Advancing Visible-Infrared Person Re-Identification: Synergizing Visual-Textual Reasoning and Cross-Modal Feature Alignment

Yuxuan Qiu, Liyang Wang, Wei Song, Jiawei Liu, Zhiping Shi, and Na Jiang, *Member, IEEE*

***Abstract*—Visible-infrared person re-identification (VI-ReID) is a critical cross-modality fine-grained classification task with significant implications for public safety and security applications. Existing VI-ReID methods primarily focus on extracting modality-invariant features for person retrieval. However, due to the inherent lack of texture information in infrared images, these modality-invariant features tend to emphasize global contexts. Consequently, individuals with similar silhouettes are often misidentified, posing potential risks to security systems and forensic investigations. To address this problem, this paper innovatively introduces natural language descriptions to learn the global-local contexts for VI-ReID. Specifically, we design a framework that jointly optimizes visible-infrared alignment plus (VIAP) and visual-textual reasoning (VTR), and introduces local-global joint measure (LJM) to enhance the metric, while proposing a human-LLM collaborative approach to incorporate textual descriptions into existing cross-modal person re-identification datasets. VIAP achieves cross-modal alignment between RGB and IR. It can explicitly utilize designed frequency-aware modality alignment and relationship-reinforced fusion to explore the potential of local cues in global features and modality-invariant information. VTR proposes pooling selection and dual-level reasoning mechanisms to force the image encoder to pay attention to significant regions based on textual descriptions. LJM proposes introducing local feature distances into the measure stage metric to enhance the relevance of matching using fine-grained information. Extensive experimental results on the popular SYSU-MM01 and RegDB datasets show that the proposed method significantly outperforms state-of-the-art approaches. The dataset is publicly available at https://github.com/qyx596/vireid-caption.**

***Index Terms*—Person re-identification, modality alignment, relation reasoning.**

Received 23 July 2024; revised 5 January 2025 and 26 January 2025; accepted 27 January 2025. Date of publication 11 February 2025; date of current version 24 February 2025. This work was supported in part by Beijing Natural Science Foundation under Grant QY24309; in part by the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, under Grant HMHAI-202407; in part by the National Natural Science Foundation of China under Grant 62376166 and Grant 62476260; and in part by the 2023 Innovation Fund of Engineering Research Center of Integration and Application of Digital Learning Technology, Ministry of Education, under Grant 1311021. The associate editor coordinating the review of this article and approving it for publication was Dr. Naser Damer. *(Corresponding author: Na Jiang.)*

Yuxuan Qiu, Liyang Wang, Wei Song, Zhiping Shi, and Na Jiang are with the Information Engineering College, Capital Normal University, Beijing 100048, China (e-mail: 1201004027@cnu.edu.cn; wly@cnu.edu.cn; wsong@cnu.edu.cn; shizp@cnu.edu.cn; jiangna@cnu.edu.cn).

Jiawei Liu is with the Department of Automation, University of Science and Technology of China, Hefei 230026, China (e-mail: jwliu6@ustc.edu.cn).

Digital Object Identifier 10.1109/TIFS.2025.3539946

I. Introduction

PERSON Re-Identification (ReID), a critical technology in video surveillance and forensics, enables precise identification of individuals across multiple camera views to enhance security monitoring capabilities. Early methods [2], [3] utilizing RGB images as input have achieved excellent performance on closed datasets where the gallery and query sets share similar environments. Nevertheless, RGB images captured in low-light or nighttime conditions cannot adequately display pedestrians' appearance or silhouettes, posing challenges for practical real-world applications.

To address identification under varying lighting environments, infrared (IR) images are incorporated, expanding ReID to the cross-modality task of visible-infrared ReID (VI-ReID) [4], [5]. The primary challenge for VI-ReID lies in handling the inherent modality differences between infrared and visible images, which exhibit distinct visual characteristics and feature representations. At the same time, global contexts are frequently captured as discriminative representations, while neglecting local information, leading to misidentification of similar silhouettes. Text-based person retrieval enables matching person images via textual descriptions, offers better alignment with human intuition than image-based person ReID. It also effectively handles semantically ambiguous cases. However, the method supports only text-to-RGB retrieval and performing poorly in low-light conditions.

To make up for these deficiencies, in this paper we propose to introduce natural language specifications for the VI-ReID task. Its motivation comes from our daily lives. As demonstrated in Figure 1, when searching for uniquely certain individual in actual scenarios, we not only provide reference photos, but also use language to describe the gender and appearance. These natural language descriptions will help us observe different local areas of pedestrians to identify whether they are retrieval persons.

The introduction of natural language transforms ReID and VI-ReID into multi-modality tasks [6], [7]. However, while additional modalities bring diversity in retrieval information, they also pose challenges in modality alignment. To address this, we propose STAR-ReID with joint visible-infrared alignment plus (VIAP), visual-textual reasoning (VTR) and local-global joint measure (LJM) for VI-ReID. VIAP is responsible for cross-modality learning between RGB and IR

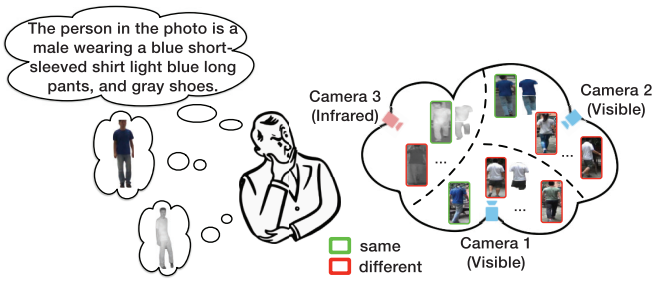

Fig. 1. Visual explanation of motivation in this paper. Taking the person (ID:0002 in SYSU-MM01 dataset [1]) in the picture as an example, humans will focus on comparing the similarity of the whole body in VI-ReID, which makes it difficult to find the same identity. But after considering natural language descriptions, humans will perform local matching according to the keywords, which makes VI-ReID easier. This prior indicates that natural language description is beneficial for establishing global-local relationships between visible and infrared modalities.

without spatial alignment. It utilizes frequency-aware modality alignment to simultaneously learn inter-modality global features and intra-modality global-local features, and then achieves a distinguishable summary of these features through a relationship-reinforced fusion (R^2 fusion). VTR leverages textual semantics to attend to local salient regions in images. It designs a pooling selection mechanism to obtain auxiliary features that can compensate for the lack of local information in vision tokens. It then employs dual-level reasoning with hard sample mining to implicitly guide VIAP in enhancing the learning of fine-grained features by providing global-local guidance. In test stage, LJM further optimizes the initial retrieval results. Unlike the existing similarity measure strategy [8], LJM leverages local similarity constraints to enhance the relevance of matches by incorporating fine-grained information.

Extensive experimental results demonstrate that the proposed method significantly outperforms most state-of-the-art approaches on the popular SYSU-MM01 and RegDB datasets.

In summary, the main contributions are as follows:

1) Propose a novel STAR-ReID architecture that introduces natural language specification into VI-ReID, with VTR to strengthen fine-grained visual representations via dual-level reasoning.
2) Design VIAP with frequency-aware modality alignment and relationship-reinforced fusion, which explicitly improves the modality-invariant features by low-frequency local information reinforced global clues.
3) Design LJM with a local-global joint similarity metric, leveraging local features for group retrieval to reduce the interference caused by spatial and semantic inconsistencies, optimizing the ranking list.

