Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Richer Semantics, Better Alignment: Aligning Visual Features with Explicit and Enriched Semantics for Visible-Infrared Person Re-Identification

Neng Dong[1] , **Shuanglin Yan**[1] , **Liyan Zhang**[2]* and **Jinhui Tang**[1]
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology
[2]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronauticsn
{neng.dong, shuanglinyan, jinhuitang}@njust.edu.cn, zhangliyan@nuaa.edu.cn

Abstract

Visible-infrared person re-identification (VIReID) retrieves pedestrian images with the same identity across different modalities. Existing methods learn visual features solely from images, failing to align them into the modality-invariant semantic space. In this paper, we propose a novel framework, termed Richer Semantics, Better Alignment (RSBA), to align visual features with explicit and enriched semantics. Specifically, we first develop an Explicit Semantics-Guided Feature Alignment (ESFA) module, which supplements textual descriptions for cross-modality images and aligns image-text pairs within each modality, alleviating the distribution discrepancy of visual features. We then devise a Consistent Similarity-Guided Indirect Alignment (CSIA) module, which constrains the similarity between intra-modality image-text pairs to be consistent with that between inter-modality text-text pairs, indirectly aligning visual features with cross-modality semantics. Furthermore, we design a Cross-View Semantics Compensation (CVSC) module, which integrates multi-view texts and improves the image-text matching of one-to-one in ESFA and CSIA to one-to-many, further strengthening the alignment of visual features within the semantic space. Extensive experimental results on three public datasets demonstrate the effectiveness and superiority of our proposed RSBA.

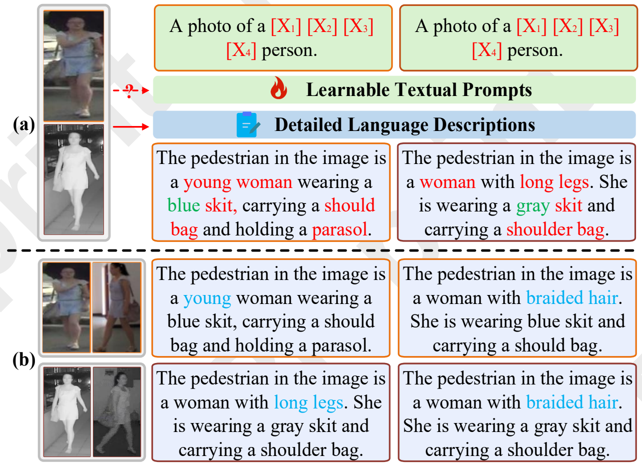

Figure 1: The core motivation of our RSBA framework: (i) Explicit semantics (red) in language descriptions generated by LLaVA enable the more effective alignment of visual features than learnable textual prompts. (ii) The conflicting semantics (green) make the alignment of images to inter-modality texts challenging. (iii) Multi-view texts provide complementary semantics (blue) that play a positive role in further enhancing the modality-invariance of visual features.

1 Introduction

Person Re-Identification (ReID) [Ye *et al.*, 2021b; Yan *et al.*, 2023a; Dong *et al.*, 2024b] aims to match images of the same individual across cameras, a critical component of intelligent security with profound research implications. Despite significant advancements [Li *et al.*, 2021; Yan *et al.*, 2022; Gong *et al.*, 2022; Dong *et al.*, 2024a], most existing algorithms focus solely on visible image retrieval, failing to meet the demands of 24-hour surveillance systems, which must also retrieve infrared images captured at night. To overcome this limitation, visible-infrared person ReID (VIReID) [Wu *et al.*, 2017] has been proposed to retrieve visible images that match the identity of a given infrared query, and vice versa.

The primary challenge in VIReID lies in aligning the feature distribution of cross-modality images, for which two main approaches have been developed. The first is generative-based methods [Dai *et al.*, 2018; Choi *et al.*, 2020; Miao *et al.*, 2021], which transfer the style of images to another modality. However, these algorithms often introduce noise during the generation process, compromising feature discriminability. The second approach, generative-free methods [Ling *et al.*, 2021; Ye *et al.*, 2021a; Li *et al.*, 2022], focuses on optimizing network structures and metric functions. Comparatively, the latter has demonstrated greater effectiveness and currently stands as the predominant solution. However, the large modality discrepancy makes it challenging to align heterogeneous features into a suitable common space.

To address this limitation, a recent study [Yu *et al.*, 2025] incorporates Contrastive Language-Image Pre-training (CLIP) [Radford *et al.*, 2021] into VIReID, demonstrating

*Corresponding author

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

that semantics represented by language descriptions of heterogeneous images exhibit no modality gap, thus aligning visual features into the semantic space is beneficial for alleviating their distribution discrepancy. However, pedestrian images typically lack accompanying language descriptions. Learning textual prompts [Zhou *et al.*, 2022] effectively addresses this issue, as illustrated in Figure 1(a), but it still presents several drawbacks: 1) Uncertainty. The set trainable words are unknown, raising questions about what the semantic information they represent; 2) Coarseness. Pedestrian images with the same identity share a common prompt, and only four learnable tokens are allocated for identity depiction, which is insufficient for the cross-view and fine-grained nature of VIReID; 3) Cumbersomeness. Rather than end-to-end, the paradigm of learnable prompts requires a meticulously designed two-stage training process.

