Visible-Infrared Person Re-Identification via Semantic Alignment and Affinity Inference

Xingye Fang[1], Yang Yang[2], Ying Fu[1]*
[1]Beijing Institute of Technology
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
{fangxingye, fuying}@bit.edu.cn, yang.yang@nlpr.ia.ac.cn

Abstract

Visible-infrared person re-identification (VI-ReID) focuses on matching the pedestrian images of the same identity captured by different modality cameras. The part-based methods achieve great success by extracting fine-grained features from feature maps. But most existing part-based methods employ horizontal division to obtain part features suffering from misalignment caused by irregular pedestrian movements. Moreover, most current methods use Euclidean or cosine distance of the output features to measure the similarity without considering the pedestrian relationships. Misaligned part features and naive inference methods both limit the performance of existing works. We propose a Semantic Alignment and Affinity Inference framework (SAAI), which aims to align latent semantic part features with the learnable prototypes and improve inference with affinity information. Specifically, we first propose semantic-aligned feature learning that employs the similarity between pixel-wise features and learnable prototypes to aggregate the latent semantic part features. Then, we devise an affinity inference module to optimize the inference with pedestrian relationships. Comprehensive experimental results conducted on the SYSU-MM01 and RegDB datasets demonstrate the favorable performance of our SAAI framework. Our code will be released at https://github.com/xiaoye-hhh/SAAI.

1. Introduction

Person re-identification (ReID) is the task of retrieving pedestrian images shot by different cameras. Most existing person ReID methods [12, 16, 17, 40] focus on matching the images shot by visible cameras, essentially addressing a single-modality pedestrian matching assignment. However, the generally visible surveillance cameras cannot capture pedestrian information well under poor illumination conditions. In response to this challenge, modern surveillance cameras can automatically switch to infrared mode for capturing images under low-light conditions. This technological advancement increases interest in researching visible-infrared person re-identification (VI-ReID).

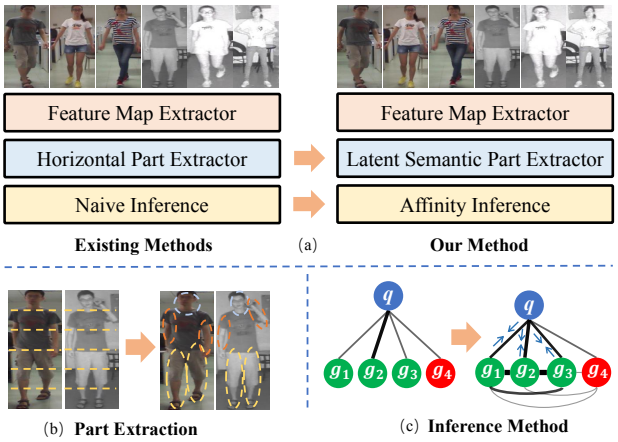


Figure 1. **Framework Differences.** (a) Most existing part-based methods extract horizontal part features from feature maps and employ naive inference (*e.g.*, Euclidean and cosine distance) to match images. Differently, our method extracts latent semantic part features and utilizes affinity information. (b) The part features extracted by our method have better semantic information. (c) Our affinity inference utilizes pedestrian relationships to optimize the distance calculation. Lines represent image relationships. Thicker lines indicate higher affinity. The green and red circles denote positive and negative gallery images, respectively.

VI-ReID suffers from substantial modality discrepancy and other factors (*e.g.*, viewpoints, backgrounds, and movements). The factors leading to a sizeable intra-class discrepancy make matching difficult. Figure 1 (a) exhibits that existing methods [13, 23, 24, 35] extract horizontal part features to alleviate this problem. Figure 1 (b) illustrates that pedestrian parts (*e.g.*, arms) are not located in a fixed position due to movements. Therefore, the simple horizontal partition by fixed height can cause part features to become

*Corresponding author

semantic misalignments, which limits the performance.

Moreover, Figure 1 (a) exhibits that most existing methods employ naive inference methods to calculate the similarity between query and gallery images. As shown in Figure 1 (c), these methods ignore the affinity information among gallery images by treating them as a single entity. The gallery images belong to the same modality without modality discrepancy, which can provide helpful affinity information to improve matching. Accordingly, SIM [14] proposes to use the similarity among the gallery images and calculate the Jaccard distance to boost matching. However, when calculating Jaccard distance, the element relationship is binary, inadequately utilizing affinity information.

We propose a Semantic Alignment and Affinity Inference framework (SAAI), which aligns latent semantic part features and better utilizes auxiliary affinity information. The SAAI framework consists of a semantic-aligned feature learning (SAFL) and an affinity inference module (AIM).

Specifically, SAFL first splits feature maps into pixel-wise features. Then, this method aggregates pixel-wise features with similar content by the similarity. This method concatenates the extracted latent semantic part features onto global features to provide local information. Finally, we use a dual-branch BNNeck to normalize features of both modalities, reducing the modality discrepancy. We devise a part diversity constraint to increase the diversity of latent semantic part features without additional annotations. Furthermore, we introduce a center separation loss to steer the network toward discerning pedestrian relationships.

Moreover, we propose the AIM to calculate the distance with the additional information from the affinity matrix. This module first calculates the query-gallery affinity matrix like most existing methods. Then, AIM calculates the gallery-gallery affinity matrix as references. Finally, AIM uses query-gallery and gallery-gallery affinity matrices to revise the distance measurement. AIM can optimize inference with affinity information among images. We propose a noise suppression algorithm to reduce the impact of inaccurate affinity values. Moreover, we devise a mean expansion method to increase the matching stability.

Our main contributions are summarized below:

- We propose an end-to-end Semantic Alignment and Affinity Inference framework (SAAI) for VI-ReID to explore the joint application of semantic-aligned feature learning and the affinity inference method.
- We propose a semantic-aligned feature learning (SAFL) to align the latent semantic part features. In addition, we devise an affinity inference module (AIM) to utilize pedestrian relationships for matching.
- Comprehensive experimental results demonstrate the effectiveness of the proposed framework. It achieves superior performance compared to state-of-the-art methods across various test settings.