Compared to the preliminary version [9], this paper introduces the following improvements. The VIA module has been redesigned to VIAP by incorporating low-frequency information from local features to enhance feature discrimination. At the same time, the local-global joint measure is proposed and the text encoder has been upgraded, which further strengthening fine-grained semantic extraction and utilization, improving retrieval performance in complex scenarios. In addition, the construction method of proposed dataset was added, and the link for access was made publicly available. These enhancements result in up to a 5% improvement in both R1 and mAP metrics. The experiments including different K settings and weighting factors in LJM were conducted to validate the value of proposed components. We believe that by introducing natural language specification into VI-ReID, fine-grained feature extraction is enhanced for modality alignment, expanding retrieval paradigms and enabling integration with large language models or AI agents.

II. Related Work

A. Visible-Infrared Person Re-Identification

Person ReID was a single-modal task initially [10], focusing primarily on effectively describing persons in RGB images. Notable works such as AMR [11] achieves this by leveraging appearance attributes to generate more comprehensive person descriptive features, while APN [12] dynamically segments pedestrian images to effectively extract local features under varying capture conditions. However, the RGB-based ReID method faces challenges in feature reliability under varying lighting conditions, which motivates the proposal of VI-ReID that integrates both RGB and IR.

VI-ReID further aims to overcome modality differences and learn modality-invariant features that are robust across the visible and infrared spectra [13], [14], [15]. Various methods have been proposed to bridge the gap between RGB and IR modalities. For instance, XIV [16] introduces an auxiliary X-modality to facilitate cross-modal learning. WRIM-Net [17] is guided to extract modality-invariant information by enhancing local region interactions between modalities. DEEN [5] employs cross-modal embedding modules to extract discriminative features that are shared across modalities. RLE [18] simulates local linear transformations of material surfaces under different modalities for data augmentation narrowing the inter-modal differences. PMWGCN [19] utilizes wavelet transform to suppress non-stationary high-frequency noise in features, improving the performance in image degradation scenarios. SGIEL [20] leverages human pose estimation as a guiding cue for extracting modality-invariant features, exploiting the structural consistency of human silhouettes across RGB and IR images.

In contrast to these approaches that focus on finding a common representation space, some researchers argue that feature differentiation should be improved by enhancing cross-modal constraints. To this end, MAUM [21] proposes a memory enhancement-based metric to suppress modal discrepancies by maintaining a dynamic feature memory bank. DART [22] designs a confidence constraint to alleviate modal noise, adaptively adjusting the contribution of each modality based on its reliability. Partmix [4] further incorporates modal entropy constraints and data augmentation techniques to improve the discriminability and generalization of learned features. IDKL [23] enhances the discriminability through the distillation of implicit information from modality-specific features.

Generative Adversarial Networks (GANs) have also shown significant performance improvements in VI-ReID tasks [24],

[25], [26], [27]. These methods often utilize transfer learning between visible and infrared images to generate cross-modal data. For example, FMCNet [28] uses GAN-generated virtual modalities to compensate for original modal information and achieve feature fusion. AlignGAN [29] aligns modalities at the pixel level through adversarial learning, while cmGAN [30] learns modality-transformed patterns for cross-modality image training, effectively reducing the domain gap. Furthermore, GML [31] designed a generative metric learning approach that improves performance by generating adversarial samples and normalizing them to produce robust distance metrics.

Although the aforementioned methods have achieved remarkable success in terms of performance, they are still constrained by the inherent limitation of infrared images: the concealment of local fine-grained cues. This limitation leads to the learning of modal-invariant features that predominantly emphasize global content, such as overall silhouettes or coarse-grained structures, while neglecting local regions that are crucial for fine-grained classification tasks like person re-identification. For instance, discriminative attributes such as accessories, clothing patterns, or gait peculiarities are often obscured or absent in infrared images. While GANs struggle with limited priors, often producing unrealistic and unreliable generated images that severely competent algorithmic generalization to unseen scenarios.

To address these issues, this paper introduces natural language specifications generated by human-GPT-4 collaborating [32] to guide both global and local attention patterns. The rationale behind this approach is that natural language descriptions can provide rich, human-interpretable information about salient local attributes (e.g., "wearing a red hat" or "carrying a backpack") that might be indiscernible in infrared images.

B. Text-Based Person Retrieval

At present, there is no existing research focused on using natural language to enhance Visible-Infrared Re-Identification (VI-ReID). However, with the rise of large models and multi-modal research, text-based person retrieval (TPR) has become a research hotspot. TPR, initially proposed in 2017 by Li et al. [33], is a retrieval task that searches for pedestrians based on textual descriptions, can be regarded as a type of cross text and visual content mining [34]. One significant challenge in TPR is effectively aligning features from two inconsistent encoding spaces. Early efforts primarily utilized well-established models from their respective domains, such as VGG [35] and ResNet50 [36] in the visual realm, and LSTM [37] and BERT [38] in the NLP domain, to learn robust representations encompassing both visual and textual modalities. In the subsequent modal alignment, these efforts typically started from a global context perspective [39], evolving into self-adaptive local semantic learning across different granularities [40], [41]. Effective semantic information includes color [42], gender, appearance [43], and other attributes.

To enhance semantic information, vision-language pretraining models have been introduced into TPR [44], [45], [46]. For instance, the renowned Contrastive Language-Image Pre-training (CLIP) model [47], trained on an extensive repository of text-image pairs, was leveraged to underscore potential modal alignment. Yan et al. [46] devised a framework that harnesses the capabilities of the CLIP model to extract nuanced information, advancing the task of text-based person retrieval. This framework optimally utilizes the vast knowledge embedded within numerous image-text pairs, thereby enhancing cross-modal transfer learning. Jiang and Ye [7] introduced the IRRA model, an evolution of the original CLIP model, enabling the acquisition of more discerning text-image embeddings. Subsequently, Bai et al. [48] proposed a relation and sensitivity-aware representation learning method, which uses Masked Language Modeling (MLM) to predict masked tokens, obtaining more discriminative and informative features.

Various cross-modal alignment or reasoning strategies mentioned in the above methods have witnessed great success in TPR. However, these methods only learn modality-invariant features between visible images and natural languages. Both modalities include fine-grained information, facilitating global-local alignment through semantic cues. In this work, we introduce natural language into VI-ReID, meaning that language, visible, and infrared modalities coexist. The fine-grained information in infrared images is latent, making it challenging to directly align natural language with infrared images. Building on prior experience, we jointly train the three modalities through two cross-modal learning processes (visible-infrared and visual-language). This approach aims to effectively bridge the gap between the modalities and enhance the performance of VI-ReID by leveraging the strengths of each modality.

III. Method

In this section, we introduce the proposed STAR-ReID architecture, elaborate on its components, and describe the method of using human-LLM collaboration to augment existing VI-ReID datasets with natural language descriptions. As shown in Figure 2, a visible image I^i_{RGB}, an infrared image I^i_{IR}, and a natural language description T^i with the same person ID form a triple input. Given a training set $\mathcal{X}$ containing M triple samples (I^i_{RGB}, I^i_{IR}, T^i) and their corresponding IDs GT^i, the objective of STAR-ReID is to leverage the natural language description T^i to guide discriminative feature extraction for VI-ReID. I^i_{RGB} and I^i_{IR} extract features through shared Transformer layers from ViT [49]. The text branch extracts features using a pretrained BERT [38] as the backbone network. However, infrared features lack obvious fine-grained cues, making it difficult to directly align with natural language or perform relational reasoning. Therefore, STAR-ReID designs VIAP to directly align the visible RGB and infrared IR modalities, then introduces text through VTR to achieve multi-modality learning indirectly. Finally, LJM optimizes the ranking list by measuring the similarity of fine-grained features optionally.

