Modality Synergy Complement Learning with Cascaded Aggregation for Visible-Infrared Person Re-Identification

Yiyuan Zhang[1], Sanyuan Zhao[1]*, Yuhao Kang[1], and Jianbing Shen[1,2]

[1] School of Computer Science, Beijing Institute of Technology
[2] SKL-IOTSC, Department of Computer and Information Science, University of Macau
{yiyuanzhang.ai, yuhaokangai, shenjianbingcg}@gmail.com
zhaosanyuan@bit.edu.cn

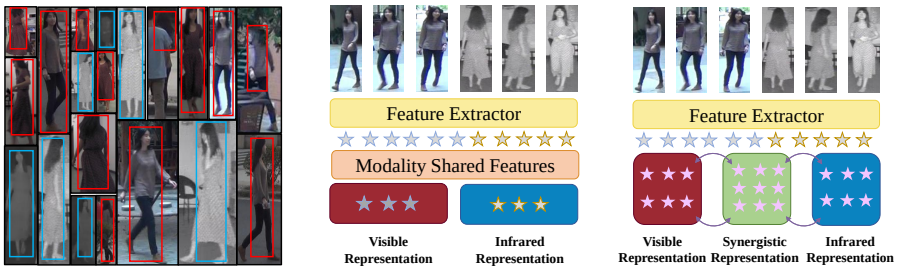
Abstract. Visible-Infrared Re-Identification (VI-ReID) is challenging in image retrievals. The modality discrepancy will easily make huge intra-class variations. Most existing methods either bridge different modalities through modality-invariance or generate the intermediate modality for better performance. Differently, this paper proposes a novel framework, named Modality Synergy Complement Learning Network (MSCLNet) with Cascaded Aggregation. Its basic idea is to synergize two modalities to construct diverse representations of identity-discriminative semantics and less noise. Then, we complement synergistic representations under the advantages of the two modalities. Furthermore, we propose the Cascaded Aggregation strategy for fine-grained optimization of the feature distribution, which progressively aggregates feature embeddings from the subclass, intra-class, and inter-class. Extensive experiments on SYSU-MM01 and RegDB datasets show that MSCLNet outperforms the state-of-the-art by a large margin. On the large-scale SYSU-MM01 dataset, our model can achieve 76.99% and 71.64% in terms of Rank-1 accuracy and mAP value. Our code will be available at https://github.com/bitreidgroup/VI-ReID-MSCLNet

Keywords: VI-ReID, Modality Synergy, Cascaded Aggregation

1 Introduction

Person re-identification (ReID) is a technique that retrieves a specific person in the gallery set shot by non-overlapping cameras [40,51,5,16]. The advancement of ReID plays an important role in smart city infrastructure and public security from the perspective of intelligent surveillance systems [28,11,20]. With the increasing demands for public security, surveillance systems are expected the ability to retrieve specific people precisely day and night. A technological requirement for Visible Infrared Person Re-Identification (VI-ReID) arises from

* Corresponding author.



(a) **Modal Characteristics** (b) **Existing Methods** (c) **Our Method**

Fig. 1: **Idea Illustration** (a) shows that, visible images usually contain discriminative fine-grained semantics, more noise and infrared images contain more similar semantics, less noise. (b) and (c) show that Synergistic features contain rich information about identities.

it. In contrast to visible person ReID [4,56], VI-ReID faces huge intra-class variations mainly due to the discrepancy between visible and infrared modalities. The modality discrepancy derives from properties of lights consisting of distinct wavelengths. Yet, their images are equivalently parsed as numerical matrices. Near-infrared is smoother and loses texture details due to longer wavelengths and more scattering. It becomes much more agnostic to skin color, albedo, and illumination. Similar texture, scatter, and color can represent different semantics. Besides, it is also difficult to ensure the perspectives of camera shooting, clothing of pedestrians, occlusion, and so on. These factors all contribute to a huge challenge in VI-ReID.

To address the aforementioned difficulties, most of the existing methods chiefly pay attention to learning modality-invariance to bridge the gap between visible and infrared images [49,9,10,11] or generating images of intermediate or the opposite modality for person retrieval [17,41]. However, GAN-based methods usually suffer from computational complexity and noise introduction. Unfortunately, pursuing modality-invariance may cause the networks to overlook feature properties of semantic diversity, as well as loss of identity discrimination.

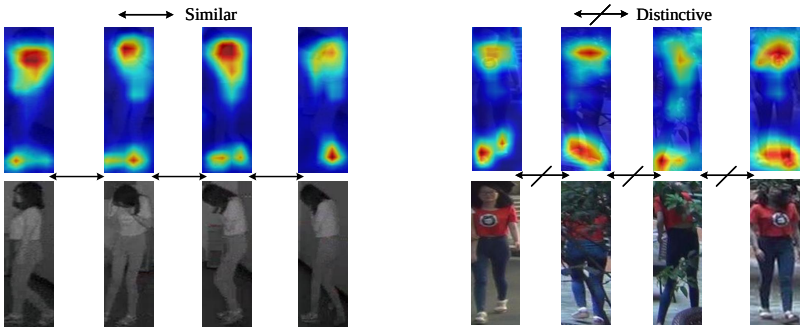
Differently, we consider the distinct representations and the semantic diversity between visible and infrared modalities. The success of visible person ReID validates that visible features are always discriminative enough to a large number of identities. Infrared cameras tend to capture thermal objects rather than non-thermal objects. The thermal sensitivity results in semantic loss and filtering of background noise. Infrared images represent relatively stable about the same identity and are comparatively immune to noise. Therefore, we conclude that *synergizing visible identity-discrimination and infrared noise-immunity can build noise-robust and retrieval-efficient representations for VI-ReID by learning*

homogeneous semantic discrimination and complementary characteristics across modalities as shown in Fig. 1.