2. Related Work

Single-Modality Person ReID. It is the task of retrieving pedestrian images captured by different visible cameras. According to different feature construction methods, current methods can be broadly categorized into two groups: hand-crafted feature construction methods [6, 9, 22, 39] and deep learning methods [4, 5, 18, 19, 21, 27, 28].

These methods achieve remarkable results in the single-modality person ReID. However, these methods only pay attention to visible images without considering the inherent modality differences between visible and infrared images. This limitation hampers the efficacy of these methods when applied to cross-modality pedestrian matching. Similar to our method, PAT [21] also adopts prototypes. However, PAT focuses on encoding pixel contexts and prototype relationships to alleviate occlusion issues. It is not suitable for cross-modality matching. Differently, our method extracts potential semantic part features shared by two modalities with shared prototypes to reduce the modality gap.

Visible-Infrared Person ReID. It pays attention to retrieving the pedestrian images shot by both visible and infrared modality cameras. Existing methods can be classified into two primary groups based on their diverse feature processing methods: generative and non-generative.

The generative methods focus on reducing differences in modality styles. Most methods use Generative Adversarial Networks (GANs) [8] to realize modality translation. Hi-CMD [2] and cmGAN [3] apply GANs to transfer different modality features into a shared space. In addition, AlignGAN [30] utilizes GANs to align the cross-modality features at both pixel and feature levels. FMCNet [37] employs GANs to realize feature-level modality compensation. These methods can reduce differences between visible and infrared styles. However, generative methods usually require extra computation and suffer from adjoint noise.

Differently, non-generative methods mainly focus on feature learning, extracting distinguishable features to bridge the cross-modality gap. Some works [10, 33, 36] rely solely on global features. They calculate the mean values of feature maps as output features for matching. These features lack local information, resulting in poor performance. To alleviate this problem, some methods typically use part features to enhance the features [13, 23, 24, 35]. MID [13] and MAUM [23] directly extract horizontal parts features as output features to increase the distinctiveness. DDAG [35] and cm-SSFT [24] further propose the attention mechanism to aggregate each horizontal part feature. With the attention mechanism, these two methods can dynamically evaluate the importance of different part features, which helps to capture more discriminative information. However, due to the varying and complex pedestrian movements, the horizontal part features can become misaligned, which can negatively impact the accuracy of these methods.

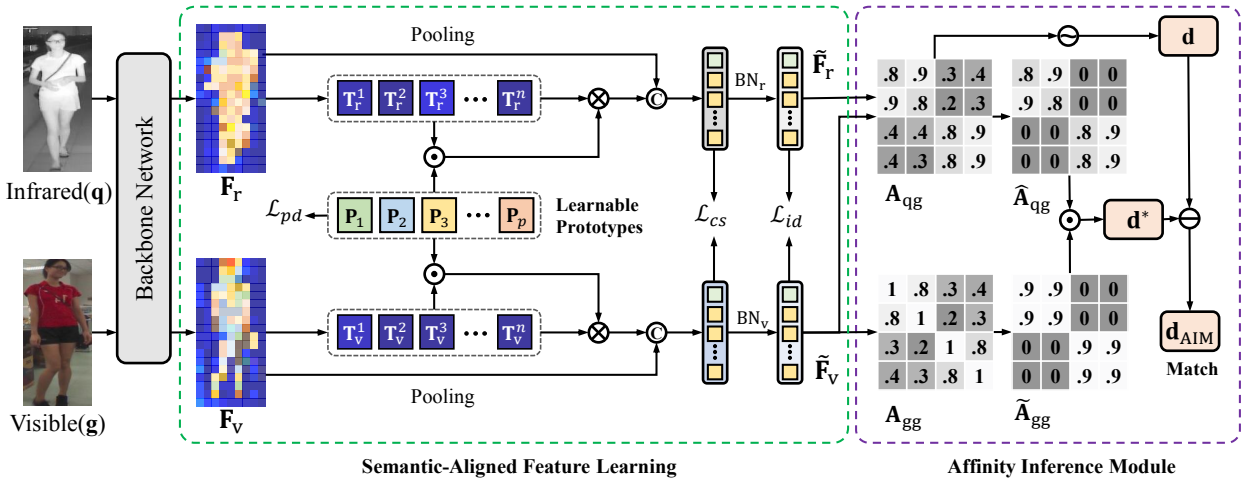


Figure 2. The framework of the proposed Semantic Alignment and Affinity Inference (SAAI). It describes the situation of querying visible images with infrared images. (1) The semantic-aligned feature learning (SAFL) first splits the feature maps $\mathbf{F}_\text{r}/\mathbf{F}_\text{v}$ into pixel-wise features $\mathbf{T}_\text{r}/\mathbf{T}_\text{v}$. Then, SAFL calculates the similarities between $\mathbf{T}_\text{r}/\mathbf{T}_\text{v}$ and learnable prototypes $\mathbf{P}$. SAFL clusters latent semantic part features with the similarity information. (2) The Affinity Inference Module (AIM) calculates the query-gallery affinity matrix $\mathbf{A}_\text{qg}$ and the gallery-gallery affinity matrix $\mathbf{A}_\text{gg}$. Then, AIM removes the noise values of $\mathbf{A}_\text{qg}$ to obtain $\hat{\mathbf{A}}_\text{qg}$. In addition, AIM removes noise values of $\mathbf{A}_\text{gg}$ and expands it with mean values to obtain $\tilde{\mathbf{A}}_\text{gg}$. Finally, AIM calculates the final distance with the affinity matrixes.

Differently, we design a semantic-aligned feature learning that dynamically extracts part features, mitigating the impact of irregular pedestrian movements.

Inference Methods for VI-ReID. Most existing methods [7, 13, 23, 33, 34, 36] use simple distance metrics for inference. Specifically, they calculate the Euclidean or cosine distance of the output features to measure the similarity. The Euclidean distance quantifies the spatial separation between two points in Euclidean space. The cosine distance gauges the similarity between two vectors based on the angle between them. These methods are intuitive and simple. But they treat each gallery image as a separate identity, ignoring the potential affinity information among the gallery images. This property can limit the matching performance, particularly in scenarios with multiple images of the same identity with variations in pose, viewpoint, or other factors.