Recently, LLaVA [Liu *et al.*, 2023], a prominent large language-vision model, has demonstrated exceptional capability in image captioning. As shown in Figure 1(a), it can generate clear and detailed descriptions for pedestrian images, whose explicit semantics, such as age and gender, are able to facilitate the effective alignment of visual features. **This inspires us to supplement specific texts with the assistance of LLaVA and align image-text pairs within each modality.** Furthermore, the alignment of images to inter-modality texts is also necessary as it can further alleviate the distribution discrepancy between visual features. However, descriptions of visible and infrared images may include conflicting semantics, such as color attributes, which makes the direct alignment inappropriate. **This motivates us to explore an indirect alignment of images to inter-modality texts.** In addition, as shown in Figure 1(b), within each modality, the descriptions corresponding to different images of the same pedestrian contain complementary content. Integrating them to acquire comprehensive semantics and accordingly guide the alignment of visual features is beneficial for further enhancing their modality invariance. **This prompts us to enrich pedestrian semantics with multi-view texts.**

In this paper, we propose a novel framework termed Richer Semantics, Better Alignment (RSBA), which aligns visual features with explicit and enriched semantics for effective VIReID. As shown in Figure 2, it consists of Explicit Semantics-Guided Feature Alignment (ESFA), Consistent Similarity-Guided Indirect Alignment (CSIA), and Cross-View Semantics Compensation (CVSC). ESFA leverages LLaVA to generate textual descriptions for visible and infrared images, respectively, and maximizes the similarity between visible (infrared) image-text pairs to align cross-modality visual features into the semantic space. CSIA constrains the similarity between intra-modality image-text pairs to be consistent with that between inter-modality text-text pairs, achieving the indirect alignment of visible visual features with infrared semantics as well as infrared visual features with visible semantics. CVSC integrates text features from another view into the current view and accordingly improves the image-text matching in ESFA and CSIA from one-to-one to one-to-many, thereby further advancing their alignment. Our RSBA is trained end-to-end, with only the visual side used to extract cross-modality image features for testing.

Our main contributions are summarized as follows:

- We explore the advantages of explicit semantics in alleviating the modality gap between visible and infrared images, and accordingly propose ESFA to align visual features into the semantic space for effective VIReID.
- We realize the alignment of visual features with inter-modality semantics, and accordingly present CSIA to address the challenge of the direct alignment resulting from conflicting semantics.
- We consider the comprehensiveness of multi-view semantics, and develop CVSC to achieve the one-to-many alignment between images and texts, further strengthening the modality invariance of visual features.
- Extensive experiments across three datasets demonstrate that RSBA achieves new state-of-the-art performance, with each component contributing effectively.

2 Related Work

2.1 Visible-Infrared Person Re-Identification

VIReID is a challenging task due to the significant modality gap between visible and infrared images. An intuitive approach is to transfer images from one modality to the style of another. For instance, JSIA [Wang *et al.*, 2020] employed feature decoupling and cycle generation to augment cross-modality image pairs. Given the substantial gap between heterogeneous data, MSA [Miao *et al.*, 2021] designed a style similarity constraint to ensure the quality of generated images. To prevent identity information loss during transfer, ACD [Pan *et al.*, 2024] introduced conditional probability density to optimize the generation network. Although generative-based methods are intuitive and effective, they are prone to model collapse and susceptible to introducing noise.

Generative-free methods have recently attracted considerable attention due to they circumvent the limitations of generative-based approaches. These methods primarily focus on aligning cross-modality features by constructing appropriate networks or metric functions. For instance, Zero-Padding [Wu *et al.*, 2017] evaluated the suitability of four networks for VIReID and proposed a one-stream structure with a zero-padding strategy. AGW [Ye *et al.*, 2021b] devised a weighted regularization triplet loss to optimize the relative distance between positive and negative pairs in both intra-modality and inter-modality. To learn informative representations, DEEN [Zhang and Wang, 2023] designed an embedding expansion network to extract diverse features. However, the large modality discrepancy still makes the feature alignment challenging. In this paper, we explore a semantic-guided approach to effectively align visual features.

2.2 Large Language-Vision Models

Large language-vision models have emerged as a significant research topic, bridging computer vision and natural language processing. CLIP, a representative model, excels in learning visual content with high-level semantic information, showcasing exceptional potential across various downstream vision tasks [Wang *et al.*, 2022; Zhao *et al.*, 2022; Yan *et al.*, 2023b; Tang *et al.*, 2024; Yan *et al.*, 2024;