Furthermore, traditional approaches to hard sample mining and feature-aggregation optimize distances of feature embeddings on the instance level. This kind of coarse-grained metric learning neglects the comprehensive distribution of all instances. We target to optimize on the different levels organized in cascaded manner. The basic idea is to subdivide instances of each identity into several subclasses according to the same shooting cameras. Instances in each subclass are much easier to aggregate, whose feature embeddings have a higher intra-class similarity. In this way, we can constrain distances between feature embeddings step-by-step.

Hence, we propose a novel framework, namely, Modality Synergy Complement Learning Network (MSCLNet). It aims at reducing the intra-class variations and boosting representations of identities discrimination. Firstly, it retains the intrinsic semantic diversity and identity relevance from visible and infrared modalities by constructing a synergistic representation with the Modality Synergy module (MS). Then, it enhances the synergistic representations by the specific advantages of the two modalities as shown in Fig. 2. MC contains these two parallel complementary processes with visible and infrared representations. On one hand, it provides guidance of fine-grained and discriminative features from the visible modality. On the other, it supplies global pedestrian statistics from the infrared modality. MS and MC greatly improve the capability of the network to represent identities across modalities. In addition, we propose the Cascaded Aggregation strategy (CA) to optimize the distribution of feature embeddings. It progressively aggregates samples into sub-class, intra-class, and inter-identities. In a cascaded manner, instance belonging to the same identities are lean to aggregation, and instances belonging to different identities are mapped to dispersion.

In conclusion, the main contributions of our work can be summarized as follows: We propose a novel framework named Modality Synergy Complement Learning Network (MSCLNet) with Cascaded Aggregation for VI-ReID. To fetch more discriminative semantics, it learns enhanced feature representations by diverse semantics and specific advantages of visible and infrared modalities. And we propose a Modality Synergy module (MS) which innovatively mines the modality-specific diverse semantics and a Modality Complement module (MC) which further enhances the feature representations by two parallel guidances of modality-specific advantages. They provide a reference for further high-level identity representation. Then we design a Cascaded Aggregation strategy (CA) to optimize the distribution of feature embeddings on a fine-grained level. It progressively aggregates the overall instances in a cascaded manner and enhances the discrimination of identities. Extensive experimental results show that our proposed framework outperforms the state-of-the-art methods by a large margin on two mainstream benchmarks of VI-ReID.



(a) **Infrared Advantages** (b) **Visible Advantages**

Fig. 2: **Demonstration of Infrared and Visible advantages.** Infrared images contain similar semantics and their feature embeddings are easier to aggregate. Visible images contain distinctive semantics even they describes the same person.

2 Related Work

Single-Modality Person Re-Identification retrieves pedestrians in the set of visible images. Visible person ReID is a reliable technique which plays an important role in daily life. These methods mainly solved the single-modal ReID problem via ranking [2,29], local and global attention [38,57], camera style [3,55,59], person key-points [36], siamese network [58], similarity graph [22], network architecture searching [18], *.etc.* Some works attempted domain adaptation [8,59]. And Some research dealt with the misalignment of human parts, such as cascaded convolutional module [39], refined part pooling [34], transformer [19] and so on. Beside, single-modality person re-identification contains several subdivided areas, for example, video person re-identification [26,44,60], unsupervised person re-identification which tackles pseudo labels [46,54], unsupervised domain adaption [1,31] and generalized person re-identification [16]. Due to the tremendous discrepancy between visible and infrared images, single-modal solutions are not suitable for cross-modality person re-identification, which creates a demand for the development of VI-ReID solutions.

Visible-Infrared Person Re-Identification focuses on narrowing the gap between visible and infrared modalities and learning appropriate representations for pedestrian retrieval across modalities. [43] proposed a deep zero-fill network to extract useful embedded features to reduce cross-modal variation. Dual-stream networks [21,48,49,50,51] simultaneously learned modal-shared and modal-specific features. [30] used Gaussian-based variational auto-encoder to distinguish the subspace of cross-modal features. [15] exploited samples similarity within modalities. A modality-aware learning approach [47] processed modality differences on the classifier level. Some works generated images of intermediate or the corresponding modality [7,17,35,37,40] to mitigate the effect of modality discrepancy. However, extracting modality-shared features causes the loss

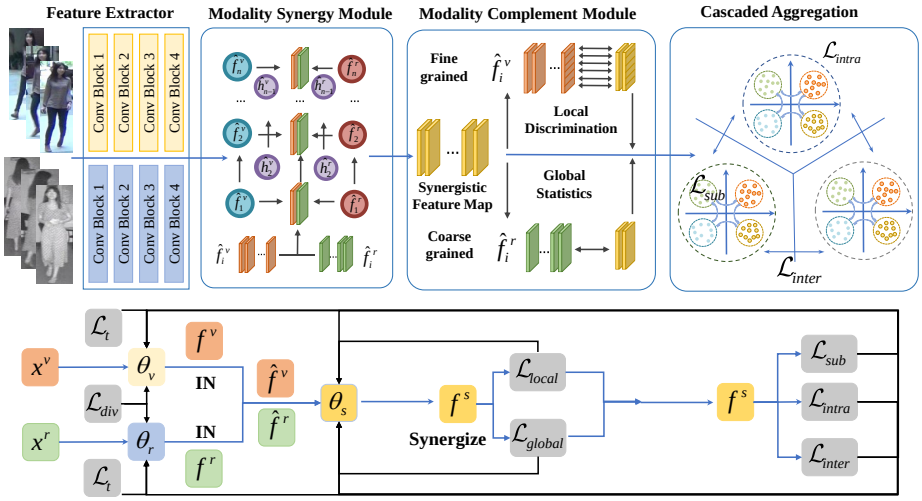


Fig. 3: **Illustration of our MSCLNet.** The images of visible and infrared modalities are fed into convolution blocks for visible and infrared representations. We synergize the single-modal features and complement synergistic features. Then, we design the Cascaded Aggregation strategy to fine-grained and progressively enhance feature embeddings.

of semantics related to identity discrimination, and GAN-based methods bring computational burden and non-original noise.