In contrast, SIM [14] utilizes pedestrian relationships for inference. SIM calculates the Jaccard distance. However, the Jaccard distance only considers the presence or absence of elements in the gallery images and ignores the specific similarity score, which can limit the effectiveness of the approach. Therefore, we propose a novel affinity inference module that maps the affinity matrix into new features to assist the distance measurement. Compared to SIM, our method can more fully utilize information.

3. Methodology

In this section, we begin by defining the problem and presenting the overall framework. Subsequently, we elaborate on the design details of the Semantic-Aligned Feature Learning (SAFL) and Affinity Inference Module (AIM). Lastly, we provide the overall loss for SAAI.

3.1. Formulation and Overview

Problem Formulation. We use $\mathcal{Q} = \{\mathbf{q}_i | i = 1, 2, ..., N_\text{q}\}$ and $\mathcal{G} = \{\mathbf{g}_j | j = 1, 2, ..., N_\text{g}\}$ to represent the query set with N_q images and the gallery set with N_g images, respectively. The images in the query and gallery sets belong to different modalities. VI-ReID aims to establish correspondences between the query and the gallery images.

Overview. Figure 2 illustrates the proposed SAAI framework, which primarily comprises two modules. (1) The semantic-aligned feature learning is devised to extract the latent semantic part features by the similarity between pixel-wise features and learnable prototypes. The intuition is that a meaningful part should consist of features with similar content forming the part consistency. (2) The affinity inference module is proposed to utilize pedestrian relationships to optimize the inference. When a query image $\mathbf{q}$ exhibits a high similarity to a gallery image $\mathbf{g}$, it is suitable to diminish the distance between $\mathbf{q}$ and other images resembling $\mathbf{g}$. Our method simulates the above process with the affinity matrix to revise the distance calculation.

3.2. Semantic-Aligned Feature Learning

To alleviate the problem of part feature misalignment caused by pedestrian movements, we propose the semantic-aligned feature learning. It uses learnable prototypes to extract latent semantic part features from feature maps $\mathbf{F}_\text{r} \in$

$\mathbb{R}^{h\times w\times c}$ and $\mathbf{F}_{\mathrm{v}} \in \mathbb{R}^{h\times w\times c}$, where h, w, c denote the height, weight, and dimension of feature maps.

Specifically, when extracting latent semantic part features from feature maps, we should give higher weights to the pixel-wise features similar to target parts. This method can form part consistency, which ensures part features are consistent and reliable. But we do not have reference features of each part. Therefore, we propose to utilize learnable prototypes $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_p] \in \mathbb{R}^{p\times c}$ as references of latent semantic part features, where $\mathbf{P}_i$ represents the i-th learnable prototype and p denotes the total number of such learnable prototypes. We utilize the learnable prototypes to learn appropriate latent semantic part features as references. Each learnable prototype aims to cluster unique latent semantic part features from the feature maps.

In order to achieve semantic alignment of part features between two modalities, we employ shared prototypes $\mathbf{P}$ for both the visible and infrared modalities. The shared prototypes enable the aggregation of latent semantic part features with the same semantic information across both modalities into the same parts. This design can achieve the modality alignment at the latent semantic part level, reducing the modality discrepancy. When extracting latent semantic part features, the pixel-wise features of the two modalities will be compared to the shared learnable prototypes to determine their weighted contribution to each latent semantic part feature. For the convenience of description, we take visible features as an example to describe details.

We first divide the feature map $\mathbf{F}_{\mathrm{v}}$ into pixel-wise features $\mathbf{T}_{\mathrm{v}} = [\mathbf{T}_{\mathrm{v}}^1, \mathbf{T}_{\mathrm{v}}^2, ..., \mathbf{T}_{\mathrm{v}}^n] \in \mathbb{R}^{n\times c}$, where $\mathbf{T}_{\mathrm{v}}^i$ represents the i-th pixel-wise feature and $n = h \times w$ represents the quantity of pixel-wise features. We add learnable position embeddings to each pixel-wise feature according to the position to provide spatial information. The spatial information can enhance the spatial stability of our latent semantic part features. Subsequently, this module calculates the similarity matrix $\mathbf{S}_{\mathrm{v}} \in \mathbb{R}^{p\times n}$ between $\mathbf{P}$ and $\mathbf{T}_{\mathrm{v}}$ as:

$$\mathbf{S}_{\mathrm{v}} = \sigma(\mathbf{P} \odot \mathbf{T}_{\mathrm{v}}^{\mathsf{T}}), \tag{1}$$

where $\odot$ represents the matrix multiplication and $\sigma(\cdot)$ denotes the Sigmoid activation function. The Sigmoid activation function is designed to enhance numerical stability.

Taking learnable prototypes $\mathbf{P}$ as references, we obtain the similarity matrix $\mathbf{S}_{\mathrm{v}}$ to describe the affinity between $\mathbf{T}_{\mathrm{v}}$ and $\mathbf{P}$. A high similarity score between the pixel-wise feature and a learnable prototype indicates that the pixel-wise feature probably belongs to that latent semantic part. Therefore, we can utilize $\mathbf{S}_{\mathrm{v}}$ as weight to aggregate pixel-wise features to form latent semantic part feature $\mathbf{p}_{\mathrm{v}}^i$ as:

$$\mathbf{p}_{\mathrm{v}}^i = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{S}_{\mathrm{v}}^{ij} \otimes \mathbf{T}_{\mathrm{v}}^j) \quad (i = 1, 2, ..., p), \tag{2}$$

where $\otimes$ denotes element-wise multiplication, and $\mathbf{S}_{\mathrm{v}}^{ij}$ indicates the similarity score between $\mathbf{P}_i$ and $\mathbf{T}_{\mathrm{v}}^j$.