Unlike existing methods that only learn modality invariant features, VIAP adopts novel frequency-aware modality alignment to synchronously learn inter-modality global features and intra-modality global-local features. Meanwhile, R^2 fusion adaptively achieves distinguishable summaries of these features. In addition, VTR utilizes CLS tokens

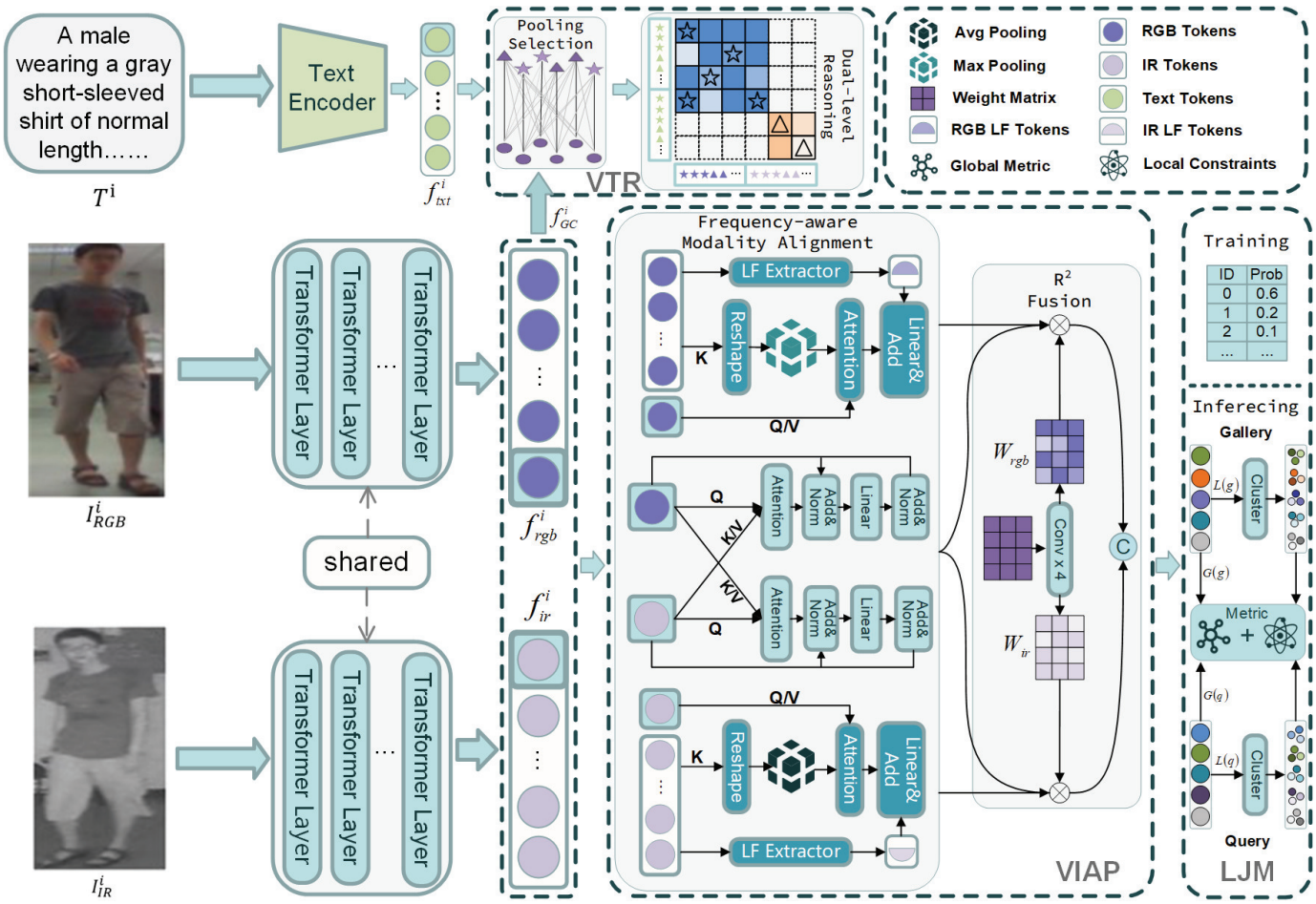

Fig. 2. Overall architecture of the proposed STAR-ReID. It is a three branches structure with VIAP and VTR. RGB and IR inputs are fed into weight-shared Transformer layers for learning modality-invariant features, and then realize visual alignment by supervised VIAP. On this basis, VTR exploits pooling selection and dual-level reasoning to enhance the attention of visual features to local fine-grained clues. And in the inference stage, LJM can be optionally applied to extend measurement dimensions by using local fine-grained information, improving the accuracy of the results. Best viewed in color.

and auxiliary features for dual-level reasoning. The auxiliary features from patch tokens of three modalities, are extracted by pooling selection. In this way, STAR-ReID can enhance fine-grained information in visual modalities by textual guidance. More details are illustrated in the following subsection.

A. Visible-Infrared Alignment Plus

Due to the lack of spatial consistency between RGB and IR images, the modality-invariant features extracted by existing methods generally contain global information. This leads to many local fine-grained features being ignored. Meanwhile, in the infrared modality, low-frequency information contains crucial cues that aid in identification. In response to these issues, frequency-aware modality alignment and R^2 fusion that can mine local salient cues are designed in VIAP.

Taking a triple input (I^i_{RGB}, I^i_{IR}, T^i) as an example, the basic visible feature f^i_{rgb} and infrared feature f^i_{ir} are extracted by backbone with shared weights. And then decompose each basic feature into a CLS token $G(f^i_{rgb/ir})$ and a patch token $L(f^i_{rgb/ir})$. Using CLS tokens as inputs, frequency-aware modality alignment first learns inter-modality global features $f^i_{Grgb/Gir}$ by (1) and (2).

$$f^i_{Grgb} = Softmax\left(\frac{G(f^i_{rgb})G(f^i_{ir})^{\mathsf{T}}}{\sqrt{d_{ir}}}\right)G(f^i_{ir}) \tag{1}$$

$$f^i_{Gir} = Softmax\left(\frac{G(f^i_{ir})G(f^i_{rgb})^{\mathsf{T}}}{\sqrt{d_{rgb}}}\right)G(f^i_{rgb}) \tag{2}$$

In (1) and (2), $d_{rgb/ir}$ means the dimension of $G(f^i_{rgb/ir})$. Through this calculation, the obtained global feature f^i_{Grgb} and f^i_{Gir} can improve the inter-modality similarity. In modality alignment, these are equally crucial as the original CLS tokens.