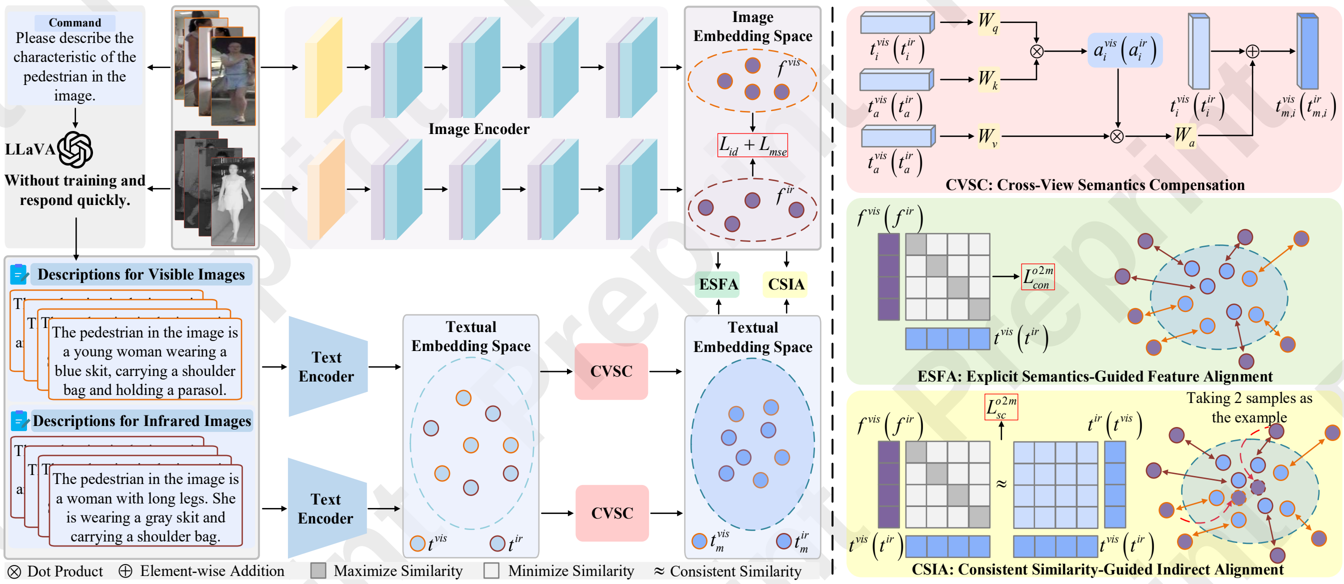

Figure 2: Overview of our RSBA. It acquires specific descriptions with LLaVA, integrates multi-view pedestrian semantics with CVSC, aligns visual features into the semantic space with ESFA, and indirectly aligns visual features with inter-modality semantics with CSIA.

Shen *et al.*, 2025]. In the field of ReID, CLIP-ReID [Li *et al.*, 2023] first introduced CLIP to advance this community. To tackle the occlusion problem, RGANet [He *et al.*, 2024] employed CLIP to generate local textual prototypes for mining discriminative part features. In VIReID, CSDN [Yu *et al.*, 2025] incorporated trainable textual prompts to acquire implicit pedestrian descriptions, aligning visual features of visible and infrared images into the semantic space. However, the semantics learned by CSDN are unknown and coarse, limiting its alignment ability. In this paper, we propose ESFA to address this limitation, and further develop CSIA and CVSC to improve our RSBA framework for more efficient VIReID.

3 Methodology

3.1 Preliminaries

Formally, we define the visible and infrared image sets as $\{x_i^{vis}\}_{i=1}^{N_v}$ and $\{x_i^{ir}\}_{i=1}^{N_r}$, where N_v and N_r represent the sizes of these two heterogeneous data, respectively. The label set is denoted as $\{y_i\}_{i=1}^{N_p}$, with N_p indicates the number of identities. In each mini-batch, N paired cross-modality images $\{x_i^{vis}, x_i^{ir}\}_{i=1}^{N}$ are randomly sampled and their visual features $\{f_i^{vis}, f_i^{ir}\}_{i=1}^{N} \in R^{N\times d}$ are extracted, where d is the dimension of features. We employ identity loss and modality-shared enhancement loss [Lu *et al.*, 2023] to optimize the network:

$$L_{id} = -\frac{1}{N}\sum_{i=1}^{N} q_i \log(p_i^{vis}) - \frac{1}{N}\sum_{i=1}^{N} q_i \log(p_i^{ir}), \quad (1)$$

where q_i is the one-hot vector of identity label y_i. p_i^{vis} and p_i^{ir} represent classification results of f_i^{vis} and f_i^{ir}, respectively.

The modality-shared enhancement loss constrains the average distance between positive pairs across modalities to be equal to that between positive pairs under the intra-modality:

$$L_{mse} = \frac{1}{2PK}\sum_{p=1}^{P}[\sum_{k=1}^{2K}(D_k^{intra} - D_k^{across})^2], \quad (2)$$

$$D^{intra} = \frac{1}{K-1}\sum_{\substack{k=1\\k\neq i}}^{K}\left\|f_i^{vis} - f_k^{vis}\right\|_2, \quad (3)$$

$$D^{across} = \frac{1}{K}\sum_{k=1}^{K}\left\|f_i^{vis} - f_k^{ir}\right\|_2, \quad (4)$$

where P and K denote P identities and K visible and K infrared images of each identity randomly sampled in each mini-batch. $\|\cdot\|_2$ represents the Euclidean distance.

3.2 Explicit Semantics-Guided Feature Alignment

Most existing frameworks treat VIReID as a pure vision task, lacking the ability to capture pedestrian semantics that is beneficial for modality alignment. Although CSDN introduces CLIP and CoOP to address this limitation, the uncertainty and coarseness of implicit semantics hinder the alignment of visual features into the semantic space. To this end, we propose ESFA, which leverages LLaVA to generate explicit textual descriptions and aligns cross-modality images with them.

As illustrated in Figure 2, given a pedestrian image, we send the request command 'Please describe the characteristics of the pedestrian in the image' to LLaVa. It responds with a natural language description 'The pedestrian in the image is a young woman wearing a blue skit, carrying a shoulder bag and holding a parasol'. This description provides clearer and more detailed explicit semantics, such as age, gender, and clothing, compared to the learnable textual prompt 'A photo of a $[X_1][X_2][X_3][X_4]$ person' in CSDN. Notably, LLaVA operates without requiring training and delivers responses quickly, taking approximately 1.2 seconds per image.

Suppose the generated language bases for visible and infrared images are $\{l_i^{vis}\}_{i=1}^{N_v}$ and $\{l_i^{ir}\}_{i=1}^{N_r}$. In each mini-batch, we sample $\{l_i^{vis}, l_i^{ir}\}_{i=1}^{N}$ corresponding to $\{x_i^{vis}, x_i^{ir}\}_{i=1}^{N}$ and input them into the textual encoder to extract features $\{t_i^{vis}, t_i^{ir}\}_{i=1}^{N} \in R^{N\times d}$. To align $\{f_i^{vis}, f_i^{ir}\}_{i=1}^{N}$ with $\{t_i^{vis}, t_i^{ir}\}_{i=1}^{N}$, we maximize the similarity between them:

$$L_{con} = L_{i2t} + L_{t2i}, \tag{5}$$

where

$$\begin{aligned} L_{i2t} = &-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(f_i^{vis}, t_i^{vis}))}{\sum_{j=1}^{N}\exp(s(f_i^{vis}, t_j^{vis}))} \\ &-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(f_i^{ir}, t_i^{ir}))}{\sum_{j=1}^{N}\exp(s(f_i^{ir}, t_j^{ir}))}, \end{aligned} \tag{6}$$

$$\begin{aligned} L_{t2i} = &-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(t_i^{vis}, f_i^{vis}))}{\sum_{j=1}^{N}\exp(s(t_i^{vis}, f_j^{vis}))} \\ &-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(t_i^{ir}, f_i^{ir}))}{\sum_{j=1}^{N}\exp(s(t_i^{ir}, f_j^{ir}))}, \end{aligned} \tag{7}$$

where $s(\cdot)$ represents the cosine similarity.