The latent semantic part features $\mathbf{p}_{\mathrm{v}}$ contain local information of pedestrians. We just concatenate $\mathbf{p}_{\mathrm{v}}$ with global features to obtain augmented features $\hat{\mathbf{F}}_{\mathrm{v}} \in \mathbb{R}^{(p+1)c}$ as:

$$\hat{\mathbf{F}}_{\mathrm{v}} = \left[\mathbf{p}_{\mathrm{v}}^1, \mathbf{p}_{\mathrm{v}}^2, ..., \mathbf{p}_{\mathrm{v}}^p, \mathrm{avg}(\mathbf{F}_{\mathrm{v}})\right], \tag{3}$$

where $[\cdot]$ denotes concatenating the features, $\mathrm{avg}(\cdot)$ denotes average pooling, and $\mathrm{avg}(\mathbf{F}_{\mathrm{v}})$ indicates the global feature. $\hat{\mathbf{F}}_{\mathrm{v}}$ contains both local and global information. Similarly, we can construct $\hat{\mathbf{F}}_{\mathrm{r}}$ from feature maps $\mathbf{F}_{\mathrm{r}}$ in the same way. Both $\hat{\mathbf{F}}_{\mathrm{v}}$ and $\hat{\mathbf{F}}_{\mathrm{r}}$ are extracted by the shared learnable prototypes $\mathbf{P}$. Hence, $\hat{\mathbf{F}}_{\mathrm{v}}$ and $\hat{\mathbf{F}}_{\mathrm{r}}$ can achieve alignment at the latent semantic part level. To mitigate the modality gap, we devise a dual-branch BNNeck to normalize the $\hat{\mathbf{F}}_{\mathrm{v}}$ and $\hat{\mathbf{F}}_{\mathrm{r}}$, respectively. The outputs are $\tilde{\mathbf{F}}_{\mathrm{v}}$ and $\tilde{\mathbf{F}}_{\mathrm{r}}$.

We utilize a classification loss $\mathcal{L}_{id}$ to steer the model to learn identity information. Further details are provided in the supplementary materials. In addition, we devise a part diversity loss $\mathcal{L}_{pd}$ to increase the latent semantic part diversity. Moreover, we propose a center separation loss $\mathcal{L}_{cs}$ to train the network to distinguish different pedestrians.

Part Diversity Loss. Extracting various features is the key to semantic-aligned feature learning. We should train each prototype to extract different part features. However, we do not have the human part annotations to train. We propose a simplified alternative strategy. We encourage the learnable prototypes to focus on different areas of pedestrians. Follow this idea, we propose part diversity loss $\mathcal{L}_{pd}$ as:

$$\mathcal{L}_{pd} = -\frac{2}{p(p-1)}\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}||\mathbf{P}_i\mathbf{T}^{\mathsf{T}} - \mathbf{P}_j\mathbf{T}^{\mathsf{T}}||_2 \tag{4}$$

where $\mathbf{P}_i$ and $\mathbf{P}_j$ denote i-th and j-th learnable prototypes, and $\mathbf{T}$ denotes $\mathbf{T}_{\mathrm{r}}$ and $\mathbf{T}_{\mathrm{v}}$ for convenience of description.

Center Separation Loss. We propose a center separation loss $\mathcal{L}_{cs}$ to guide the network to identify the pedestrian relationships. $\mathcal{L}_{cs}$ aims to bring the samples with the same identity closer to each other while pushing the instance centers belonging to different identities apart. In order to reduce the network overfitting problem and increase the feature diversity, we only gather samples to the range ρ_1 of their respective center. The $\mathcal{L}_{cs}$ can be represented as:

$$\begin{aligned}\mathcal{L}_{cs} = &\frac{1}{N}\sum_{i=1}^{N}[-\rho_1 + ||\hat{\mathbf{F}}_i - \mathbf{c}_{y_i}||_2]_+ \\ &+ \frac{2}{M(M-1)}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}[\rho_2 - ||\mathbf{c}_{y_j} - \mathbf{c}_{y_k}||_2]_+,\end{aligned} \tag{5}$$

where N denotes the batch size, $\hat{\mathbf{F}}_i$ denotes the i-th feature, y_i represents the i-th label, $\mathbf{c}_{y_i}$ represents the center of y_i, M represents number of centers, ρ_1 is the margin between samples to centers, and ρ_2 is the margin among centers.

3.3. Affinity Inference Module

The effectiveness of matching pedestrian images based on Euclidean or cosine distance is limited. These methods overlook the affinity information and treat gallery images as distinct entities during the inference process. Therefore, we propose the affinity inference module (AIM), which utilizes pedestrian relationships to revise distances.

The intuition behind AIM is that we can leverage gallery images with high affinity to the query image to revise distances. AIM can capture the potential affinity information among the gallery images and incorporate it into the distance calculation, optimizing the matching performance. Algorithm 1 outlines the primary steps of the AIM.

Algorithm 1 Framework of AIM

Input: query features $\mathbf{F}_{\mathrm{q}}$, gallery features $\mathbf{F}_{\mathrm{g}}$, hyper-parameters k_1 and k_2

Output: $\mathbf{d}_{\mathrm{AIM}}$

1: Calculate the query-gallery affinity matrix $\mathbf{A}_{\mathrm{qg}}$
2: Calculate the gallery-gallery affinity matrix $\mathbf{A}_{\mathrm{gg}}$
3: Remove the noise values in $\mathbf{A}_{\mathrm{qg}}$ and $\mathbf{A}_{\mathrm{gg}}$ according to k_1 to obtain $\hat{\mathbf{A}}_{\mathrm{qg}}$ and $\hat{\mathbf{A}}_{\mathrm{gg}}$
4: Expand $\hat{\mathbf{A}}_{\mathrm{gg}}$ according to k_2 to obtain $\tilde{\mathbf{A}}_{\mathrm{gg}}$
5: Calculate the base distance $\mathbf{d}$ as Eq. (8)
6: Calculate the amended distance $\mathbf{d}^*$ as Eq. (9)
7: Calculate the final distance $\mathbf{d}_{\mathrm{AIM}}$ as Eq. (10)
8: **return** $\mathbf{d}_{\mathrm{AIM}}$