In addition, considering that both I^i_{RGB} and I^i_{IR} in the a training sample may not be captured simultaneously by the same device, intra-modality features should also be learned. The frequency-aware modality alignment take basic feature $f^i_{rgb/ir}$ as input, calculating intra-modality global-local feature f^i_{GLrgb} and f^i_{GLir} according to (3) and (4).

$$f^i_{GLrgb} = Softmax\left(\frac{G(f^i_{rgb})RMax(L(f^i_{rgb}))^{\mathsf{T}}}{\sqrt{d_{Lrgb}}}\right)G(f^i_{rgb}) \tag{3}$$

$$f^i_{GLir} = Softmax\left(\frac{G(f^i_{ir})RAvg(L(f^i_{ir}))^{\mathsf{T}}}{\sqrt{d_{Lir}}}\right)G(f^i_{ir}) \tag{4}$$

In this part, the CLS tokens act as query and value, while patch tokens serve as the key. $RMax()$ denotes max pooling after reshaping, $RAvg()$ denotes average pooling after reshaping, $d_{Lrgb/Lir}$ indicates the dimension of $L(f^i_{rgb/ir})$ after pooling. Due to the dimensional differences between local vector $L(f^i_{rgb/ir})$ and global token $G(f^i_{rgb/ir})$, a reshape operation is required to change $L(f^i_{rgb/ir})$ from $D{\times}HW$ to $D{\times}H{\times}W$. Then the reshaped vectors are conducted in the pooling operation. For visible image I^i_{RGB}, max pooling is adopted to highlight discriminative color or texture details. While for infrared image I^i_{IR}, average pooling is used to consolidate robust shape context. In this way, the achieved f^i_{GLrgb} and f^i_{GLir} can reflect global-local relationships, thus facilitating modality alignment.

To further enhance feature discrimination, especially in infrared modality, local low-frequency information is extracted to facilitate fine-grained cross-modality alignment, as shown in (5) and (6),

$$c^i_{rgb} = H_l(L(f^i_{rgb})) \tag{5}$$

$$c^i_{ir} = H_l(L(f^i_{ir})) \tag{6}$$

where H_l denotes the Haar wavelet low-frequency extractor, and $c^i_{rgb/ir}$ represents the corresponding low-frequency tokens.

Subsequently, low-frequency tokens are reshaped and subjected to average or max pooling, followed by separate projectors to enhance the cross-modality global-local features, as depicted in (7) and (8),

$$f^i_{GLFrgb} = Proj_{rgb}(RMax(c^i_{rgb})) + f^i_{GLrgb} \tag{7}$$

$$f^i_{GLFir} = Proj_{ir}(RAvg(c^i_{ir})) + f^i_{GLir} \tag{8}$$

where $Proj_{rgb/ir}$ denotes the linear projector of low-frequency feature, f^i_{GLFrgb} and f^i_{GLFir} represent the features strengthened by low-frequency information.

At this moment, VIAP can obtain the original global features $G(f^i_{rgb/ir})$, inter-modality features f^i_{Grgb} and f^i_{Gir}, intra-modality features f^i_{GLFrgb} and f^i_{GLFir}. They are adaptively fused by R^2 fusion, and the weights $W_{rgb/ir}$ are learned from the STAR-ReID. As shown in the subgraph of figure 2, firstly initialize a 2*3*768 one matrix, and then perform 4 layers of 1*1 convolution to obtain weights suitable for RGB and IR modalities. The calculation of the fusion features f^i_{gl} is defined as,

$$\begin{aligned} f^i_{gl} &= Cat(Sum(W^0_{rgb} \cdot f^i_{Grgb}, W^1_{rgb} \cdot G(f^i_{rgb}), W^2_{rgb} \cdot f^i_{GLFrgb}), \\ &\quad Sum(W^0_{ir} \cdot f^i_{Gir}, W^1_{ir} \cdot G(f^i_{ir}), W^2_{ir} \cdot f^i_{GLFir})) \end{aligned} \tag{9}$$

where $Cat(a, b)$ means a concatenate with b, $Sum(a, b, c)$ represents matrix addition. The integrated f^i_{gl} reflects inter-modality and global-local alignment, which effectively improves the discrimination of modality-invariant features. Table I summarizes the notations used above.

TABLE I
SUMMARY OF NOTATIONS

Notation	Description
I^i_{RGB}	visible image
I^i_{IR}	infrared image
T^i	natural language description
f^i_{rgb}	visible feature
f^i_{ir}	infrared feature
$f^i_{Grgb/Gir}$	inter-modality global feature
$d_{rgb/ir}$	dimension of inter-modality global feature
$f^i_{GLrgb/GLir}$	intra-modality global-local feature
$f^i_{GLFrgb/GLFir}$	low-frequency strengthened global-local feature
$d_{Lrgb/Lir}$	dimension of intra-modality global-local feature
$c^i_{rgb/ir}$	low-frequency component
$W_{rgb/ir}$	fusion weights
$G(f)$	extract CLS token from f
$L(f)$	extract patch tokens from f
$RAvg(f)$	reshape and apply average pooling to the feature f
$RMax(f)$	reshape and apply max pooling to the feature f
$H_l(f)$	extract low-frequency token from f
$Proj_{rgb/ir}(f)$	extract projected features from f

B. Visual-Textual Reasoning

In retrieval tasks, natural language descriptions that include gender or clothing can help quickly locate the object range by local salient clues. Thus, this paper introduces natural language specification to improve VI-ReID, which is also the first re-identification work that simultaneously handles RGB, IR, and text. However, existing VI-ReID datasets lack textual descriptions. To address this, we coarsely annotate unified attributes (e.g., bag, clothing) for images with the same ID and upload them in JSON format to GPT-4 for generating standardized language descriptions.

Using the above natural language specifications to guide visual images to extract fine-grained features inevitably requires cross-modality learning between vision and language. Considering the significant differences between these two modalities and the lack of obvious fine-grained information in the infrared modality. VTR is proposed to conduct relation reasoning between textual f^i_{txt} and the visual fusion of f^i_{rgb} and f^i_{ir}. VTR contains a pooling selection and dual-level reasoning. The pooling selection extracts auxiliary feature PTokens $P(f^i_{rgb/ir/txt})$ from patch tokens by channel-wise max pooling, which provides local features for relation reasoning. Unlike existing reasoning methods that only depend on CLS tokens, dual-level reasoning exploits CLS tokens and PTokens for global and local vision-language supervision. This can implicitly assist VIAP in enhancing the learning of fine-grained features.

In vision-language supervision, global visual-to-textual relation reasoning adopts the successful contrastive loss L^G_{V2T} defined in (10),

$$L^G_{V2T} = -\log \frac{exp(S(f^i_{GC}, G(f^i_{txt})))}{\sum_{k=1}^{N} exp(S(f^i_{GC}, G(f^k_{txt})))} \tag{10}$$

where f^i_{GC} refer to $Cat(G(f^i_{rgb}), G(f^i_{ir}))$, $S(a, b)$ computes the similarity between a and b, N denotes the batchsize. Local visual-to-textual loss L^L_{V2T} only needs to replace CLS tokens in (10) with PTokens $P(f^i_{rgb/ir/txt})$.

