusion method for VI Re-ID, which includes three major components: dual-stream backbone network, GPFF, and loss function. The dual-stream backbone network contains two modal-specific sub-modules with separate parameters and one modal-shared sub-module with shared parameters. Subsequently, on the one hand, the 3D feature map is directly used as a global feature map without any operations. On the other hand, 3D feature maps are divided into \( p \) 3D tensors along horizontal direction, where \( p = 4 \). The global feature map and the \( p \) 3D tensors are pooled into vectors by the GeM pooling operation. For the global vector and each part-level vector, a 1x1 convolutional block decreases the feature dimension. Next, the global feature vector and all part-level feature vectors are concatenated (cat) in the channel dimension to obtain the output features. Finally, each global and part-level feature vector is supervised by both identification loss \( L_{id} \) and the hetero-center triplet loss \( L_{hc} \), respectively.

GeM can be expressed as

\[
\hat{x} = \left( \frac{1}{|X|} \sum_{x_i \in X} x_i^\pi \right)^\frac{1}{\pi}
\]  

where \( \hat{x} \in \mathbb{R}^{C \times 1 \times 1} \) is the outcome of pooling, \( |\cdot| \) represents the number of elements, and \( n \) is the pooled hyper-parameter.

c) After step b), 1x1 convolutional blocks are employed to decrease the dimension of the global feature vector and part-level feature vectors. These 1x1 convolutional blocks include a 1x1 convolutional layer; their output channel number is set to \( \text{dim} \), where \( \text{dim} \) is a hyperparameter that can be preset. This layer is followed by a batch normalization layer and a ReLU layer.

2) ONE GLOBAL AND SEVERAL PART-LEVEL FEATURE VECTORS FUSION

One global feature vector and \( p \) part-level feature vectors are concatenated (cat) in the channel dimension to form the final person feature for the similarity measure during testing. Therefore, the key point is the determination of a method to aggregate one global feature vector and \( p \) part-level feature vectors. The body structure is an inherent characteristic of a person that does not depend on the modality. Thus, the aggregation order of \( p \) part-level feature vectors is unique, but the index of the global feature vector in the final feature sequence can be changed, as shown in Figure 2.

For simplicity, \( g \) denotes the global feature vector, \( p0, p1, p2, \) and \( p3 \) represent \( p \) part-level feature vectors, where \( p = 4 \). Moreover, \( f_i \) is the \( i \)th final feature obtained by combining the global feature and \( p \) part-level feature vectors, where \( i \in \{0, 1, 2, 3, 4\} \) is the index of the global feature vector in the final feature sequence.

C. LOSS FUNCTION

After obtaining the features from our proposed model, we use ID loss and hetero-center triplet loss to guide the model to learn. The ID loss is applied on classifier \( C(\cdot) \) to predict the identities:

\[
L_{id} = -\frac{1}{N} \sum_{i=1}^{N} \log P(y_i | C(f_i | \theta))
\]  

where \( N \) denotes the total number of person images in a mini-batch, \( C(f_i | \theta) \) is the identity predictions for both visible and infrared image features \( f_i \) with the same parameters \( \theta \).
The heterogeneous center triplet loss focuses on only one cross-modal positive center pair and the hardest (intra-modal and inter-modal) negative center pair [22]. The hetero-center triplet loss function is formulated as

$$L_{hc}(f) = \sum_{i=1}^{P} \left[ mc + \| fc_i^v - fc_i^m \|_2 - \min_{m \neq i} \| fc_i^v - fc_m^v \|_2 \right] + \sum_{i=1}^{P} \left[ mc + \| fc_i^f - fc_i^m \|_2 - \min_{m \neq i} \| fc_i^f - fc_m^f \|_2 \right]$$

where $mc$ is the margin; $fc_i^v = \frac{1}{K} \sum_{j=1}^{K} f_{i,j}^v$ and $fc_i^f = \frac{1}{K} \sum_{j=1}^{K} f_{i,j}^f$ are the visible and infrared centers of identity, respectively. Moreover, $f_{i,j}^v$ and $f_{i,j}^f$ respectively indicate the $j^{th}$ visible and infrared image features of $i^{th}$ identity.