3.3 Consistent Similarity-Guided Indirect Alignment

ESFA achieves the alignment of images and texts in each modality; however, it ignores the alignment of images and inter-modality texts. A straightforward approach is to maximize the similarity between them similar to the above process. However, cross-modality texts describe the same object with conflicting attributes due to visual ambiguity. For example, the clothing exhibits 'blue' in the visible text while being seen as 'gray' in the infrared one. Forcing the maximization of similarity between images and inter-modality texts may disrupt the expressiveness of semantics. To this end, we develop CSIA to explore the indirect alignment between them.

As illustrated in Figure 2, for the visible visual feature f_i^{vis}, CSIA constrains its similarity with the visible text feature t_i^{vis} to be equal to the similarity between the infrared text feature t_i^{ir} and visible text feature t_i^{vis}, thereby indirectly establishing the alignment relationship between f_i^{vis} and t_i^{ir}. Similarly, infrared visual features f_i^{ir} and visible text features t_i^{vis} are indirectly aligned by constraining the similarity between f_i^{ir} and t_i^{ir} to be consistent with that between t_i^{vis} and t_i^{ir}:

$$\begin{aligned} L_{sc} = &\frac{1}{N}\sum_{i=1}^{N}(s(f_i^{vis}, t_i^{vis}) - s(t_i^{ir}, t_i^{vis}))^2 \\ &+\frac{1}{N}\sum_{i=1}^{N}(s(f_i^{ir}, t_i^{ir}) - s(t_i^{vis}, t_i^{ir}))^2. \end{aligned} \tag{8}$$

This similarity consistency loss not only achieves the alignment of images with inter-modality texts but also indirectly maximizes the similarity between infrared and visible texts, which helps alleviate cross-modality semantic discrepancy, thus facilitating more effective alignment of visual features.

3.4 Cross-View Semantics Compensation

The above two alignments are based on the one-to-one matching between image and text. However, within each modality, variations in camera views result in descriptions for different images of the same pedestrian emphasizing distinct objects. For example, the description for a front-facing image may highlight age and gender, while that for a rear-facing image is more likely to focus on hairstyle and backpack. As a result, semantics derived solely from single-view text are one-sided and contribute limited to the robustness of visual features. To address this limitation, we design CVSC to explore the one-to-many correspondence between images and texts.

As illustrated in Figure 2, we introduce an attention fusion module to integrate information in the textual feature from another view into the textual feature of the current view. Specifically, for the visible textual feature t_i^{vis}, we randomly select a textual feature t_a^{vis} that shares the same identity with t_i^{vis} while from different views. We compute the similarity between t_i^{vis} and t_a^{vis} to derive the attention weight a_i^{vis}:

$$a_i^{vis} = softmax\left(\frac{W_q(t_i^{vis})(W_k(t_a^{vis}))^T}{\sqrt{d}}\right), \tag{9}$$

where W_q and W_k are two linear projection layers. We multiply a_i^{vis} and t_a^{vis} to determine the contribution of t_a^{vis}, and add the the resulting weighted feature to t_i^{vis}:

$$t_{m,i}^{vis} = t_i^{vis} + W_a(a_i^{vis}W_v(t_a^{vis})), \tag{10}$$

where W_a and W_v are also linear projection layers. $t_{m,i}^{vis}$ represents the multi-view textual feature corresponding to l_i^{vis}, which contains richer pedestrian semantics as it compensates for the missing cross-view information in t_i^{vis}. Similarly, we can acquire the multi-view infrared textual feature $t_{m,i}^{ir}$.

We reformulate Equations (5), (6), and (7) as the following Equations (11), (12), and (13), which maximize the similarities between f_i^{vis} and $t_{m,i}^{vis}$, as well as between f_i^{ir} and $t_{m,i}^{ir}$:

$$L_{con}^{o2m} = L_{i2t}^{o2m} + L_{t2i}^{o2m}, \tag{11}$$

$$\begin{aligned} L_{i2t}^{o2m} = &-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(f_i^{vis}, t_{m,i}^{vis}))}{\sum_{j=1}^{N}\exp(s(f_i^{vis}, t_{m,j}^{vis}))} \\ &-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(f_i^{ir}, t_{m,i}^{ir}))}{\sum_{j=1}^{N}\exp(s(f_i^{ir}, t_{m,j}^{ir}))}, \end{aligned} \tag{12}$$

$$\begin{aligned} L_{t2i}^{o2m} = &-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(t_{m,i}^{vis}, f_i^{vis}))}{\sum_{j=1}^{N}\exp(s(t_{m,i}^{vis}, f_j^{vis}))} \\ &-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(t_{m,i}^{ir}, f_i^{ir}))}{\sum_{j=1}^{N}\exp(s(t_{m,i}^{ir}, f_j^{ir}))}. \end{aligned} \tag{13}$$