When computing textual-to-visual loss, we find that any triple input in a batch probably has more than one positive. This means that the same description T^i may correspond to different (I^i_{RGB}, I^i_{IR}). Facing these positives, the better the reasoning between vision and language, the smaller the contrastive loss. VTR should pay more attention to the training samples with large errors, as they are difficult to prompt between textual and visual modalities. Therefore, a reasoning power w^i_p is designed to modify contrastive loss as L_{T2V}. Taking global constraint as an example, the relevant calculations are defined in (11) and (12),

$$w^i_p = \frac{L_C(f^i_{GC}, G(f^i_{txt}))}{\sum_{p\in P(T)} L_C(f^p_{GC}, G(f^i_{txt}))} \tag{11}$$

$$L^G_{T2V} = -w^i_p log \frac{exp(S(f^i_{GC}, G(f^i_{txt})))}{\sum_{k=1}^{N} exp(S(f^k_{GC}, G(f^i_{txt})))} \tag{12}$$

where L_C refers to the standard contrastive loss, $P(T)$ is the set of indices of all positives for T^i in the batch, w^i_p stands for the normalized weight that reflects the impact on VTR updates. The same calculation is applies to L^L_{T2V} with $P(f^i_{rgb/ir/txt})$. They work together to help VTR implement hard example mining. The total loss of vision-language supervision L_{VLS} is be defined as (13),

$$L_{VLS} = L^G_{V2T} + \alpha L^L_{V2T} + L^G_{T2V} + \alpha L^L_{T2V} \tag{13}$$

where α is set to 0.3 based on experience. And then the object function L_{OB} is summarized as,

$$L_{OB} = L_{id} + L_{tri} + L_{VLS} \tag{14}$$

where L_{tri} represents the triplet loss function with f^i_{gl}, L_{id} is the sum of cross entropy loss function using $G(f^i_{rgb/ir})$. They are jointly responsible for updating STAR-ReID.

C. Local-Global Joint Measure

Besides paying sufficient attention to local information in the feature part, the joint measure of global and local features is also used in the retrieval phase.

In this phase, a person image q from query and a gallery g containing a large number of images are given, and the purpose of this phase is to match a large number of images from the gallery to a person image similar to the selected q.

The common retrieval strategy is to input the images into the trained model to extract features and get the retrieval list based on the feature similarity of q and g. In a unimodal re-recognition task, it is necessary to solve the recognition error problem caused by the overall similarity but not the local feature similarity. And in cross-modal retrieval, the impact of this lack of attention to local details can be even greater.

Therefore, in the cross-modal re-identification task, there is a need to further cross the impact of feature similarity metric brought by modal differences and further optimize the ranking strategy. Figure 3 show the ranking list.

We utilize the joint global-local optimization idea. The neighboring sample similarities based on global features and pooled local features are obtained separately, and then the two sample similarities are summed and fused in the form of a distance matrix to obtain a joint constrained distance matrix for ranking retrieval.

Fig. 3. Visual comparison with and without LJM. The first column is the query images. Columns 2–11 are the top ten results. Green box indicates consistency with groundtruth, while the red box denotes error. The yellow markings indicate local cues that play a role in LJM. Best viewed in color.

The distance matrix is first obtained at the global feature level by the similarity, and the overall pose is used as a constraint to ensure the recognition accuracy. And for local feature, the modal difference between infrared and visible images leads to the modal imbalance problem. Due to the excessive semantic information of visible features, the infrared features may not be matched accurately even for the same identity when performing feature metrics. To solve this retrieval imbalance caused by modal differences, we perform a pooling operation for local feature of both modalities to reduce the effect of complex background noise of visible features while enhancing the foreground information of the person contained in infrared features.

Thus, the constraint operation for the local features is specifically described as follows: we apply average pooling to the local features extracted from the query and gallery, and use the k-reciprocal method to obtain the k-nearest neighbors of the gallery and query images based on the pooled features. Finally, these two matrices are summed and fused with specific weights to achieve our joint global-local measurement strategy, which is only used during the inference stage. This strategy including local constraints can be defined as:

$$D_{Joint} = D_{Global} + \beta R_{Local} \tag{15}$$

$$R_{Local} = 0.3 \cdot \Delta_{Ja}(q_l, g_l, K) + 0.7 \cdot \Theta(q_l, g_l) \tag{16}$$

where D_{Joint} represent joint metric constraint, D_{Global} represent global metric, R_{Local} represent local constraints, The parameters β used to represents weighting factors for global and local features. Δ_{Ja} denotes using local features to compute the Jaccard distance, Θ is the cosine distance, q_l and g_l denotes query and gallery of local feature. K represents the number of samples selected based on distance, empirically set to 10. Through Δ_{Ja}, sample expansion is enabled using local features, converting single-to-single similarity measurement into multi-to-multi group retrieval.

The use of local features in VI-ReID helps address spatial and semantic inconsistencies caused by occlusion or viewpoint changes, as representative local features offer better discriminative power and provide accurate reference information. Additionally, group-based metric optimization enhances performance in complex scenarios.

As shown in Figure 3, the first and third rows are the initial VI-ReID results, while the second and fourth rows are the results via LJM. From these results, it is evident that local fine-grained features such as backpacks, hats, and water bottles are taken into consideration into LJM along with global features. The LJM improves initial results with global-local similarity.

Algorithm 1 STAR-ReID Processing Procedure

Input: Triple input (I^i_{RGB}, I^i_{IR}, T^i) with the same ID
Parameter: Textual branch $\mathcal{T}_{txt}$, I^i_{RGB} and I^i_{IR} use shared visual branch $\mathcal{T}_{img}$
1: Initialization part of $\mathcal{T}_{img}$ from *ViT*
Initialization part of $\mathcal{T}_{txt}$ from *BERT*
2: **while** in train stage **do**
3: Extract $f^i_{rgb/ir} = \mathcal{T}_{img}(I^i_{RGB/IR})$, $f^i_{txt} = \mathcal{T}_{txt}(T^i)$
4: Calc. inter-modality global features f^i_{Grgb} and f^i_{Gir} according to (1)–(2)
5: Calc. intra-modality global-local features f^i_{GLrgb} and f^i_{GLir} according to (3)–(4)
6: Calc. low-frequency strengthened intra-modality features f^i_{GLFrgb} and f^i_{GLFir} according to (5)–(8)
7: Achieve fusion feature f^i_{gl} and realize frequency-aware modality alignment according to (9)
8: Achieve PTokens $P(f^i_{rgb/ir/txt})$ and conduct visual-textual dual-level reasoning according to (10)–(13)
9: Optimize $\mathcal{T}_{txt}$, $\mathcal{T}_{img}$ according to (14)
10: **end while**
11: **while** in test stage **do**
12: **if** mix-retrieve **then**
13: Calc. the inner product between $G(f^i_{txt})$ of probe and f^i_{gl} of galley
14: Sort by inner product and retain the Top-N of the gallery
15: Achieve results with f^i_{gl} of retained gallery and probe
16: **end if**
17: **if** local-global joint measure **then**
18: Achieve distance with $RMax(L(f^i_{rgb}))$ and $RAvg(L(f^i_{ir}))$ according to (15)–(16)
19: **end if**
20: Achieve results with f^i_{gl} of gallery and probe
21: **end while**

Ultimately, the whole process procedure of the proposed STAR-ReID, combined with the equations presented above, is demonstrated step by step in Algorithm 1. In the training stage, VIAP and VTR is utilizes to achieve the distinguishable modality-invariant features. VIAP conducts modality alignment, which is described on lines 3-6. The VTR incorporates natural language specifications to enhance attention to local fine-grained features. Its process is shown in lines 7-8. The L_{OB} defined in (14) is responsible for updating parameters.

During the testing phase, this work provides two retrieval modes. One is standard similarity measurement in VI-ReID, and the other is the mix-retrieve with natural language specification. For the second mode, we firstly calculate the inner product of the text tokens $G(f_{txt})$ of the probe and the fusion features f_{gl} of the gallery, and then retain the Top-N gallery images. On this basis, achieve the cosine distance with f_{gl} of the probe and the reserved gallery. The textual descriptions can effectively narrow the search scope, which is crucial for the promotion of STAR-ReID in practical scenarios.