Differently, our work pays more attention to deep supervised knowledge synergy [32], which explores explicit information interaction between the supervised branches. We propose to make the most use of the intrinsic information of visible and infrared modalities, which learns diverse semantics and enhances feature representations by a modality synergy and complement learning scheme. To better discriminate identities, we introduce a cascaded feature aggregation strategy.

3 Modality Synergy Complement Learning

In this section, we formulate the VI-ReID problem and introduce the framework of our proposed MSCLNet (§ 3.1). It mainly contains three major components: Modality Synergy module (MS, § 3.2), Modality Complement module (MC, § 3.3), and Cascaded Aggregation strategy (CA, § 3.4). We utilize MS to synergize modality-specific diverse semantics from the extractors, and then use MC to enhance feature representations under the guidance of advantages from the two modalities. To optimize the distribution of the features and aggregate instances of the same identity, we exploit CA to constrain the feature distribution in a fine-grained and progressive way. Finally, we summarize the proposed loss function (§ 3.5).

3.1 Problem Formulation

We take $\mathcal{V} = \{x_i^v | x_i^v \in \mathcal{V}\}$ and $\mathcal{R} = \{x_i^r | x_i^r \in \mathcal{R}\}$ to denote visible and infrared images, respectively. $\mathcal{Y}_v = \{y_i^v | x_i^v \in \mathcal{V}\}$ and $\mathcal{Y}_r = \{y_i^r | x_i^r \in \mathcal{R}\}$ indicates the corresponding identity labels. Given a query person image x_Q^v or x_Q^r, VI-ReID aims to retrieve the most precise result in the gallery set x_G^r or x_G^v. Existing methods extract modality-shared features at the cost of discarding modality-specific semantics of diversity which can well depict the person. Therefore, we take these intrinsic diverse semantics and the special advantages of each modality into consideration, to learn more precise and better discriminative representation for identities.

Fig. 3 illustrates the framework of Modality Synergy Complement Learning Network (MSCLNet) with Cascaded Aggregation. It adopts a dual-stream network as the feature extractor. Firstly, based on the extracted feature representations f^v and f^r from visible and images, MSCLNet constructs synergistic representations f^s by constraining the diversity of the feature distributions between the two modalities. The synergistic feature will be further enhanced by modality complement guidance. The visible modality provides fine-grained discriminative semantics, while the infrared modality supplies with stable global pedestrian statistics. Then we aggregate feature embeddings of the same class via Cascaded Aggregation strategy which optimizes the comprehensive distribution of feature embeddings progressively on three aspects.

3.2 Modality Synergy Module

According to the differences in imaging principles and the heterogeneity of the image contents, visible and infrared images reveal quite different semantics to depict the same person. In our work, we design the network to learn and synergize the diverse semantics of the two modalities. Given a pair of visible and infrared images $x_i^v \in \mathcal{V}$, $x_i^r \in \mathcal{R}$, the dual-stream network extracts their features f_i^v and f_i^r. With the prerequisite of precise pedestrian re-identification, we concentrate on acquiring the semantic diversity to the largest extent. Features f_i^v and f_i^r are normalized by the following operations.

$$\hat{f}_i^v = \frac{f_i^v - \mathrm{E}\left[f_i^v\right]}{\sqrt{\mathrm{Var}\left[f_i^v\right] + \epsilon^v}} \times \gamma + \beta, \mathrm{E}\left[f_i^v\right] = \frac{1}{HW} \sum_{l=1}^{W} \sum_{m=1}^{H} f_{itlm}, \tag{1}$$

Where $\mathrm{Var}\left[f_i^v\right] = \frac{1}{HW} \sum_{l=1}^{W} \sum_{m=1}^{H} \left(f_{itlm} - \mathrm{E}\left[f_i^v\right]\right)^2$ are calculated per-dimension separately for each instance in a mini-batch. Let $\mathcal{S}(\cdot)$ indicate the Modality Synergy module to construct synergistic feature f_i^s with label y_i on the basis of f_i^v, f_i^r:

$$f_i^s = \mathcal{S}(\hat{f}_i^v, \hat{f}_i^r, y_i, \theta_s), \tag{2}$$

where θ_s acts as parameters of the Modality Synergy module $\mathcal{S}(\cdot)$. We utilize Mogrifier LSTM [25] as a synergistic feature encoder to maximize the effect of modality synergy learning, and the synergistic feature f_i^s is encoded with visible

and infrared features with their shared ground-truth label. To construct f_i^s with diverse semantics, we exploit KL-Divergence to constrain the logistic distribution of visible and infrared features f_i^v, f_i^r, which can be formulated as follows:

$$\mathcal{L}_{div} = -\,\mathrm{KL}(\hat{f^v} \parallel \hat{f^r}) = -\frac{1}{N}\sum_{i=1}^{N}(\hat{f_i^v} \cdot \log \frac{\hat{f_i^v}}{\hat{f_i^r}}, \theta_v, \theta_r), \tag{3}$$

where N denotes the number of samples in a batch. θ_v and θ_r act as learned feature extractors of visible and infrared modalities respectively, which aim to maximize the diversity of semantic representation across modalities. f^v and f^r are firstly designed in the representation spaces to maximize the modality-specific discrimination among identities. Then, the synergistic feature extractor θ_s projects $\hat{f_i^v}$, $\hat{f_i^r}$ to a shared representation space and constructs synergistic features f_i^s.