This process achieves the one-to-many alignment between images and texts within each modality. In addition, we also redefine the Equation (8) to the following Equation (14), indirectly aligning images with multi-view inter-modality texts:

$$\begin{aligned} L_{sc}^{o2m} = &\frac{1}{N}\sum_{i=1}^{N}(s(f_i^{vis}, t_{m,i}^{vis}) - s(t_i^{ir}, t_{m,i}^{vis}))^2 \\ &+\frac{1}{N}\sum_{i=1}^{N}(s(f_i^{ir}, t_{m,i}^{ir}) - s(t_i^{vis}, t_{m,i}^{ir}))^2. \end{aligned} \tag{14}$$

3.5 Training and Inference

The proposed RSBA is trained in an end-to-end manner, with the total loss expressed as:

$$L_{total} = L_{id} + L_{mse} + \lambda_1 L_{con}^{o2m} + \lambda_2 L_{sc}^{o2m}, \quad (15)$$

where λ_1 and λ_2 are two hyper-parameters used to balance the relative importance of L_{con}^{o2m} and L_{sc}^{o2m}, respectively.

Notably, the generation of language descriptions is only performed in the training phase, ensuring the practicality of our framework. During inference, the textual encoder and attention fusion module are not required, reducing the model complexity and inference time of our framework.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. **SYSU-MM01** [Wu *et al.*, 2017] contains 30,071 visible images captured by 4 RGB cameras and 15,792 infrared images captured by 2 IR cameras. The training set includes 22,258 visible images and 11,909 infrared images corresponding to 395 identities. The testing set comprises 3,803 infrared images of 96 identities and either 301 or 3,010 randomly sampled visible images for single-shot or multi-shot settings, respectively. **RegDB** [Nguyen *et al.*, 2017] is a small-scale VIReID dataset with 4,120 visible images and 4,120 infrared images from 412 pedestrians. Following the standard protocol, 2,060 visible and 2,060 infrared images of 206 identities are allocated for training, while the remaining images are used for testing. **LLCM** [Zhang and Wang, 2023] is a recently released challenging VIReID dataset collected under low-light conditions. Its training set includes 16,946 visible images and 13,975 infrared images of 713 identities, and its testing set consists of 8,680 visible images and 7,166 infrared images corresponding to 351 identities.

Evaluation Metrics. We assess the retrieval performance using the general indicators named mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC).

4.2 Implementation Details

We conduct experiments using the PyTorch library on a single RTX 4090 GPU. The proposed RSBA framework incorporates a training-free LLaVA, a CLIP model comprising a visual encoder and a textual encoder, with ResNet50 [He *et al.*, 2016] serving as the backbone for the visual encoder, and an attention fusion module consisting of four linear projection layers. Following AGW [Ye *et al.*, 2021b], we train two parallel first convolutional layers of ResNet50 for each modality while sharing the parameters of the subsequent four blocks. During training, we randomly sample 8 identities, each with 4 visible and 4 infrared images. All input images are resized to 288×144 and subjected to data augmentation techniques, including random padding, cropping, and flipping. The training process spans 120 epochs, with the initial learning rate set to 3e-4 for the visual encoder and 1e-6 for the textual encoder and attention fusion module, decaying by a factor of 0.1 at the 40th and 70th epochs, respectively. The hyper-parameters are configured as $\lambda_1 = 0.25$ and $\lambda_2 = 0.2$.

4.3 Comparison with State-of-the-Art Methods

SYSU-MM01. Table 1 presents the comparison results with the state-of-the-art methods on the SYSU-MM01 dataset, showing that RSBA consistently outperforms them across all settings. Specifically, in the all-search testing mode, our Rank-1 accuracy and mAP surpass those of the best generative-based method, ACD, by 4.0% (4.2%) and 3.7% (3.1%), respectively, while in the indoor-search mode, the improvements are 8.8% (6.3%) and 5.1% (5.2%). These gains are attributed to our approach aligning modalities at the feature level, which circumvents performance limitations imposed by the generated low-quality images. Compared to generative-free methods, under the single-shot mode, our Rank-1 accuracy exceeds that of CycleTrans by 1.9% (0.5%), and our mAP surpasses HOS-Net by 0.6% (2.1%). This advantage arises from the proposed RSBA aligns visual features with the semantic space, which are beneficial for alleviating the modality gap. Furthermore, our RSBA also outperforms CSDN across all settings, benefiting from its ability to capture clear, detailed, and rich semantics, as opposed to the coarse and ambiguous semantics learned by CSDN.

RegDB. We further evaluate the performance of RSBA on the RegDB dataset, with the quantitative results summarized in Table 2. Our method achieves superior recognition rates compared to existing generative-based approaches. For instance, in the visible-to-infrared testing mode, RSBA outperforms TSME in Rank-1 accuracy by 7.9% and surpasses ACD in mAP by 7.6%. Similarly, our method exhibits significant performance advantages over state-of-the-art generative-free methods, such as MBCE and HOS-Net. In comparison with CSDN, RSBA improves the Rank-1 and mAP by 2.1% and 4.0% in the visible-to-infrared testing mode.

LLCM. We also evaluate the proposed RSBA on the challenging LLCM dataset to provide a comprehensive assessment. As detailed in Table 3, in the visible-to-infrared testing mode, RSBA achieves a Rank-1 accuracy and mAP that are 1.1% and 0.9% higher, respectively, than those of the state-of-the-art HOS-Net. Similarly, in the infrared-to-visible testing mode, RSBA outperforms HOS-Net with improvements of 0.9% in Rank-1 accuracy and 1.0% in mAP. These results further highlight the superiority of our approach.

4.4 Ablation Studies

We evaluate the effectiveness of each component in our proposed RSBA, with the results presented in Table 4. The Rank-1 and mAP of Baseline ('0') are 71.9% and 67.6% under the single-shot and 80.0% and 61.9% under the multi-shot.