D. Human-LLM Collaborative Natural Language Description Generation

Due to the absence of natural language specifications in existing visible-infrared re-identification datasets such as SYSU-MM01 and RegDB, we annotated these datasets at the pedestrian ID level.

Relying solely on manual annotation requires different annotation personnel to ensure consistent description rules, which is time-consuming and laborious. It is also difficult to obtain the accurate descriptions for the cropped images in ReID datasets by directly using existing Image-to-Text generation algorithms. The reason behind this is that these algorithms lack objective guidance and cannot perform parsing and language organization that is conducive to ReID task.

For this, we perform two-stage annotation. Provide semantic orientation for algorithm-based automatic annotation through manual annotation, while reducing labor and time costs by automatic generation algorithm.

In manual annotation, the rough tags include eleven dimensions: gender, top color, top type, top length, pants color, pants type, bag type, hat type, glasses type, shoe color and notes. The 'notes' dimension is used to describe special patterns, accessories and items to enhance the ability of title description. This tag information will be stored in a structured JSON format using the corresponding dimension names as keys to generate detailed descriptions for GPT-4.

In GPT-4 generation, the previously annotated coarse labels are used as the part of the prompts for GPT-4, which aims to produce smoother natural language descriptions. This turn facilitates subsequent text encoders in semantically supervising the model. The prompt is as follows:

A photo taken by a surveillance camera shows an individual, whose attire characteristics will be described using a structured JSON data as follows: [JSON]. Please succinctly describe the person in the photo in English. In the JSON, the 'top color' field represents the color of this person's clothing; a field value of 'none' indicates that he is not carrying corresponding items; The "top type" field could be either 'short sleeve' or 'long sleeve'; The "top length" field refers to its clothing length, which can be either 'normal' or 'long'; The 'pants color' field is his pants' color. The description should start with 'The person in the photo is'.

Where *[JSON]* is located, it will be replaced with the specific JSON data. It's important to note that the temperature parameter of GPT-4 needs to be adjusted to around 0.6, in order to prevent it from automatically filling in details not present in the image. At the same time, due to limited

TABLE II
THE DISTRIBUTION OF CAPTIONS

Dataset	Quantity	Min WC.	Max WC.
SYSU-MM01	491	29	70
RegDB	412	29	62

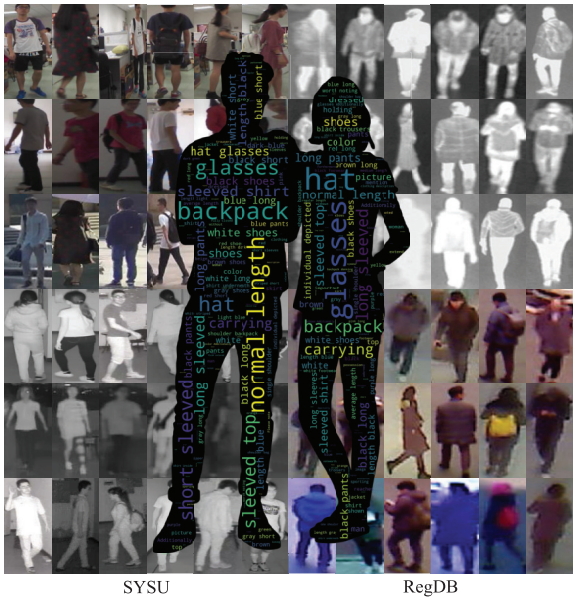

Fig. 4. Word cloud distribution of natural language descriptions on SYSU-MM01 and RegDB datasets. Best viewed in color.

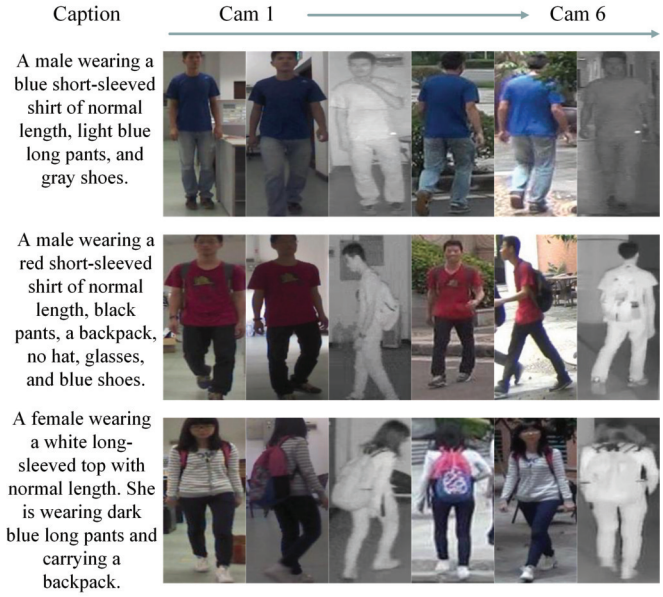

Fig. 5. Sample from SYSU-MM01 dataset. The leftmost part is the description text, and the corresponding silhouette images of the person with ID are from Cam1 to Cam6 from left to right. Best viewed in color.

detail dimensions, there may be situations where different IDs correspond to identical captions after generation is complete. In such cases, manual modification of captions is required for differentiation. These measures help avoid incorrect supervision of network by textual semantic information.

IV. EXPERIMENTS

A. Dataset and Implementation Details

To validate the efficacy of STAR-ReID, we conducted experiments on two benchmark datasets: SYSU-MM01 [1] and RegDB [50]. The SYSU-MM01 dataset comprises 30,071 visible and 15,792 infrared images, encompassing 491 unique pedestrians. In contrast, the RegDB dataset contains 412 identities, with each identity represented by 10 visible and 10 infrared images. Figure 5 presents representative examples from these datasets.

Table II elucidates the distribution of natural language descriptions at the identity level across both datasets. Notably, there is a substantial variance in word count, primarily attributed to the complexity of certain scenarios (e.g., striped clothing, carried objects) that necessitate more detailed descriptions. Figure 4 visualizes the word cloud distribution of captions, with the left and right halves representing SYSU-MM01 and RegDB, respectively.

It is worth noting that SYSU-MM01 presents a more challenging and authoritative benchmark compared to RegDB. All reported results in subsequent tables represent the average of ten independent experimental runs. We employ Rank-1 accuracy (R1) and mean average precision (mAP) as our evaluation metrics.

The proposed STAR-ReID framework was implemented using PyTorch and trained on a single NVIDIA RTX3090Ti GPU. We utilized a batch size of 64 and trained the model for 150 epochs. The optimization process employed the AdamW optimizer in conjunction with a cosine annealing learning rate scheduler. The initial learning rate was set to $3e^{-4}$, with a weight decay of $1e^{-4}$.

B. Comparison With State-of-the-Art Methods

The comparison with outstanding algorithms on the SYSU-MM01 and RegDB datasets is shown in Table III and Table IV, respectively. The SYSU-MM01 dataset involves all-search and indoor-search settings, while the RegDB is evaluated on infrared to visible (I to V) and visible to infrared (V to I). The best and second best results are highlighted in bold and underline.