Furthermore, we constrain the diverse semantics by identity-relevance, which introduces cross entropy constraining the logistic probability of visible and infrared features p_i^v and p_i^r and the ground truth label y_i.

$$\mathcal{L}_t = -\frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i \cdot \log \hat{p}_i^v(\hat{f}_i^v, \theta_v)] - \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i \cdot \log \hat{p}_i^r(\hat{f}_i^r, \theta_r)] \tag{4}$$

where λ_{div} and λ_t are hype-parameters to balance the contributions of individual loss terms. The optimization processes of θ_v, θ_r separately track the the gradient of $(\frac{\partial f^v}{\partial x^v}, \frac{\partial f^s}{\partial x^v})$ and $(\frac{\partial f^r}{\partial x^r}, \frac{\partial f^r}{\partial x^r})$.

$$\mathcal{L}_{Synergy} = \mathcal{L}(\theta_v, \theta_r) = \lambda_{div} \cdot \mathcal{L}_{div} + \lambda_t \cdot \mathcal{L}_t \tag{5}$$

3.3 Modality Complement Module

Although synergistic representation contains more identity-relevant diverse semantics, it is uncertain whether synergistic feature outperforms the combination of visible and infrared features $\mathrm{Concat}(f_i^v, f_i^r)$. Due to infrared images containing global pedestrian statistics with less noise and visible images containing fine-grained discriminative semantics, we enhance the representation effectiveness of synergistic feature f_i^s from two aspects. Considering fine-grained semantics, we enhance synergistic features with advantages of visible features f_i^v in terms of local parts. And considering coarse-grained semantics, we enhance synergistic features with advantages of infrared features f_i^r about global parts.

On the fine-grained level, we split visible and synergistic features into $n = 6$ parts as MPANet [45] and get separate feature blocks as $f_i^v = [b_1^v, b_2^v \cdots, b_n^v]$, $f_i^s = [b_1^s, b_2^s \cdots, b_n^s]$. The local discrimination of synergistic features can be boosted with nuanced regions of visible modality. Cosine similarity $cos(\cdot, \cdot)$ is utilized for the optimization process.

$$\mathcal{L}_{local} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}(cos(b_j^v, b_j^s) + \sqrt{2 - 2cos(b_j^v, b_j^s)}) \tag{6}$$

In parallel, on the coarse-grained level, we supervise f_i^s by keeping the statistic centers of synergistic features consistent with that of the infrared feature f_i^r. The global statistics of synergistic features can get optimized with center consistency of infrared modality.

$$\mathcal{L}_{global} = \frac{1}{N}\sum_{i=1}^{N} ||C_{y_i}^s - C_{y_i}^r||_2^2, \tag{7}$$

where $C_{y_i}^s, C_{y_i}^r$ denote the center of the $y_i{}^{th}$ class for synergistic features f_i^s, f_i^r. $\mathcal{L}_{global}$ helps to coordinate semantics of the synergistic and the infrared feature and filter identity-irrelevance of the synergistic representation.

In the progress of Modality Complement module, we update the parameters of synergistic feature extractor θ_s, which aims to construct features with less noise, more diverse and more precise semantic description for each identity. θ_s is optimized as follows:

$$\mathcal{L}_{Com}(\theta_s) = \lambda_{local} \cdot \mathcal{L}_{local} + \lambda_{global} \cdot \mathcal{L}_{global}, \hat{\theta}_s = \arg\min_{\theta_s} \mathcal{L}(\theta_s), \tag{8}$$

where $\lambda_{local}, \lambda_{global}$ are hyper-parameters to balance the contributions of individual loss terms.

3.4 Cascaded Aggregation Strategy

Due to factors like shooting perspectives, clothing, and occlusion, the results of person retrieval will easily be affected [53,33]. To cope with this problem, center loss [23] and triplet loss [14] are widely adopted in ReID problems to simultaneously learn the centralized representation of feature embeddings and mine hard samples. Center loss $\mathcal{L}_c$ and Triplet loss $\mathcal{L}_{tri}$ can be formulated as:

$$\begin{aligned} \mathcal{L}_c &= \frac{1}{N}\sum\nolimits_{i=1}^{N} \|f_i - C_{y_i}\|_2^2, \\ \mathcal{L}_{tri} &= \sum_{i}^{N} \left[\|f\left(x_i^a\right) - f\left(x_i^{pos}\right)\|_2^2 - \|f\left(x_i^a\right) - f\left(x_i^{neg}\right)\|_2^2 + \alpha\right]_+ \end{aligned} \tag{9}$$

where x_i denotes the i^{th} input sample, C_{y_i} is the $y_i{}^{th}$ class center, f_i is the feature embedding, x_i^a is the anchor. Center loss pays attention to aggregating feature embeddings but neglects the intrinsic differences and diverse semantics existing in the visible and the infrared modalities. Triplet loss specializes in handling hard samples separately rather than considering the comprehensive distribution across modalities, which limits the performance. Considering the diverse semantics and structural distribution across modalities, we propose Cascaded Aggregation to progressively optimize the features distribution of, as shown in Fig. 4.

1) Aggregation on Sub-class level. We utilize the identity of shooting cameras for each image as the natural sub-class, since images of the same person shot by the same camera have high similarities with each other, where C_{s_i} denotes the

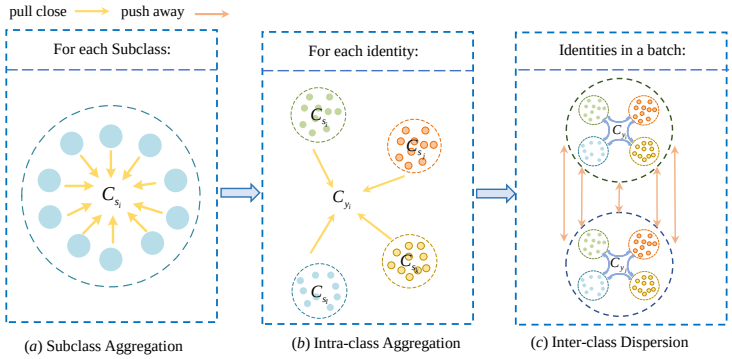


Fig. 4: **Cascaded Aggregation demonstration** (a) indicates the optimization for subclass aggregation, (b) indicates the intra-class aggregation, and (c) indicates the inter-class dispersion.

s_i^{th} sub-class center:

$$\mathcal{L}_{sub} = \frac{1}{N}\sum_{i=1}^{N}||f_i^s - C_{s_i}||_2^2, \tag{10}$$

2) Aggregation on the intra-class level, which keeps the structural priors of the features during the training progress. The formulation of the aggregation can be represented as follows, where N_s denotes the number of the sub-classes of each identity.