The total loss $L$ of GPFF is calculated as

$$L = L_{id} + \beta L_{hc}$$

where $\beta$ is trade-off parameter to balance the influence of each loss function.

### IV. EXPERIMENT

In this chapter, we verify the effectivity of our proposed methods for VI Re-ID tasks on the SYSU-MM01 [32] and RegDB [31] public datasets.

#### A. EXPERIMENT SETTING

1) DATASETS

SYSU-MM01 [32] is a large-scale and widely used dataset, which was captured in the SYSU campus by six camcorders, including four visible and two infrared camcorders deployed in indoor and outdoor environments. The training set contains 395 person identities, including 22,258 visible images and 11,909 infrared images. Moreover, the testing set contains 96 additional person identities, consisting of 3,803 thermal images for querying and 301 randomly selected visible images as a gallery set. The gallery set includes all visible images taken by the four visible camcorder in full search mode. While, in the indoor-search mode, the gallery set includes solely visible pictures taken by two visible camcorder installed in the room. We followed the current approaches to conduct 10 trials of gallery set selection in the context of a single-shot [33], [34] and recorded the mean values.

RegDB [31] is another widely used VI Re-ID dataset, which was built using two-camcorder systems (one visible and one infrared camcorder) and contained 412 person identities. For every identity, 10 visible pictures were taken by the visible camcorder, meanwhile, 10 infrared images were taken by infrared camcorder. Following the evaluation program [15], [17], where the datasets are randomly divided into two halves, One part is used for training, and the other part is used for testing. In the test, images of one modal (visible/infrared) are used as the gallery set and images of the other modal (infrared/visible) are used as the query set. For a fair comparison, the procedure was repeated for 10 trials, and the mean values were recorded.

#### Evaluation Metrics.

Based on previous studies, we use the following evaluation metrics were used: cumulative Matching Characteristics (CMC), mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP). All person features are L2-normalized before testing.

2) IMPLEMENTATION DETAILS

We adopt the PyTorch framework to implement our method. Following previous studies, we use the ResNet50 model as a backbone network. We change the stride of the last convolutional block from 2 to 1. We used the stochastic gradient descent (SGD) as optimizer in the experiment, and set the momentum parameter to 0.9. We set the initial learning rate to 0.1 on both datasets. The warm-up learning rate strategy was employed to guide the network and improve the performance. The learning rate (lr) at epoch $e$ was computed as follows:

$$lr(e) = \begin{cases} 0.1 \times \frac{e+1}{10}, & 0 \leq e < 10 \\ 0.1, & 10 \leq e < 20 \\ 0.01, & 20 \leq e < 50 \\ 0.001, & 50 \leq e \end{cases}$$

The hyper-parameters are listed in Table 1.

#### B. ABLATION EXPERIMENTS

We demonstrate the validity of our presented approach, consisting of three modules: global and part-level feature learning, global and part-level feature fusion, and hetero-center triplet loss. To make a fair comparison, the mean values of 10 trials were reported on both the SYSU-MM01 and RegDB datasets during the ablation experiments.

1) GLOBAL AND PART-LEVEL FEATURE LEARNING

To enhance the learning of discriminative features that would help to identify a person, we propose joint learning of global features and part-level features with the same steps as shown in Figure 2.

Three points are considered: The first point is the number of strips. The second point is that the GeM pooling can be used for replacing the conventional average pooling or maximum pooling. The third point is that the global and part

---

**TABLE 1.** Settings of hyper-parameters.

| Dataset   | mc  | batch | dim | p  | index | $\beta$ |
|-----------|-----|-------|-----|----|-------|---------|
| SYSU-MM01 | 0.3 | 96    | 256 | 6  | 1     | 1.0     |
| RegDB     | 0.3 | 64    | 256 | 3  | 2     | 2.0     |
features dimension (dim) corresponds to the $1 \times 1$ Conv output channel number.