In Table III, our proposed STAR-ReID method demonstrates superior performance across both search settings. Without the LJM module, STAR-ReID achieves a R1 of 79.92% and an mAP of 77.23% in the all-search setting, as well as a R1 of 86.74% and an mAP of 88.12% in the indoor-search mode, already surpassing previous state-of-the-art methods. Additionally, [17], [18] achieved competitive results by enriching feature representations with local features, while the proposed method further enhances performance through global-local alignment. With the addition of the LJM module (denoted as STAR-ReID (Ours)), the performance further improves to a R1 of 82.93% and an mAP of 80.47% for all-search, and a R1 of 88.04% and an mAP of 89.58% for indoor-search. This significant improvement over both previous methods and our

TABLE III
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE SYSU-MM01 DATASET

Methods	Venue	SYSU-MM01 [1]			
		All Search		Indoor Search	
		R1	mAP	R1	mAP
FMCNet [28]	CVPR'22	66.34	62.51	68.15	74.09
DART [22]	CVPR'22	68.72	66.29	72.52	74.94
MAUM [21]	CVPR'22	71.68	68.79	76.97	81.94
CMT [51]	ECCV'22	71.88	68.57	76.90	79.91
CIFT [52]	ECCV'22	74.08	74.79	81.82	85.61
MSCLNet [53]	ECCV'22	76.99	71.64	78.49	81.17
CAJ+ [54]	TPAMI'23	71.48	68.15	78.36	81.98
ProtoHPE [55]	MM'23	71.92	70.59	77.81	81.31
DEEN [5]	CVPR'23	74.70	71.80	80.30	83.30
PartMix [4]	CVPR'23	77.78	74.62	81.52	84.38
PMWGCN [19]	TIFS'24	66.82	64.88	72.64	76.19
DARD [27]	TIFS'24	69.33	65.65	77.21	81.91
MFCS [56]	TMM'24	70.59	67.49	75.98	80.24
RLE [18]	NIPS'24	75.40	72.40	84.70	87.70
WRIM-Net [17]	ECCV'24	77.40	75.40	86.20	88.10
J-ReID [9]	ICME'24	77.83	75.68	82.52	85.72
STAR-ReID (w/o LJM)		79.92	77.23	86.74	88.12
STAR-ReID (Ours)		**82.93**	**80.47**	**88.04**	**89.58**

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE REGDB DATASET

Methods	Venue	RegDB [50]			
		I to V		V to I	
		R1	mAP	R1	mAP
DART [22]	CVPR'22	81.97	73.78	83.60	75.67
MSCLNet [53]	ECCV'22	83.86	78.31	84.17	80.99
MAUM [21]	CVPR'22	86.95	84.34	87.87	85.09
FMCNet [28]	CVPR'22	88.38	83.86	89.12	84.43
CAJ+ [54]	TPAMI'23	84.88	78.55	85.69	79.70
PartMix [4]	CVPR'23	84.93	82.52	85.66	82.27
ProtoHPE [55]	MM'23	88.69	81.99	88.74	83.72
MFCS [56]	TMM'24	83.88	75.16	85.34	76.39
DARD [27]	TIFS'24	85.53	85.09	86.19	85.39
PMWGCN [19]	TIFS'24	88.77	81.61	90.61	84.53
J-ReID [9]	ICME'24	89.37	87.63	89.42	87.64
STAR-ReID (w/o LJM)		89.52	88.17	90.69	89.57
STAR-ReID (Ours)		**90.48**	**91.77**	**91.89**	**93.31**

own baseline without LJM demonstrates the effectiveness of our proposed approach.

In Table IV, our proposed method significantly outperforms all mentioned works on the mAP. These further demonstrate that VIAP and VTR have effectively improved modality-invariant features.

It is particularly noteworthy that local cues were also explored in competitive PartMix [4]. This proves that local fine-grained alignment plays an important role in VI-ReID. Meanwhile, this indirectly indicates that using text to pay attention to local salient regions is more reasonable than existing exploration of local information.

Moreover, PMWGCN [19] also demonstrates competitive results, as it similarly incorporates low-frequency information, akin to the proposed STAR-ReID. However, the proposed method further explores low-frequency information within local features, mitigating spatial and semantic inconsistencies. Finally, the close integration of proposed VTR, VIAP and LJM leads to the best results.

C. Ablation Study

In this section, an ablation study is performed to evaluate the contribution of each component in STAR-ReID, as shown in Table V. 'Base' represents PMT [57] which is the baseline of our proposed method. 'FMA' refers to frequency-aware modality alignment.

Compared with the baseline, the introduction of VIAP with local low-frequency information extractor further strengthens the feature discriminability, increased R1 and mAP by 7.59% and 9.27%, respectively. The addition of LJM further improved performance, with R1 reaching 76.92% and mAP 75.99%.

Incorporating BERT for text feature extraction also showed notable improvements, increasing R1 to 72.09% and mAP to 70.78%. When combining VTR and VIAP, significantly boosted performance, achieving 79.92% in R1 and 77.23%

TABLE V
ABLATION STUDY ON SYSU-MM01 DATASET

Settings	R1	mAP
Base [57]	67.53	64.98
Base+FMA	74.68	73.09
Base+VIAP	75.12	74.25
Base+VIAP+LJM	76.92	75.99
Base+BERT	72.09	70.78
Base+VTR	72.93	72.21
Base+VTR+VIAP	79.92	77.23
STAR-ReID (Ours)	82.93	80.47
STAR-ReID (Mix-retrieve)	**83.11**	**80.93**

TABLE VI
RESULTS OF DIFFERENT K VALUES IN LJM ON SYSU-MM01 DATASET

K	mAP	mAP INC. (%)	Duration (ms)	Duration INC. (%)
0	77.23	0%	67445	0%
5	78.16	1.20%	69367	2.85%
10	**80.47**	**4.20%**	69610	3.21%
20	80.39	4.09%	70069	3.89%
30	79.34	2.73%	70702	4.83%

TABLE VII
RESULTS OF DIFFERENT K VALUES IN LJM ON REGDB DATASET

K	mAP	mAP INC. (%)	Duration (ms)	Duration INC. (%)
0	88.17	0%	67923	0%
5	90.03	2.11%	69905	2.92%
10	**91.77**	**4.08%**	70118	3.23%
20	91.22	3.46%	70653	4.02%
30	90.15	2.25%	71349	5.04%

in mAP. This indicates that both frequency-aware modality alignment and visual-text reasoning are indispensable for our method.

Our complete method, STAR-ReID, achieves state-of-the-art performance with 82.93% R1 and 80.47% mAP. This demonstrates that each component is essential and contributes to the overall effectiveness of the approach.

In addition, the mix-retrieval results that filter the gallery using probe text are displayed in the last row of Table IV. The Top-N in the retrieval mode is simply set to 50% of the total number of gallery items. It is not difficult to find that using natural language specification can further improve R1 to 83.11% and mAP to 80.93%. In a real scenario with a large amount of gallery data, mixing retrieval will become more effective by narrowing down the search scope.

D. Hyperparameter Analysis

To evaluate the value of LJM, experiments with different K settings and time consumption were conducted on the SYSU-MM01 test set in all search mode and RegDB test set in infrared to visible mode, as shown in Table VI and Table VII.