Effectiveness of ESFA. ESFA aims to introduce explicit semantics to guide the alignment of cross-modality visual features. As shown in Table 4, it improves the Rank-1 and mAP by 4.4% and 5.0% under the single-shot mode, which validates that aligning visual features into the semantic space is reasonable and effective for mitigating the modality gap.

Effectiveness of CSIA. CSIA constrains the consistent similarity between intra-modality image-text pairs and inter-modality text-text pairs to establish the correspondence between images and cross-modality texts. As detailed in Table 4, under the single-shot test mode, it improves the Rank-1 accuracy from 76.3% to 77.4%, which indicates that the align-

Methods	Ref	All-Search				Indoor-Search			
		Single-Shot		Multi-Shot		Single-Shot		Multi-Shot	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP
cmGAN [Dai *et al.*, 2018]	IJCAI'18	26.9	27.8	31.4	22.2	31.6	42.1	37.0	32.7
Hi-CMD [Choi *et al.*, 2020]	CVPR'20	34.9	35.9	-	-	-	-	-	-
JSIA [Wang *et al.*, 2020]	AAAI'20	38.1	36.9	45.1	29.5	43.8	52.9	52.7	42.7
MSA [Miao *et al.*, 2021]	IJCAI'21	63.1	59.2	-	-	67.1	72.7	-	-
TSME [Liu *et al.*, 2022b]	TCSVT'22	64.2	61.2	70.3	54.3	64.8	71.5	76.8	65.0
ACD [Pan *et al.*, 2024]	TIFS'24	74.4	71.1	80.4	66.9	78.9	82.7	86.0	78.6
AGW [Ye *et al.*, 2021b]	TPAMI'21	47.5	47.6	-	-	54.1	62.9	-	-
MCSL [Ling *et al.*, 2021]	IJCAI'21	64.8	60.8	68.0	51.4	-	-	-	-
CAJ [Ye *et al.*, 2021a]	ICCV'21	69.8	66.8	-	-	76.2	76.7	-	-
MMN [Zhang *et al.*, 2021]	MM'21	70.6	66.9	-	-	76.2	79.6	-	-
MAUM [Liu *et al.*, 2022a]	CVPR'22	71.6	68.7	-	-	76.9	81.9	-	-
CIFT [Li *et al.*, 2022]	ECCV'22	71.7	67.6	78.0	62.4	78.6	82.1	86.9	77.0
MBCE [Cheng *et al.*, 2023]	AAAI'23	74.7	72.0	78.3	65.7	83.4	86.0	88.4	80.6
DEEN [Zhang and Wang, 2023]	CVPR'23	74.7	71.8	-	-	80.3	83.3	-	-
SEFL [Feng *et al.*, 2023]	CVPR'23	75.1	70.1	-	-	78.4	81.2	-	-
HOS-Net [Qiu *et al.*, 2024]	AAAI'24	75.6	74.2	-	-	84.2	86.7	-	-
CSCL [Liu *et al.*, 2025]	TMM'24	75.7	72.0	-	-	80.8	83.5	-	-
CycleTans [Wu *et al.*, 2025]	TNNLS'24	76.5	72.6	82.8	68.5	87.2	84.9	91.2	81.4
CSDN [Yu *et al.*, 2025]	TMM'25	76.7	73.0	83.5	67.9	84.5	86.8	91.3	82.2
Ours (RSBA)	IJCAI'25	**78.4**	**74.8**	**84.6**	**70.0**	**87.7**	**87.8**	**92.3**	**83.8**

Table 1: Performance comparison with state-of-the-art methods on SYSU-MM01. '-' denotes that no reported result is available.

Methods	Visible to Infrared		Infrared to Visible	
	R1	mAP	R1	mAP
Hi-CMD	70.9	66.0	-	-
JSIA	48.1	48.9	48.5	49.3
MSA	84.8	82.1	-	-
TSME	87.3	76.9	86.4	75.7
ACD	84.7	83.2	87.1	84.7
AGW	70.0	66.4	-	-
MCSL	93.8	87.5	91.5	85.2
CAJ	85.0	65.3	84.7	61.5
MMN	91.6	84.1	87.5	80.5
MAUM	87.8	85.0	86.9	84.3
CIFT	92.1	86.9	90.1	84.8
MBCE	93.1	88.3	93.4	87.9
DEEN	91.1	85.1	89.5	83.4
SEFL	91.0	85.2	92.1	86.5
HOS-Net	94.7	90.4	93.3	89.2
CSCL	92.1	84.2	89.6	85.0
CycleTrans	90.6	85.6	81.8	87.0
CSDN	95.4	87.7	92.3	85.5
Ours (RSBA)	**95.2**	**90.8**	**94.4**	**89.5**

Table 2: Performance comparison on RegDB.

Methods	Visible to Infrared		Infrared to Visible	
	R1	mAP	R1	mAP
AGW	51.5	55.3	43.6	51.8
CAJ	56.5	59.8	48.8	56.6
MMN	59.9	62.7	52.5	58.9
DEEN	62.5	65.8	54.9	62.9
HOS-Net	64.9	67.9	56.4	63.2
Ours (RSBA)	**66.0**	**68.8**	**57.3**	**64.2**

Table 3: Performance comparison on LLCM.

	ESFA	CSIA	CVSC	Single-Shot		Multi-Shot	
				R1	mAP	R1	mAP
0				71.9	67.6	80.0	61.9
1	✓			76.3	72.6	82.1	66.5
2	✓	✓		77.4	73.2	82.7	68.2
3	✓		✓	77.6	73.7	83.1	68.8
4	✓	✓	✓	**78.4**	**74.8**	**84.6**	**70.0**

Table 4: Ablation studies of our RSBA.

ment of inter-modality image-text pairs plays a positive role in the further effective alignment of visual features.