$$\mathcal{L}_{intra} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_s}||C_{s_j} - C_{y_i}||_2^2, \tag{11}$$

3) Aggregation on the inter-class level. Our method of aggregation not only maximizes the similarity of intra-class instances but also maximizes the dissimilarity of inter-class instances on the whole. The dispersion between different identities and the two types of aggregation in 1) and 2) of the same identities are independent of each other. Formally, the dispersion between different identities can be represented as:

$$\mathcal{L}_{inter} = -\frac{1}{\binom{N}{2}}\sum_{i=1}^{N}\sum_{j\neq i}^{N}||C_{y_i} - C_{y_j}||_2^2. \tag{12}$$

The loss function of CA for metric learning can be represented as:

$$\begin{aligned}\mathcal{L}_{cascade} &= \mathcal{L}_{sub} + \mathcal{L}_{intra} + \mathcal{L}_{inter} \\ &= \frac{1}{N}\sum_{i=1}^{N}||f_i^s - C_{s_i}||_2^2 + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_s}||C_{s_j} - C_{y_i}||_2^2 - \frac{1}{\binom{2}{N}}\sum_{i=1}^{N}\sum_{j\neq i}^{N}||C_{y_i} - C_{y_j}||_2^2.\end{aligned} \tag{13}$$

Compared with Center Loss, our method begins with only a few samples of high similarity for the same shooting cameras and it will become much easier to learn sub-center representations.
Compared with Triplet Loss, our method deals with negative samples simultaneously by guiding the negative samples to the correspondent sub-class instead of easily pushing away alongside the gradient.

3.5 Objective Function

Firstly, we utilize Synergy Loss $\mathcal{L}_{Synergy}$ to enrich the representation on diverse semantics. The parameters of feature extractors θ_v and θ_r are updated as:

$$\mathcal{L}_{Synergistic} = \mathcal{L}(\theta_v, \theta_r) = \lambda_{div} \cdot \mathcal{L}_{div} + \lambda_t \cdot \mathcal{L}_t. \tag{14}$$

Then, we enhance the synergistic feature representation with the advantages of two modalities, namely, the discriminative local parts from the visible feature and global identity statistics from the infrared feature. We utilize Complementary Loss $\mathcal{L}_{Com}$ to update the modality synergy feature extractor θ_s:

$$\mathcal{L}_{com} = \mathcal{L}(\theta_s) = \lambda_{local} \cdot \mathcal{L}_{local} + \lambda_{global} \cdot \mathcal{L}_{global}. \tag{15}$$

Finally, we constrain the distribution of visible, infrared and synergistic feature f^v, f^r, f^s with cascaded aggregation strategy $\mathcal{L}_{cascaded}$:

$$\mathcal{L}_{cascaded} = \mathcal{L}(\theta_v, \theta_r, \theta_s) = \mathcal{L}_{sub} + \mathcal{L}_{intra} + \mathcal{L}_{inter}. \tag{16}$$

Overall, the objective function of our MSCLNet can be summarized as follows:

$$\mathcal{L}_{total} = \lambda_{div}\mathcal{L}_{div} + \lambda_t\mathcal{L}_t + \lambda_{local}\mathcal{L}_{local} + \lambda_{global}\mathcal{L}_{global} + \mathcal{L}_{sub} + \mathcal{L}_{intra} + \mathcal{L}_{inter} \tag{17}$$

4 Experiment

4.1 Datasets and Evaluation Protocol

SYSU-MM01 [43] is a large-scale dataset for VI-ReID which contains 491 pedestrians with total 287,628 visible images and 15,792 infrared images. It collects samples by 6 cameras, *i.e.* 4 visible and 2 infrared cameras, in the outdoor and indoor environments. It contains two different testing modes, *all-search* and *indoor-search* modes. Compared with RegDB, SYSU-MM01 is more challenging due to the large variations between samples.

RegDB [28] collects 412 identities, and each identity has 10 visible images and 10 infrared images. We randomly choose 206 identities for training and the left for testing [48]. There are two modes in testing, *visible-to-infrared* and *infrared-to-visible*.The former denotes that the model retrieves the person in the infrared gallery when given a visible image, and vice versa. We average the results for 10 trials for stable performance [40].

Evaluation Protocol. The cumulative matching characteristics (CMC) [27], and mean average precision (mAP) are used as evaluation metrics.

4.2 Implement Details

Training. We implement MSCLNet with PyTorch on a single NVIDIA RTX 2080 Ti GPU and deal with 64 images consisting of 32 visible and 32 infrared images of 8 identities in a mini-batch by randomly selecting 4 visible and 4 infrared images for each identity. Our baseline is AGW*, which means AGW [51] with Random Erasing. We adopt pre-trained ResNet-50 [13] on ImageNet as the backbone network. Then, we pre-process each image by re-scaling in to 288×144 and augment images through random cropping with zero-padding, random horizontal flipping and random erasing (80% probability, 80% max-area, 20% min-area) . During the training process, we optimize the feature extractors θ_v, θ_r and modality synergy module θ_s with SGD optimizer. We set the initial learning rate $\eta = 0.1$, the momentum parameter $p = 0.9$. The learning rate is changed as $\eta = 0.05$ at 21-50 epoch, $\eta = 0.01$ at 51-100 epoch, and $\eta = 0.001$ at 101-200 epoch. The hyper-parameters $\lambda_{div}, \lambda_t, \lambda_{local}, \lambda_{global}$ are set to 0.5, 1.25, 0.8, and 1.5, respectively. We synergize visible and infrared instances to train a concise end-to-end network, which retrieves specific person across modalities.

Testing. For testing, the model works in *Single-shot* mode by extracting the query and the gallery features from a single modality by the feature extractor θ_v or θ_r. Besides, MS and MC modules **do not** participate in testing stage.

4.3 Ablation Study

In this subsection, we conduct an ablation study to evaluate the effectiveness of each component of MSCLNet, as summarized in Eq. 17. The results are demonstrated in Tab. 1. We evaluate how much improvement can be made by each component on the *all-search* mode of SYSU-MM01 dataset.