Calculating Affinity Matrix. The framework transfers the query image $\mathbf{q}$ and gallery image $\mathbf{g}$ into to features $\mathbf{F}_{\mathrm{q}}$ and $\mathbf{F}_{\mathrm{g}}$, respectively. AIM first calculates the query-gallery affinity matrix $\mathbf{A}_{\mathrm{qg}} \in \mathbb{R}^{N_{\mathrm{q}} \times N_{\mathrm{g}}}$ by cosine similarity between $\mathbf{F}_{\mathrm{q}}$ and $\mathbf{F}_{\mathrm{g}}$. Similarly, AIM can also calculate the gallery-gallery affinity matrix $\mathbf{A}_{\mathrm{gg}} \in \mathbb{R}^{N_{\mathrm{g}} \times N_{\mathrm{g}}}$ from all gallery image pairs. Unlike most existing methods that directly use $\mathbf{A}_{\mathrm{qg}}$ to match pedestrian images, our method utilizes $\mathbf{A}_{\mathrm{qg}}$ and $\mathbf{A}_{\mathrm{gg}}$ to revise distances before matching.

Removing Noise Values. $\mathbf{A}_{\mathrm{qg}}$ and $\mathbf{A}_{\mathrm{gg}}$ contain rich affinity information. However, there are noise values in matrices. $\mathcal{L}_{cs}$ encourages images of the same identity to be clustered together while dispersing centers corresponding to distinct identities. Instances with different labels are separated, and the affinity information between them is too weak to use. These noise values can mislead our following distance calculation, resulting in inaccurate results.

To address this issue, we propose an effective noise suppression strategy. These noise values are numerically small. Based on this property, we can remove noise values by clearing small values in the affinity matrix. For the convenience of description, we take $\mathbf{A}_{\mathrm{qg}}$ as an example. We first identify the k_1-th largest value v in each row of $\mathbf{A}_{\mathrm{qg}}$ and set any value less than v to 0 as:

$$\hat{\mathbf{A}}_{\mathrm{qg}}^{ij} = \begin{cases} \mathbf{A}_{\mathrm{qg}}^{ij} & \text{if } \mathbf{A}_{\mathrm{qg}}^{ij} \geq v_i \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $\hat{\mathbf{A}}_{\mathrm{qg}}^{ij}$ denotes the output removing the noise value, $\mathbf{A}_{\mathrm{qg}}^{ij}$ denotes the element of $\mathbf{A}_{\mathrm{qg}}$ on row i and column j, and v_i denote the k_1-th largest value in row i of $\mathbf{A}_{\mathrm{qg}}$. Similarly, AIM can remove noise values of $\mathbf{A}_{\mathrm{gg}}$ to obtain $\hat{\mathbf{A}}_{\mathrm{gg}}$.

Expanding Representation. Replacing the current value with the average value of the most similar neighbors can provide a more stable representation of the affinity information. $\hat{\mathbf{A}}_{\mathrm{gg}}$ contains the affinity information among gallery images, which can be used to expand the representation. Specifically, given a gallery image $\mathbf{g}_i$, we can find the k_2 gallery images $\mathcal{I}_i$ that are most similar to $\mathbf{g}_i$. Then, we can expand the original affinity information with the average affinity score of the k_2 gallery images as:

$$\tilde{\mathbf{A}}_{\mathrm{gg}}^{i} = \frac{1}{k_2} \sum_{l \in \mathcal{I}_i} \hat{\mathbf{A}}_{\mathrm{gg}}^{l}, \tag{7}$$

where $\tilde{\mathbf{A}}_{\mathrm{gg}}^{i}$ denotes the i-th row of the output, $\hat{\mathbf{A}}_{\mathrm{gg}}^{l}$ denotes the affinity score of the l-th most similar neighbours.

Final Distance. $\mathbf{A}_{\mathrm{qg}}$ presents the cosine similarity between query and gallery images. We can covert the cosine similarity to base distance $\mathbf{d}$ with $\mathbf{A}_{\mathrm{qg}}$ as:

$$\mathbf{d} = 1 - \mathbf{A}_{\mathrm{qg}}. \tag{8}$$

In addition, AIM utilizes $\hat{\mathbf{A}}_{\mathrm{qg}}$ and $\tilde{\mathbf{A}}_{\mathrm{gg}}$ to assist the distance measurement. If the query $\mathbf{q}_i$ image is similar to a gallery image $\mathbf{g}_j$, this module reduces the distance between $\mathbf{q}_i$ and gallery images being similar to $\mathbf{g}_j$. The reduced distances depend on the affinity between these images and $\mathbf{q}_i$/$\mathbf{g}_j$. The amended distance $\mathbf{d}^*$ can be calculated as:

$$\mathbf{d}^* = \hat{\mathbf{A}}_{\mathrm{qg}} \tilde{\mathbf{A}}_{\mathrm{gg}}. \tag{9}$$

Finally, We can subtract $\mathbf{d}^*$ from $\mathbf{d}$ to revise the distances. The final distance $\mathbf{d}_{\mathrm{AIM}}$ can be calculated as:

$$\mathbf{d}_{\mathrm{AIM}} = \mathbf{d} - \mathbf{d}^*. \tag{10}$$

3.4. Training and Inference

SAAI is trained by minimizing:

$$\mathcal{L} = \mathcal{L}_{id} + \mathcal{L}_{cs} + \lambda \mathcal{L}_{pd}, \tag{11}$$

where λ is a hype-parameter to balance the loss items. AIM only works in the inference stage, which utilizes affinity information to assist in distance measurement.

Table 1. Comparison with the state-of-the-art methods on SYSU-MM01. The comparison indicators are CMC (%) and mAP (%).