2) VALIDITY OF STRIPS
The number of strips reflects the granularity of the local features. The results for the different number of strips are shown in Figure 4 on the SYSU-MM01 and RegDB datasets. We observe that $p = 6$ is the optimal setting for strips setting to obtain part-level features.

3) VALIDITY OF GeM
This part demonstrates the effectiveness of GeM pooling compared with the traditional maximum pooling and average pooling methods. Table 2 gives the outcomes of the diverse pooling approaches on the SYSU-MM01 and RegDB datasets. The results show that maximum pooling outperforms average pooling, whereas the GeM pooling has the best performance. Experimental results demonstrate that the features obtained by Gem operation are more robust and discriminative.

4) VALIDITY OF THE FEATURE DIMENSION
This subsection represents the effect of the feature dimension (dim). Table 3 illustrates the outcomes with various dimensions of the feature (dim) on the SYSU-MM01 and RegDB datasets. For the SYSU-MM01 dataset, dim = 256 performs the best under the rank1 criteria; dim = 512 is slightly better under the mAP and mINP criteria. For the RegDB dataset, dim = 512 achieves the highest performance under the rank1 standard, and dim = 256 obtains the optimal performance with the mAP and mINP standards. Taking into account both computational cost and the performance, we set dim = 256 for the SYSU-MM01 and RegDB datasets.

2) Global and part-level feature fusion in the channel dimension: In this subsection, we verify the effectiveness of our proposed method of global and part-level feature fusion from two aspects. On the one hand, global and part-level fusion is compared to separate global or part-level features. On the other hand, we compare the effectiveness of different aggregation methods by changing the index of the global feature in the final feature, as shown in Figure 3.

5) GLOBAL + PART-LEVEL FEATURES vs. GLOBAL FEATURES vs. PART-LEVEL FEATURES
This experiment evaluates the effectiveness of GPFF compared with single global features or single part-level features. Table 4 shows the results of the different features on the SYSU-MM01 and RegDB datasets. We observe that part-level features outperform global features, whereas global + part-level features performs the best.

6) COMPARISON OF DIFFERENT FUSION METHODS
In this part, we compare the effectiveness of different fusion methods by changing the index of the global feature in the final feature. Table 5 lists the results of different fusion methods on the SYSU-MM01 and RegBD datasets. We observe that on the SYSU-MM01 dataset, index = 1 performs the best with the rank1, mAP, and mINP standards. On the RegDB
FIGURE 5. The matching results on the SYSU-MM01 datasets under “Infrared to visible” mode, the true/false searching results are shown in the green/red boxes.

TABLE 6. The comparative results of hetero-center triplet loss and sample-based triplet loss on SYSU-MM01 and RegDB datasets. Note that the loss $L_{tr}$ dose not converge on SYSU-MM01.

| Dataset | $L_{tr}$ | mAP | mINP | rank1 | mAP | mINP |
|---------|---------|-----|------|-------|-----|------|
| SYSU-MM01 | - | - | - | 86.32 | 82.10 | 73.25 |
| RegDB | 63.83 | 59.42 | 41.90 | 91.22 | 84.33 | 72.12 |

7) HETERO-CENTER TRIPLET LOSS
In this part, we verify the efficacy of our introduced hetero-center triplet loss ($L_{hc}$) by comparing it with the conventional triplet loss ($L_{tr}$).

8) HETERO-CENTER TRIPLET LOSS VS TRIPLET LOSS
We train our proposed GPFF with $L_{hc}$ and $L_{tr}$, respectively. The results are shown in Table 6. We can observe that the $L_{hc}$ is significantly better than that with index = 4 on SYSU-MM01 datasets. According to the above performance, we set index = 1 for SYSU-MM01 datasets, and index = 3 for RegBD datasets.

C. COMPARISON WITH THE STATE-OF-THE-ART
In this chapter, our method is compared with the state-of-the-art VI Re-ID methods. The experimental results on the two datasets (SYSU-MM01 and RegDB) are reported in Tables 6 and 7, respectively. Note that we did 10 experiments and took the average value as the experimental result.