Effectiveness of CVSC. CVSC integrates multi-view texts to capture comprehensive semantics that are beneficial for improving the alignment in ESFA and CSIA. As illustrated in Table 4, when it is equipped with ESFA, the Rank-1 accuracy is improved by 1.3% and 1.0% under the two test modes, respectively. In addition, when incorporating it with both ESFA and CSIA, the recognition performance reaches a peak. These results fully demonstrate the reasonableness of motivation behind CVSC and the effectiveness of its technology.

4.5 Parameters Analysis

We introduce the hyper-parameters λ_1 and λ_2 to regulate the relative importance of the loss terms L_{con}^{o2m} and L_{sc}^{o2m}. The former optimizes the model to align image-text pairs within each modality, while the latter drives the model to mine the correspondence between image-text pairs across modalities.

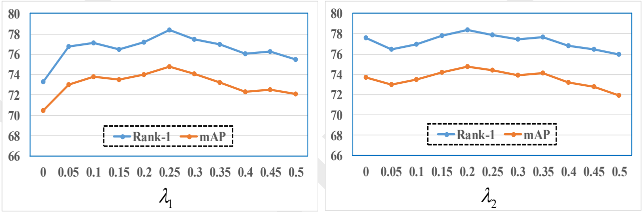

Figure 3: Parameters analysis of λ_1 and λ_2.

As shown in Figure 3, the optimal values for λ_1 and λ_2 are 0.25 and 0.2. In addition, excessively large values diminish the contributions of the identity loss and modality-shared enhancement loss, while overly small values result in insufficient alignment, both leading to performance degradation.

4.6 Further Discussions

In this section, we further discuss each proposed module, with the experimental results presented in Table 5.

Modules		Single-Shot			Multi-Shot		
		R1	R10	mAP	R1	R10	mAP
ESFA	1	74.1	95.0	69.9	80.9	96.5	64.3
	2	**76.3**	**96.8**	**72.6**	**82.1**	**98.1**	**66.5**
CSIA	1	75.6	95.9	72.1	80.9	97.5	65.6
	2	**77.4**	**97.9**	**73.2**	**82.7**	**98.3**	**68.2**
CVSC	1	**78.4**	**98.6**	**74.8**	**84.6**	**99.0**	**70.0**
	2	77.8	98.1	74.3	83.5	98.2	68.9
	3	76.7	96.9	72.9	81.8	97.6	67.1

Table 5: Further discussions of each proposed module.

The superiority of explicit semantics

Different from the implicit semantics in CSDN [Yu *et al.*, 2025], our ESFA acquires explicit pedestrian semantics to align cross-modality visual features. In contrast, the latter is more representative and thus guides the more efficient alignment. As shown in Table 5, the recognition performance achieved by ESFA (2) is higher than that achieved based on implicit semantic alignment (1), with improvements of 2.1% in Rank-1 accuracy and 2.7% in mAP under the single-shot test mode. Notably, we observe that the descriptions generated by LLaVA follow the fixed sentence structure of 'The pedestrian in the image is a [age group] [gender] wearing [clothing], carrying [accessory]', which may cause the model to overfit to the non-differentiated semantic pattern, limiting the effect of alignment. This motivates us to explore acquiring diverse pedestrian semantics in the future.

The advantage of indirect alignment

(1) Why align visual features with inter-modality semantics? The proposed ESFA achieves alignment of visible visual features and visible semantics, as well as infrared visual features and infrared semantics. If we further align visible visual features and infrared semantics, as well as infrared visual features and visible semantics, the distribution discrepancy between visual features of visible and infrared can be further reduced. (2) Why indirectly align them? Different from image-text pairs within each modality, which naturally correspond to each other, images and inter-modality texts are not completely matched. Therefore, aligning them directly by maximizing the similarity between them may destroy the expressiveness of semantics, thereby weakening the alignment between intra-modality image-text pairs. As shown in Table 5, the direct alignment (1) reduces the Rank-1 accuracy and mAP of ESFA from 76.3% to 75.6% and from 72.6% to 72.1%. In contrast, our designed indirect alignment strategy (2) improves the Rank-1 and mAP by 1.1% and 0.6%. This proves the rationality and effectiveness of the approach.

The number of cross-view texts

The proposed CVSC aims to enrich pedestrian semantics with multi-view texts, and we achieve this by integrating text with that from an additional view. It is also feasible to integrate it with texts from multiple additional views. However, we observe that the recognition performance degrades as the number of views increases (2 and 3). This is because the generated descriptions may contain some inaccurate content, amplifying the noisy semantics during the information integration. In addition, CVSC is achieved through the attention fusion network, which requires more parameters as the number of views increases, making model optimization challenging.

4.7 Limitations

This paper acquires explicit and enriched semantics to effectively alleviate the modality gap between visible and infrared pedestrian images. However, as we discussed above, on the one hand, the rigid semantic pattern weakens the effect of alignment. On the other hand, this paper initially explores the enrichment of pedestrian semantics with multi-view texts, while we ignore the quality of texts, the number of cross-view texts, and the strategy of text fusion, which all affect the richness of the semantics. These limitations motivate us to explore the semantics of diversity and richness more deeply.