Method	All-Search				Indoor-Search			
	Single-Shot		Multi-Shot		Single-Shot		Multi-Shot	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Zero-Padding [33]	14.80	15.95	19.13	10.89	20.58	26.92	24.43	18.86
cmGAN [3]	26.97	27.80	31.49	22.27	31.63	42.19	37.00	32.76
JSIA-ReID [31]	38.10	36.90	45.10	29.50	43.80	52.90	52.70	42.70
AlignGAN [30]	42.40	40.70	51.50	33.90	45.90	54.30	57.10	45.30
AGW [36]	47.50	47.65	-	-	54.17	62.97	-	-
LbA [26]	55.41	54.14	-	-	58.46	66.33	-	-
NFS [1]	56.91	55.45	63.51	48.56	62.79	69.79	70.03	61.45
MID [13]	60.27	59.40	-	-	64.86	70.12	-	-
cm-SSFT [24]	61.60	63.20	63.40	62.00	70.50	72.60	73.00	72.40
CM-NAS [7]	61.99	60.02	68.68	53.45	67.01	72.95	76.48	65.11
MCLNet [10]	65.40	61.98	-	-	72.56	76.58	-	-
FMCNet [37]	66.34	62.51	73.44	56.06	68.15	74.09	78.86	63.82
SMCL [32]	67.39	61.78	72.15	54.93	68.84	75.56	79.57	66.57
MPANet [34]	70.58	68.24	75.58	62.91	76.74	80.95	84.22	75.11
MAUM [23]	71.68	68.79	-	-	76.97	81.94	-	-
CMT [15]	71.88	68.57	80.23	63.13	76.90	79.91	84.87	74.11
CIFT [20]	74.08	74.79	79.74	75.56	81.82	85.61	88.32	86.42
MSCLNet [38]	**76.99**	71.64	-	-	78.49	81.17	-	-
SAAI(ours)	75.90	**77.03**	**82.86**	**82.39**	**83.20**	**88.01**	**90.73**	**91.30**

4. Experiments

In this section, we first introduce datasets and experiment implementation. Then, we conduct experiments on two public datasets. Finally, we analyze SAAI carefully.

4.1. Datasets and Settings

SYSU-MM01 [33] is the pioneering large-scale benchmark dataset in the VI-ReID domain. It has a total of 287,628 visible and 15,792 infrared images. It records 491 pedestrians with four visible cameras and two infrared cameras. The capture environment encompasses both indoor and outdoor settings. It has multiple evaluation modes based on different configurations. The modes are categorized into all-search and indoor-search settings based on whether including outdoor-captured images. Besides, based on the varying number of images within the gallery, the modes can be further subdivided into single-shot and multi-shot scenarios.

RegDB [25] contains 8,240 images of 412 pedestrians. All images are collected by a visible camera and an infrared camera. Ten infrared images and ten visible images were taken for each pedestrian. RegDB is randomly split into two non-overlapping sets based on the pedestrian identities: one set is utilized for training purposes, while the other serves as the testing set. It contains two evaluation settings: Visible2Infrared and Infrared2Visible. The first setting denotes retrieving infrared images according to visible ones, and the second setting entails the reverse operation.

Metrics. The experiments adhere to the standard evaluation settings, encompassing two fundamental assessment metrics: the Cumulative Matching Characteristic (CMC) and the Mean Average Precision (mAP).

Implementation Details. The SAAI framework is realized using the PyTorch framework and executed on a single RTX3090 GPU. We employ the ResNet-50 [11] as the backbone. For each batch, we randomly sample 16 identities and each identity contains 8 images. The input images are initially resized to a consistent dimension of 288×144. Then, we apply a series of augmentation techniques, including random cropping, random erasing, random horizontal flipping, and random grayscale. The network is optimized by Adam with a linear warmup strategy. The initial learning rate is set to 3.5×10^{-4} and is decreased by factors of 0.1 and 0.01 at 80 and 120 epochs, respectively. The training procedure spans a total of 160 epochs.

λ in Eq. (11) is set to 0.5. The margin parameters ρ_1 and ρ_2 are set to 0.01 and 0.7, respectively. The number of learnable prototypes p is set to 7. For SYSU-MM01, k_1/k_2 are set to 4/1 under the single-shot setting and 20/6 under the multi-shot setting. For RegDB, k_1/k_2 are set to 8/2.

4.2. Comparison with State-of-the-art Methods

We evaluate our SAAI framework on SYSU-MM01 and RegDB. We compare the proposed SAAI with numerous state-of-the-art (SOTA) methods, including four generative methods [3, 30, 31, 37] and fourteen non-generative methods [1, 7, 13, 15, 20, 23, 24, 26, 32, 33, 34, 36, 37, 38].

Table 2. Comparison with the state-of-the-art methods on RegDB. The comparison indicators are CMC (%) and mAP (%).

Method	Visible2Infrared		Infrared2Visible	
	Rank-1	mAP	Rank-1	mAP
Zero-Padding [33]	17.75	18.90	16.63	17.82
JSIA-ReID [31]	48.50	49.30	48.10	48.90
AlignGAN [30]	57.90	53.60	56.30	53.40
AGW [36]	70.05	66.37	70.49	65.90
cm-SSFT [24]	72.30	72.90	71.00	71.70
LbA [26]	74.17	67.64	72.43	65.46
MCLNet [10]	80.31	73.07	75.93	69.49
NFS [1]	80.54	72.10	77.95	69.79
MPANet [34]	83.70	80.90	82.80	80.70
SMCL [32]	83.93	79.83	83.05	78.57
MSCLNet [38]	84.17	80.99	83.86	78.31
CM-NAS [7]	84.54	80.32	82.57	78.31
MID [13]	87.45	84.85	84.29	81.41
MAUM [23]	87.87	85.09	86.95	84.34
FMCNet [37]	89.12	84.43	88.38	83.86
CIFT [20]	91.96	**92.00**	90.30	90.78
CMT [15]	**95.17**	87.30	91.97	84.46
SAAI(ours)	91.07	91.45	**92.09**	**92.01**

Comparison on SYSU-MM01. As shown in Table 1, our model outperforms SOTAs in mAP under all test settings. Our model achieves 77.03% in mAP, improving the mAP by 2.24% over the best SOTA (CIFT) under the all-search and single-shot setting. In addition, our model also achieves considerable results in Rank-1. Under the indoor-search and single-shot setting, our model achieves 83.20% in Rank-1, outperforming the best SOTA (CIFT) by 1.38%.