Table 1: Analysis of the effectiveness of MS, MC, CA on SYSU-MM01 dataset in the *all-search* mode. Rank-1 accuracy(%) and mAP(%)are reported.

Methods								Metric	
	MS		MC			CA			
B	$\mathcal{L}_{div}$	$\mathcal{L}_t$	$\mathcal{L}_{glocal}$	$\mathcal{L}_{local}$	$\mathcal{L}_{sub}$	$\mathcal{L}_{intra}$	$\mathcal{L}_{inter}$	Rank-1	mAP
✓								59.82	56.07
✓	✓							60.32	56.79
✓		✓						60.67	58.12
✓			✓					62.14	59.94
✓				✓				61.33	59.23
✓					✓			61.74	59.88
✓						✓		61.96	60.40
✓							✓	63.55	60.97
✓	✓	✓						62.82	60.25
✓			✓	✓				64.84	61.00
✓					✓	✓	✓	66.13	61.99
✓	✓	✓	✓	✓				71.16	66.30
✓	✓	✓			✓	✓	✓	69.78	65.29
✓			✓	✓	✓	✓	✓	72.81	67.66
✓	✓	✓	✓	✓	✓	✓	✓	76.99	71.64

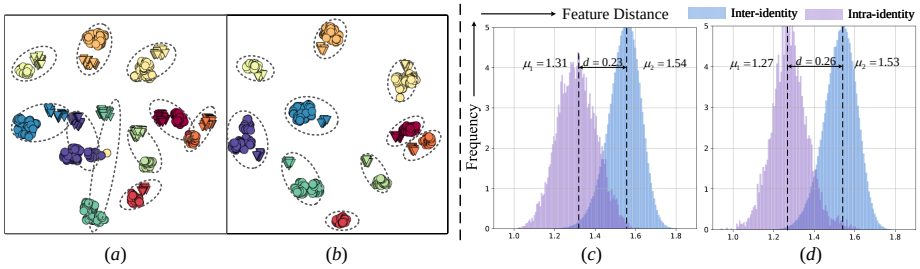


Fig. 5: **Visualization Results**. (a) and (b) show the feature embeddings distribution of baseline and MSCLNet via t-SNE [24], where circles and triangles in different colors denote visible and infrared modalities. (c) and (d) show the intra-and-inter distribution of feature distance.

Effectiveness of MS. Referring to the ninth row, we add the MS structure to the baseline, and the baseline obtains a rank-1 score of 62.82% and a mAP of 60.25%, improved by 3% and 19.75%. Meanwhile, we add MS to other combinations as shown in the 12th, 13th, 15th rows. MS also brings different degrees of enhancement to the model.

Effectiveness of MC. The experimental setting of Base+MC acquires 64.84% at Rank-1 and 61.00% at mAP. When baseline works with MS+MC, a further improvement is reached, where rank-1 is 71.16% and mAP is 66.3%. This illustrates that Modality Synergy Complement Learning effectively improves performance.

Effectiveness of CA. The settings of Base+MS+CA and Base+MC+CA work better than merely utilizing one of the three modules. Base+MS+MC+CA reaches the best result, in which rank-1 is 69.78%, mAP is 65.29%.

Overall. The results show that each component of MSCLNet can improve precision. At the same time, they work better when cooperating, which reveals that the three components focus on different aspects of optimization.

4.4 Visualization Analysis

To present the effectiveness of MSCLNet, we visualize the feature distribution via t-SNE [24] as shown in Fig. 5. Different colors denote different identities. For the baseline, feature embeddings of some identities entangle with each other, which indicates the baseline is confused about these identities. In comparison, MSCLNet discriminates and aggregates these feature embeddings of the same identity separately and clearly.

Meanwhile, we also visualize the feature distances analysis between baseline and MSCLNet in Fig. 5. After numerical analysis, our conclusions are as follows: 1) Distances between these distribution increase $d = 0.23 \rightarrow 0.26$ and the mean distance of intra-identity reduces $\mu = 1.31 \rightarrow 1.27$. 2) Variance of intra-identity distribution reduces prominently $\sigma = 9.5 \times 10^{-2} \rightarrow 8.8 \times 10^{-2}$ and the distribution of intra-identity aggregates better.

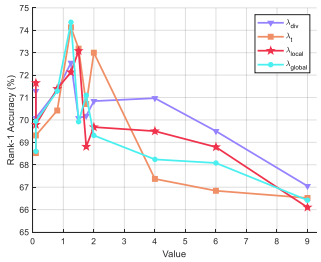


Fig. 6: **Hyper-parameters Sensitive Graph**

4.5 Parameters Analysis

We analyze the parameter sensitivity under the condition that, single hyper-parameter was selected as a variable and all other hyper-parameters are kept constant. Thus, we can obtain the curve of the effect of changing hyper-parameters on Rank-1 accuracy (%). In turn, by continuously changing the variables, we can get the sensitivity graphs of all hyper-parameters $\lambda_{div}, \lambda_t, \lambda_{global}, \lambda_{local}$ as shown in Fig 6. It clearly shows that five hyper-parameters present different sensitivities and most of their optimal intervals of these parameters are in $[1, 2]$.

4.6 Comparison with State-of-the-Art Methods

Table 2: Comparison with the state-of-the-arts on SYSU-MM01 dataset. Rank-k accuracy (%) and mAP (%) are reported.