5 Conclusion

In this paper, we propose a novel Richer Semantics, Better Alignment (RSBA) framework for effective VIReID. It focuses on aligning visible and infrared visual features with explicit and enriched semantics and achieves this through Explicit Semantics-Guided Feature Alignment (ESFA), Consistent Similarity-Guided Indirect Alignment (CSIA), and Cross-View Semantics Compensation (CVSC). ESFA supplements language descriptions for pedestrian images and builds the correspondence of image-text pairs, aligning visual features into the semantic space. CSIA introduces the similarity consistency constraint to indirectly align visual features with inter-modality semantics, further alleviating the distribution discrepancy of visual features. CVSC mines comprehensiveness semantics to further facilitate ESFA and CSIA. Experimental results highlight the advancements RSBA achieves over state-of-the-art methods. In the future, we will further explore the assistance of semantic information for VIReID.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62172212 and Grant 62427819, the Natural Science Foundation of Jiangsu Province under Grant BK20230031, the Jiangsu Provincial Science and Technology Major Project under Grant BG2024042.

References

[Cheng *et al.*, 2023] De Cheng, Xiaolong Wang, Nannan Wang, Zhen Wang, Xiaoyu Wang, and Xinbo Gao. Cross-modality person re-identification with memory-based contrastive embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 425–432, 2023.

[Choi *et al.*, 2020] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10257–10266, 2020.

[Dai *et al.*, 2018] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, volume 1, page 6, 2018.

[Dong *et al.*, 2024a] Neng Dong, Shuanglin Yan, Hao Tang, Jinhui Tang, and Liyan Zhang. Multi-view information integration and propagation for occluded person re-identification. *Information Fusion*, 104:102201, 2024.

[Dong *et al.*, 2024b] Neng Dong, Liyan Zhang, Shuanglin Yan, Hao Tang, and Jinhui Tang. Erasing, transforming, and noising defense network for occluded person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):4458–4472, 2024.

[Feng *et al.*, 2023] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22752–22761, 2023.

[Gong *et al.*, 2022] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4313–4322, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[He *et al.*, 2024] Shuting He, Weihua Chen, Kai Wang, Hao Luo, Fan Wang, Wei Jiang, and Henghui Ding. Region generation and assessment network for occluded person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:120–132, 2024.

[Li *et al.*, 2021] Huafeng Li, Neng Dong, Zhengtao Yu, Dapeng Tao, and Guanqiu Qi. Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2814–2830, 2021.

[Li *et al.*, 2022] 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 *European Conference on Computer Vision*, pages 381–398. Springer, 2022.

[Li *et al.*, 2023] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1405–1413, 2023.

[Ling *et al.*, 2021] Yongguo Ling, Zhiming Luo, Yaojin Lin, and Shaozi Li. A multi-constraint similarity learning with adaptive weighting for visible-thermal person re-identification. In *IJCAI*, pages 845–851, 2021.

[Liu *et al.*, 2022a] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19366–19375, 2022.

[Liu *et al.*, 2022b] Jianan Liu, Jialiang Wang, Nianchang Huang, Qiang Zhang, and Jungong Han. Revisiting modality-specific feature compensation for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7226–7240, 2022.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023.

[Liu *et al.*, 2025] Min Liu, Zhu Zhang, Yuan Bian, Xueping Wang, Yeqing Sun, Baida Zhang, and Yaonan Wang. Cross-modality semantic consistency learning for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 27:568–580, 2025.

[Lu *et al.*, 2023] Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1835–1843, 2023.

[Miao *et al.*, 2021] Ziling Miao, Hong Liu, Wei Shi, Wanlu Xu, and Hanrong Ye. Modality-aware style adaptation for rgb-infrared person re-identification. In *IJCAI*, pages 916–922, 2021.

[Nguyen *et al.*, 2017] 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.

[Pan *et al.*, 2024] Honghu Pan, Wenjie Pei, Xin Li, and Zhenyu He. Unified conditional image generation for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:9026–9038, 2024.

[Qiu *et al.*, 2024] Liuxiang Qiu, Si Chen, Yan Yan, Jing-Hao Xue, Da-Han Wang, and Shunzhi Zhu. High-order structure based middle-feature learning for visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4596–4604, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[Shen *et al.*, 2025] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6795–6804, 2025.

[Tang *et al.*, 2024] Hao Tang, Zechao Li, Dong Zhang, Shengfeng He, and Jinhui Tang. Divide-and-conquer: Confluent triple-flow network for rgb-t salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Wang *et al.*, 2020] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12144–12151, 2020.

[Wang *et al.*, 2022] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695, 2022.

[Wu *et al.*, 2017] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5380–5389, 2017.

[Wu *et al.*, 2025] Qiong Wu, Jiaer Xia, Pingyang Dai, Yiyi Zhou, Yongjian Wu, and Rongrong Ji. Cycletrans: Learning neutral yet discriminative features via cycle construction for visible- infrared person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):5469–5479, 2025.

[Yan *et al.*, 2022] Yuming Yan, Huimin Yu, Shuzhao Li, Zhaohui Lu, Jianfeng He, Haozhuo Zhang, and Runfa Wang. Weakening the influence of clothing: Universal clothing attribute disentanglement for person re-identification. In *IJCAI*, pages 1523–1529, 2022.

[Yan *et al.*, 2023a] Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. Learning comprehensive representations with richer self for text-to-image person re-identification. In *ACM International Conference on Multimedia (ACM MM)*, page 6202–6211, 2023.

[Yan *et al.*, 2023b] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32:6032–6046, 2023.

[Yan *et al.*, 2024] Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. Image-specific information suppression and implicit local alignment for text-based person search. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):17973–17986, 2024.

[Ye *et al.*, 2021a] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021.

[Ye *et al.*, 2021b] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2021.

[Yu *et al.*, 2025] Xiaoyan Yu, Neng Dong, Liehuang Zhu, Hao Peng, and Dapeng Tao. Clip-driven semantic discovery network for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, pages 1–13, 2025.

[Zhang and Wang, 2023] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2153–2162, 2023.

[Zhang *et al.*, 2021] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 788–796, 2021.

[Zhao *et al.*, 2022] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European Conference on Computer Vision*, pages 159–175. Springer, 2022.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.