The test results on the SYSU-MM01 datasets (Table 7) indicate that our approaches realize the best performance compared with the current state-of-the-art results obtained by HcTri [7], and it is superior to all other compared approaches. Specifically, in the more challenging all-search mode, our method considerably outperforms HcTri under the two key criteria, rank1/mAP: 63.83%/61.68% vs. 59.62%/57.51%. In the indoor search mode, our method considerably outperforms HcTri under the key criteria rank1: 66.67%/63.41%.

The test results on the RegDB dataset (Table 8) prove that our approaches outperform the baseline with the three standards, rank1/mAP/mINP: 91.22%/84.33%/72.12% for the visible-to-thermal query set, when the baseline achieved the following best performance in rank1/mAP/mINP: 91.05%/83.28%/68.84%.

We did not introduce additional parameters during testing, indicating that our method is suitable for practical application scenarios. We also show the qualitative retrieval results in Figure 5. Extensive experiments have demonstrated that discriminative features can be better learned by using our GPFF.

D. DISCUSSION
The greatest strength of this paper is the proposal of a simple and effective network architecture. Compared with [39], [40], and [42], our approach is significantly better than these methods. Moreover, our method does not require a complex network structure [40], nor does it employ multiple objective functions like [39], which increases the training difficulty. Still less does it require complicated adversarial learning [42]. Compared with the baseline [7], simply adding global features and subtly changing the position of the global features in the target features can lead to significant improvements.
in network performance. However, as can be seen from the experimental results in Tables 3, 5 and Figure 4, our network is more sensitive to hyperparameters, which makes the network less intelligent, and we need to further improve the network to make the extracted features more robust and discriminative.

V. CONCLUSION

This study aimed to extract discriminative features to identify a person using GPFF for VI Re-ID. First, the 3D feature map of visible and thermal images of a person was obtained using a dual-stream network. This network included a modal-specific submodule with different parameters and a modal-shared submodule with shared parameters. Second, the 3D feature map was divided into \( p \) part-level feature maps, which were learned together with the global feature map to obtain part-level feature vectors and the global feature vector. Finally, the final feature vector was formed by fusing the global feature vector and \( p \) part-level feature vectors in the channel dimension. The experimental results demonstrated the effectiveness of our proposed approach compared with the state-of-the-art methods. Our approach is simple but effective, and it can enhance high-quality research in the future.

REFERENCES

[1] Y. Chen, W. Zheng, and J. Lai, “Mirror representation for modeling view-specific transform in person re-identification,” in Proc. 24th Int. Conf. Artif. Intell., 2015, pp. 3402–3408.

[2] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 2197–2206, doi: 10.1109/CVPR.2015.7298832.

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

[4] L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 402–419.

[5] M. Ye, J. Shen, G. Lin, T. Xiang, L. Liu, and S. C. Hoi, “Deep learning for person re-identification: A survey and outlook,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 6, pp. 2872–2906, Jun. 2022, doi: 10.1109/TPAMI.2021.3054775.

[6] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu, “Cross-modality person re-identification with shared-specific feature transfer,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 13376–13386, doi: 10.1109/CVPR42600.2020.01339.

[7] H. Liu, X. Tan, and X. Zhou, “Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 13376–13386, doi: 10.1109/CVPR42600.2020.01339.

[8] H. Liu, Y. Chai, X. Tan, D. Li, and X. Zhou, “Strong but simple baseline with dual-granularity triplet loss for visible-thermal person re-identification,” IEEE Trans. Multimedia, vol. 23, pp. 4414–4425, 2021, doi: 10.1109/TMM.2021.3054775.

[9] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, “Relation-aware global attention for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 3183–3192, doi: 10.1109/CVPR42600.2020.00325.

[10] Y. Sun, Z. Y. L. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 480–496.

[11] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, “Learning discriminative features with multiple granularities for person re-identification,” in Proc. 26th ACM Int. Conf. Multimedia, Oct. 2018, pp. 274–282.
X. Fan, H. Luo, C. Zhang, and W. Jiang, “Cross-spectrum dual-subspace pairing for RGB-infrared cross-modality person re-identification,” 2020, arXiv:2003.00213.