Settings		*All Search*				*Indoor Search*			
Method	Venue	R1	R10	R20	mAP	R1	R10	R20	mAP
Zero-Pad [43]	ICCV17	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
HCML [48]	AAAI18	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
cmGAN [7]	IJCAI18	26.97	67.51	80.56	27.80	31.63	77.23	89.18	42.19
HSME [12]	AAAI19	20.68	32.74	77.95	23.12	-	-	-	-
AliGAN [37]	ICCV19	42.40	85.00	93.70	40.70	45.90	87.60	94.40	54.30
CMSP [42]	IJCV20	43.56	86.25	-	44.98	48.62	89.50	-	57.50
JSIA [35]	AAAI20	38.10	80.70	89.90	36.90	43.80	86.20	94.20	52.90
XIV [17]	AAAI20	49.92	89.79	95.96	50.73	-	-	-	-
MACE [47]	TIP20	51.64	87.25	94.44	50.11	57.35	93.02	97.47	64.79
MSR [9]	TIP20	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
Hi-CMD [6]	CVPR20	34.94	77.58	-	35.94	-	-	-	-
cm-SSFT [21]	CVPR20	47.70	-	-	54.10	-	-	-	-
AGW [51]	TPAMI21	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
MCLNet [11]	ICCV21	65.40	93.33	97.14	61.98	72.56	96.88	99.20	76.58
SMCL [41]	ICCV21	67.39	92.87	96.76	61.78	68.84	96.55	98.77	75.56
NFS [5]	CVPR21	56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
CM-NAS [10]	CVPR21	61.99	92.87	97.25	60.02	67.01	97.02	99.32	72.95
MPANet [45]	CVPR21	70.58	96.21	98.80	68.24	76.74	98.21	99.57	80.95
MSCLNet	**Ours**	**76.99**	**97.63**	**99.18**	**71.64**	**78.49**	**99.32**	**99.91**	**81.17**

We compare the proposed MSCLNet with state-of-the-art methods. Tab. 2 and Tab. 3 illustrate the comparison results on the SYSU-MM01 and the RegDB

datasets. MSCLNet outperforms the other methods on both of the benchmarks. On the SYSU-MM01 dataset, MSCLNet achieves the rank-1 scores of 76.99% and mAP score of 71.64% in the *all-search* mode, higher than MPANet [45] by 6.41% and 3.40%. On the RegDB dataset, MSCLNet achieves Rank-1 scores of 84.17% and 83.86% in visible-to-infrared and infrared-to-visible modes, better than NFS [10] by 3.63% and 5.91%, respectively.

Table 3: Comparison with the state-of-the-arts on RegDB dataset. Rank-1 accuracy (%) and mAP(%) are reported.

Settings		*Visible to Infrared*		*Infrared to Visible*	
Method	Venue	Rank-1	mAP	Rank-1	mAP
Zero-Pad [43]	ICCV'17	17.75	18.90	16.63	17.82
HCML [48]	AAAI'18	24.44	20.08	21.70	22.24
HSME [12]	AAAI'19	50.85	47.00	50.15	46.16
AliGAN [37]	ICCV'19	57.90	53.60	56.30	53.40
CMSP [42]	IJCV'20	65.07	64.50	-	-
JSIA [35]	AAAI'20	48.10	48.90	48.50	49.30
XIV [17]	AAAI'20	62.21	60.18	-	-
DG-VAE [30]	ACM MM'20	72.97	71.78	-	-
HAT [52]	TIFS'20	71.83	67.56	70.02	66.30
MSR [9]	TIP'20	48.43	48.67	-	-
MACE [47]	TIP'20	72.37	69.09	72.12	68.57
DDAG [50]	ECCV'20	69.34	63.46	68.06	61.80
Hi-CMD [6]	CVPR'20	70.93	66.04	-	-
AGW [51]	TPAMI'21	70.05	66.37	70.49	65.90
MCLNet [11]	ICCV'21	80.31	73.07	75.93	69.49
NFS [5]	CVPR'21	80.54	72.10	77.95	69.97
MPANet [45]	CVPR'21	83.70	80.90	82.80	**80.70**
MSCLNet	**Ours**	**84.17**	**80.99**	**83.86**	78.31

5 Conclusion and Discussion

In this paper, we propose a novel VI-ReID framework, which has the capability to make full use of the visible and the infrared modality semantics and learn discriminative representation of identities by synergizing and complementing instances of visible and infrared modalities. Different from existing methods pursuing modal-shared information at the risk of identity-relevant semantics loss, MSCLNet provides an innovative approach exploring high-level unity in VI-ReID task. Meanwhile, we propose Cascaded Aggregation strategy to fine-grained and progressively optimize the distribution of feature embeddings, which assists the network discriminate identities and extract more precise and more comprehensive features. Experimental results validate the merit of the framework, as well as the effectiveness of each component in this framework. In the future work, we plan to explore background scenes, gender, and appearances to construct better different sub-classes.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 61902027, and the Start-up Research Grant (SRG) of University of Macau.

References

1. Ahmed, S.M., Lejbolle, A.R., Panda, R., Roy-Chowdhury, A.K.: Camera on-boarding for person re-identification using hypothesis transfer learning. In: CVPR. pp. 12144–12153 (2020)
2. Bai, S., Tang, P., Torr, P.H., Latecki, L.J.: Re-ranking via metric fusion for object retrieval and person re-identification. In: CVPR. pp. 740–749 (2019)
3. Chen, G., Lin, C., Ren, L., Lu, J., Zhou, J.: Self-critical attention learning for person re-identification. In: ICCV. pp. 9637–9646 (2019)
4. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. In: CVPR. pp. 8351–8361 (2019)
5. Chen, Y., Wan, L., Li, Z., Jing, Q., Sun, Z.: Neural feature search for rgb-infrared person re-identification. In: CVPR. pp. 587–597 (June 2021)
6. Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C.: Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: CVPR. pp. 10257–10266 (2020)
7. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: IJCAI. pp. 677–683 (2018)
8. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: CVPR. pp. 994–1003 (2018)
9. Feng, Z., Lai, J., Xie, X.: Learning modality-specific representations for visible-infrared person re-identification. IEEE TIP **29**, 579–590 (2019)
10. Fu, C., Hu, Y., Wu, X., Shi, H., Mei, T., He, R.: Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In: ICCV. pp. 11823–11832 (October 2021)
11. Hao, X., Zhao, S., Ye, M., Shen, J.: Cross-modality person re-identification via modality confusion and center aggregation. In: ICCV. pp. 16403–16412 (October 2021)
12. Hao, Y., Wang, N., Li, J., Gao, X.: Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In: AAAI. pp. 8385–8392 (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
15. Jia, M., Zhai, Y., Lu, S., Ma, S., Zhang, J.: A similarity inference metric for rgb-infrared cross-modality person re-identification. arXiv preprint arXiv:2007.01504 (2020)
16. Jin, X., Lan, C., Zeng, W., Chen, Z., Zhang, L.: Style normalization and restitution for generalizable person re-identification. In: CVPR. pp. 3143–3152 (2020)
17. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: AAAI. pp. 4610–4617 (2020)
18. Li, H., Wu, G., Zheng, W.S.: Combined depth space based architecture search for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6729–6738 (2021)
19. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: Occluded person re-identification with part-aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2898–2907 (2021)