Comparisons on RegDB. We evaluate our model in a smaller dataset as Table 2. Under the infrared2visible setting, our model achieves a Rank-1 of 92.09% and an mAP of 92.01%. When switching to the visible2infrared setting, our method is slightly inferior to SOTAs. The reason could be attributed to the limited size of RegDB, which may result in incomplete training of the learnable prototypes.

4.3. Ablation Study

We adopt the ResNet-50 with horizontal part features as the backbone of the baseline. The baseline is trained by $\mathcal{L}_{id}$. We keep other settings consistent with our method.

Table 3 shows the performance improvements achieved by SAFL and AIM. SAFL enhances Rank-1 and mAP by 8.24% and 10.10%, respectively. AIM improves Rank-1 and mAP by 1.09% and 6.19%, respectively. When combined, SAFL and AIM further boost the overall performance, proving the effectiveness of SAFL and AIM.

Table 4 illustrates the impact of L_{dp} and L_{cs}. L_{dp} results in a Rank-1 improvement of 0.43% and an mAP improvement of 0.39%. On the other hand, L_{cs} boosts the Rank-1 by 2.90% and the mAP by 3.97%. Moreover, the combined usage of L_{dp} and L_{cs} further enhances the overall performance, proving of the effectiveness of these losses.

Table 3. Analysis of the proposed SAFL and AIM on SYSU-MM01 under the all-search and single-shot mode.

Base	SAFL	AIM	Rank-1	mAP
✓	✗	✗	66.79	61.59
✓	✓	✗	75.03	71.69
✓	✗	✓	67.88	67.78
✓	✓	✓	**75.90**	**77.03**

Table 4. Analysis of the proposed L_{dp} and L_{cs} on SYSU-MM01 under the all-search and single-shot.

L_{id}	L_{dp}	L_{cs}	Rank-1	mAP
✓	✗	✗	71.81	72.15
✓	✓	✗	72.24	72.54
✓	✗	✓	74.71	76.12
✓	✓	✓	**75.90**	**77.03**

As an inference module, AIM can be added to other methods. To further prove the effectiveness of AIM, we add AIM to three methods, including one generative method [30] and two non-generative methods [34, 36]. As shown in Table 5, AIM works well in both generative method and non-generative methods. AIM improves both Rank-1 and mAP of these methods. These results further prove the effectiveness of the affinity inference module.

Table 5. Analysis of AIM on other methods on SYSU-MM01 under the multi-shot setting. We retrain the models.

Method	all-search		indoor-search	
	Rank-1	mAP	Rank-1	mAP
AlignGAN* [30]	48.31	34.47	57.75	45.62
AlignGAN*+AIM	**51.63**	**50.65**	**58.32**	**61.94**
AGW* [36]	52.47	41.48	60.59	54.31
AGW*+AIM	**55.01**	**55.18**	**63.84**	**67.51**
MPANet* [34]	77.14	62.92	84.64	75.41
MPANet*+AIM	**78.52**	**78.27**	**85.96**	**87.31**

4.4. Model Analysis

Parameters Analysis. We first evaluate the effect of λ in Eq. (11) on SYSU-MM01 dataset under the all-search and single-shot mode. In Figure 3, we show results of Rank-1 and mAP of different λ. The most suitable parameter setting is 0.5. Then, we evaluate the number of learnable prototypes p. As shown in Figure 4, before p reaches 7, the performance increases with the increase of p. A larger p allows the model to pay attention to more parts. However, once p exceeds the required value. It leads to competition among learnable prototypes, affecting the ability to discover latent semantic parts and reducing performance.

Visualization Analysis. We utilize t-SNE [29] to visualize the feature distribution of baseline and SAAI. As shown in Figure 5 (a), there are large modality discrepancies in the

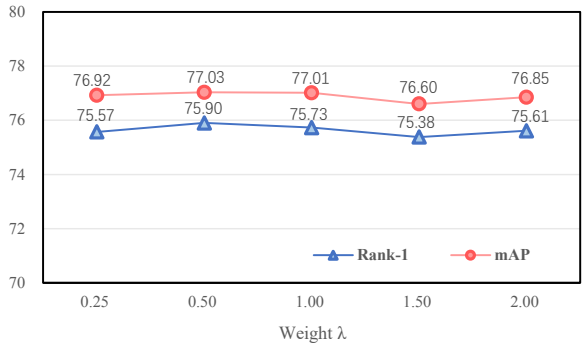


Figure 3. The sensitive graph of the weight λ in Eq. (11).

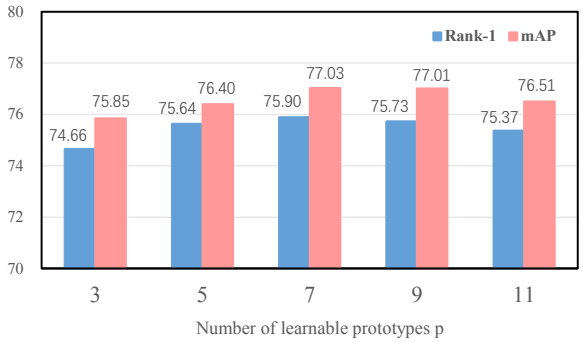
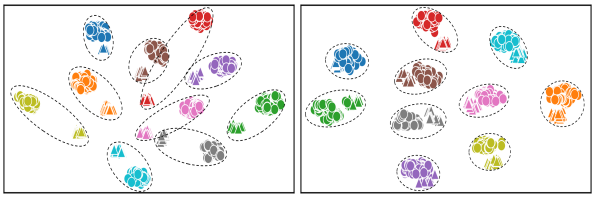


Figure 4. The impact of the number of learnable prototypes p.



(a) (b)

Figure 5. The results of the features distribution on SYSU-MM01. Colors represent identities. Circles represent visible features. Triangles represent infrared features. (a) displays the horizontal part features. (b) displays the features extracted by SAAI.

features extracted by the baseline. In contrast, as shown in Figure 5 (b), the discrepancy among the features extracted by SAAI is much smaller. The images are closely clustered by identities. This proves the effectiveness SAAI.