20. Lin, Y., Xie, L., Wu, Y., Yan, C., Tian, Q.: Unsupervised person re-identification via softened similarity learning. In: CVPR. pp. 3390–3399 (2020)
21. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N.: Cross-modality person re-identification with shared-specific feature transfer. In: CVPR. pp. 13379–13389 (2020)
22. Luo, C., Chen, Y., Wang, N., Zhang, Z.: Spectral feature transformation for person re-identification. In: CVPR. pp. 4976–4985 (2019)
23. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPR Workshops. pp. 0–0 (2019)
24. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
25. Melis, G., Kočiskỳ, T., Blunsom, P.: Mogrifier lstm. arXiv preprint arXiv:1909.01792 (2019)
26. Meng, J., Zheng, W.S., Lai, J.H., Wang, L.: Deep graph metric learning for weakly supervised person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1–1 (2021)
27. Moon, H., Phillips, P.J.: Computational and performance aspects of pca-based face-recognition algorithms. Perception **30**(3), 303–321 (2001)
28. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017)
29. Paisitkriangkrai, S., Shen, C., Van Den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: CVPR. pp. 1846–1855 (2015)
30. Pu, N., Chen, W., Liu, Y., Bakker, E.M., Lew, M.S.: Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. In: ACMMM. pp. 2149–2158 (2020)
31. Ren, C.X., Liang, B.H., Lei, Z.: Domain adaptive person re-identification via camera style generation and label propagation. IEEE Transactions on Information Forensics and Security **15**, 1290–1302 (2019)
32. Sun, D., Yao, A., Zhou, A., Zhao, H.: Deeply-supervised knowledge synergy. In: CVPR. pp. 6997–7006 (2019)
33. Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: CVPR. pp. 608–617 (2019)
34. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV. pp. 480–496 (2018)
35. Wang, G.A., Yang, T.Z., Cheng, J., Chang, J., Liang, X., Hou, Z., et al.: Cross-modality paired-images generation for rgb-infrared person re-identification. In: AAAI. pp. 12144–12151 (2020)
36. Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun, J.: High-order information matters: Learning relation and topology for occluded person re-identification. In: CVPR. pp. 6449–6458 (2020)
37. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: ICCV. pp. 3623–3632 (2019)
38. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR. pp. 2275–2284 (2018)
39. Wang, Y., Chen, Z., Feng, W., Gang, W.: Person re-identification with cascaded pairwise convolutions. In: IEEE Conference on Computer Vision and Pattern Reconigtion (2018)

40. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: CVPR. pp. 618–626 (2019)
41. Wei, Z., Yang, X., Wang, N., Gao, X.: Syncretic modality collaborative learning for visible infrared person re-identification. In: ICCV. pp. 225–234 (October 2021)
42. Wu, A., Zheng, W.S., Gong, S., Lai, J.: Rgb-ir person re-identification by cross-modality similarity preservation. IJCV pp. 1–21 (2020)
43. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: ICCV. pp. 5380–5389 (2017)
44. Wu, D., Ye, M., Lin, G., Gao, X., Shen, J.: Person re-identification by context-aware part attention and multi-head collaborative learning. IEEE Transactions on Information Forensics and Security **17**, 115–126 (2021)
45. Wu, Q., Dai, P., Chen, J., Lin, C.W., Wu, Y., Huang, F., Zhong, B., Ji, R.: Discover cross-modality nuances for visible-infrared person re-identification. In: CVPR. pp. 4330–4339 (June 2021)
46. Xuan, S., Zhang, S.: Intra-inter camera similarity for unsupervised person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11926–11935 (2021)
47. Ye, M., Lan, X., Leng, Q., Shen, J.: Cross-modality person re-identification via modality-aware collaborative ensemble learning. IEEE TIP **29**, 9387–9399 (2020)
48. Ye, M., Lan, X., Li, J., Yuen, P.C.: Hierarchical discriminative learning for visible thermal person re-identification. In: AAAI. pp. 7501–7508 (2018)
49. Ye, M., Lan, X., Wang, Z., Yuen, P.C.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. IEEE TIFS **15**, 407–419 (2019)
50. Ye, M., Shen, J., Crandall, D.J., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: ECCV (2020)
51. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook. arXiv preprint arXiv:2001.04193 (2020)
52. Ye, M., Shen, J., Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE TIFS (2020)
53. Yu, S., Li, S., Chen, D., Zhao, R., Yan, J., Qiao, Y.: Cocas: A large-scale clothes changing person dataset for re-identification. In: CVPR. pp. 3400–3409 (2020)
54. Zhang, X., Ge, Y., Qiao, Y., Li, H.: Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3436–3445 (2021)
55. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In: CVPR. pp. 10407–10416 (2020)
56. Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z.: Relation-aware global attention for person re-identification. In: CVPR. pp. 3186–3195 (2020)
57. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal person re-identification via multi-loss dynamic training. In: CVPR. pp. 8514–8522 (2019)
58. Zheng, M., Karanam, S., Wu, Z., Radke, R.J.: Re-identification with consistent attentive siamese networks. In: CVPR. pp. 5735–5744 (2019)
59. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: CVPR. pp. 5157–5166 (2018)
60. Zhu, X., Jing, X.Y., You, X., Zuo, W., Shan, S., Zheng, W.S.: Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix. IEEE Transactions on Information Forensics and Security pp. 1–1 (2017)