We visualize the attention maps of baseline and SAAI. As shown in Figure 6, SAAI focus more accurately on the main body parts than the baseline. The results show the capacity of SAFL to effectively extract latent semantic parts, resulting in improved feature localization and alignment.

Table 6. Evaluation of removing and expanding steps on SYSU-MM01. AIM$^-$ has no removing and expanding.

AIM$^-$	Removing	Expanding	Rank-1	mAP
✓	✗	✗	79.58	67.69
✓	✓	✗	82.72	79.67
✓	✗	✓	80.58	74.27
✓	✓	✓	**82.86**	**82.39**

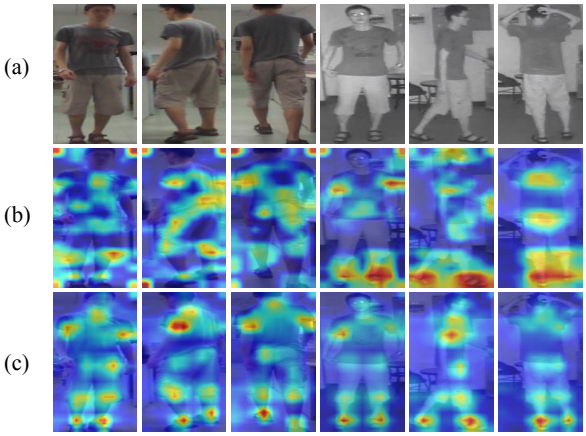


Figure 6. The visual comparison of attention maps. (a) displays raw images. (b) and (c) show results of baseline and SAAI.

Analysis of AIM. As shown in Table 6, the removing step is pivotal in mitigating the impact of noise values in matrices, leading to the improvements in both Rank-1 and mAP, respectively. Additionally, the expanding step contributes to a more stable representation and can also boost the matching performance. When removing and expanding steps are combined, this module achieves the best performance. These results strongly demonstrate the effectiveness of removing and expanding steps in AIM.

5. Conclusion

We propose the Semantic Alignment and Affinity Inference framework (SAAI) for visible-infrared person ReID. We first devise a semantic-aligned feature learning process that aligns latent semantic parts by considering the similarity between pixel-level features and learnable prototypes. This method effectively reduces the impact of pedestrian movements and enhances the distinctiveness of the feature representation. Furthermore, we present an affinity inference module that leverages pedestrian relationships to revise distance calculations. Through extensive experiments conducted on the SYSU-MM01 and RegDB datasets, our framework exhibits substantial efficacy for VI-ReID.

Limitation and future work. Our framework only leverages pedestrian relationships during the inference stage. Nevertheless, affinity information remains untapped during the training phase, which hinders the advancement of the training process. Our future work is combining our affinity inference method with network training.

Acknowledgments. This work was supported by the National Key R&D Program of China (2022YFC3300700), the National Natural Science Foundation of China (62206276, 62171038, 61931008, 62171042, and U21B2024), the R&D Program of Beijing Municipal Education Commission (KZ202211417048), and the Fundamental Research Funds for the Central Universities.

References

[1] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for rgb-infrared person re-identification. In *CVPR*, pages 587–597, 2021. 6, 7

[2] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*, pages 10257–10266, 2020. 2

[3] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, pages 677–683, 2018. 2, 6

[4] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Pan Ji, Lars Petersson, and Mehrtash Harandi. Attention in attention networks for person retrieval. *TPAMI*, 44(9):4626–4641, 2021. 2

[5] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *ICCV*, pages 8030–8039, 2019. 2

[6] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010. 2

[7] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *ICCV*, pages 11823–11832, 2021. 3, 6, 7

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[9] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275, 2008. 2

[10] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *ICCV*, pages 16403–16412, 2021. 2, 6, 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[12] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021. 1

[13] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In *AAAI*, pages 1034–1042, 2022. 1, 2, 3, 6, 7

[14] Mengxi Jia, Yunpeng Zhai, Shijian Lu, Siwei Ma, and Jian Zhang. A similarity inference metric for rgb-infrared cross-modality person re-identification. In *IJCAI*, page 1026–1032, 2020. 2, 3

[15] Kongzhu Jiang, Tianzhu Zhang, Xiang Liu, Bingqiao Qian, Yongdong Zhang, and Feng Wu. Cross-modality transformer for visible-infrared person re-identification. In *ECCV*, pages 480–496, 2022. 6, 7

[16] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, pages 3143–3152, 2020. 1

[17] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, pages 1062–1071, 2018. 1

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 2

[19] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 2

[20] Xulin Li, Yan Lu, Bin Liu, Yating Liu, Guojun Yin, Qi Chu, Jinyang Huang, Feng Zhu, Rui Zhao, and Nenghai Yu. Counterfactual intervention feature transfer for visible-infrared person re-identification. In *ECCV*, pages 381–398, 2022. 6, 7

[21] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *CVPR*, pages 2898–2907, 2021. 2

[22] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 2

[23] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *CVPR*, pages 19366–19375, 2022. 1, 2, 3, 6, 7

[24] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, pages 13379–13389, 2020. 1, 2, 6, 7

[25] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 6

[26] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *ICCV*, pages 12046–12055, 2021. 6, 7

[27] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *ICCV*, pages 1025–1034, 2021. 2

[28] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 2

[29] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11):2579–2605, 2008. 7

[30] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, pages 3623–3632, 2019. 2, 6, 7

[31] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *AAAI*, pages 12144–12151, 2020. 6, 7

[32] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *ICCV*, pages 225–234, 2021. 6, 7

[33] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017. 2, 3, 6, 7

[34] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *CVPR*, pages 4330–4339, 2021. 3, 6, 7

[35] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, pages 229–247, 2020. 1, 2

[36] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 44(6):2872–2893, 2021. 2, 3, 6, 7

[37] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *CVPR*, pages 7349–7358, 2022. 2, 6, 7

[38] Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In *ECCV*, pages 462–479, 2022. 6, 7

[39] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 2

[40] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019. 1