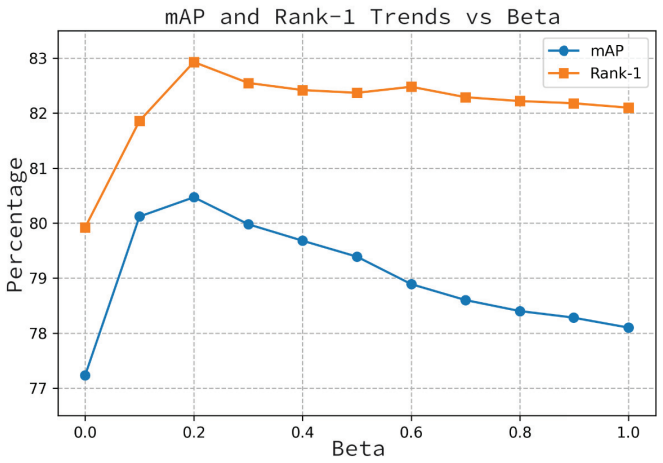

Fig. 6. Weighting factors for global and local feature joint measurement. The blue line represents mAP, the orange line represents Rank-1. $\beta \in [0.0, 1.0]$. Best viewed in color.

As presented in the tables, LJM with different K values generally enhances mAP, owing to its use of local features in constructing group retrieval, which provides reference information during the measurement. However, as K increases, computational costs also rise, making it crucial to select an appropriate K based on the application scenario and gallery size to balance performance and efficiency.

For example, as depicted in Table VI, when $K = 10$, accuracy improved by 4.20% relative to the original result, while inference time increased by 3.21%. At this point, time and accuracy reached a favorable cost-performance ratio. In large-scale scenarios, combining textual descriptions for mix-retrieve can first narrow down the gallery before applying LJM, thus achieving accuracy improvements with manageable time costs while maintaining the main workflow.

The experiment to analyze the impact of weighting factors on LJM also be conducted as illustrated in Figure 6. As β increases from 0 to 0.2, both mAP and Rank-1 metrics improve significantly, indicating that incorporating local features enhances model performance. Beyond $\beta = 0.2$, Rank-1 remains relatively stable while mAP gradually declines. This trend suggests that while maintaining high Rank-1 accuracy, overemphasizing local features may slightly reduce overall ranking quality. The divergence between mAP and Rank-1 curves reveals the trade-off between top-rank accuracy and overall retrieval performance, providing insights for model optimization in various scenarios. Therefore, $\beta = 0.2$ was chosen as the weight of joint measurement.

E. Visualization Analysis

In addition to quantitative analysis, four visualization analysis are presented in this section.

Firstly, for a better understanding the value of Visual-Textual Reasoning (VTR), we visualized the attention features in the last Transformer layer using Grad-CAM [58], as shown in Figure 8. This figure compares textual descriptions (a), raw input images (b), baseline attention maps (c), and attention maps with VTR applied (d) for four sample subjects.

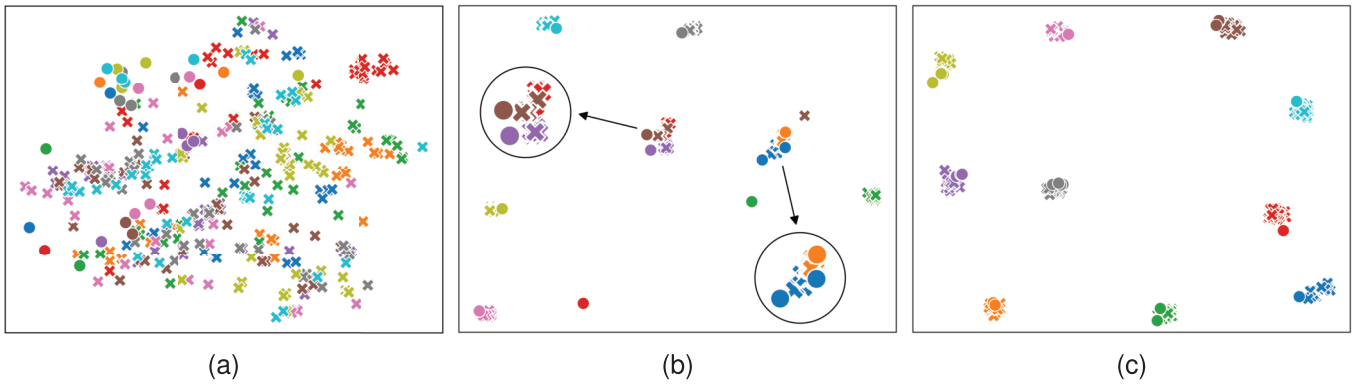

Fig. 7. T-SNE visualization on ten randomly test IDs in SYSU-MM01. (a) Original. (b) Baseline. (c) Ours. The same ID is marked with a unified color. Cross and circle denotes visible and infrared features, respectively. Best viewed in color.

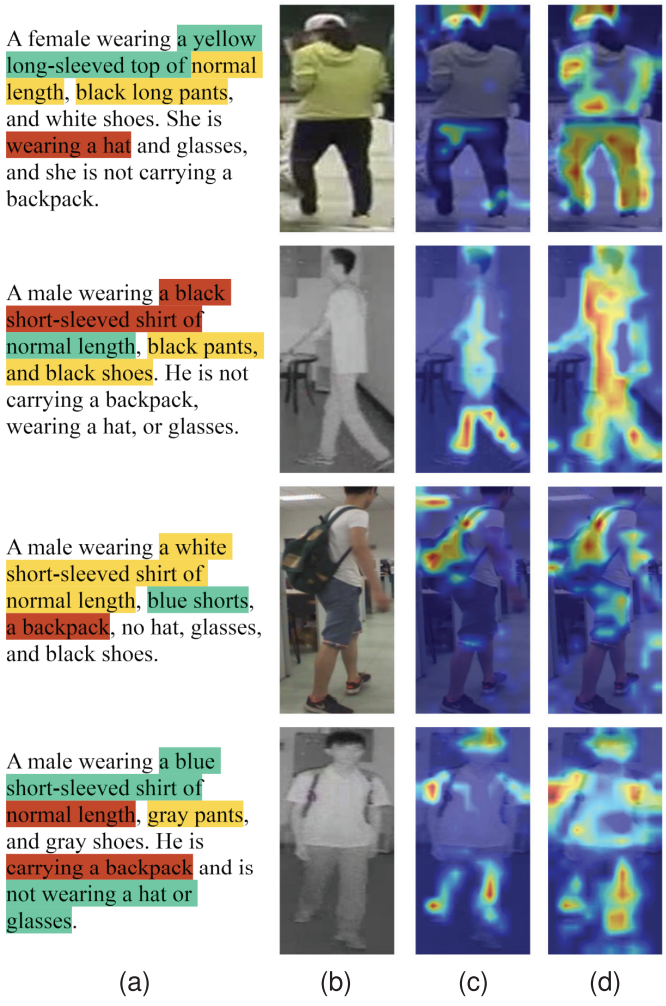

Fig. 8. Attention visualization through Grad-CAM [58].(a) Caption (b) Raw input. (c) Baseline. (d) Baseline + VTR. Different colors highlight guiding descriptions. Yellow is more prominent than green. Best viewed in color.

Comparing columns (c) and (d) in Figure 8, we observe that the text containing attribute descriptions guides the attention to local salient regions through VTR. This is evident in the more focused and precise heat maps in column (d) compared to the more diffuse attention in column (c). The color coding in the textual descriptions (yellow being more prominent than green) corresponds to the intensity of attention in the heat maps, demonstrating how the model prioritizes certain attributes over others.

This attention optimization improves modality-alignment between RGB and IR representations while enhancing fine-grained expression in modality-invariant features. The visualizations demonstrate the significance of introducing natural language into VI-ReID (Visible-Infrared Re-Identification) and the effectiveness of VTR. By leveraging textual descriptions, the model focuses on specific attributes crucial for distinguishing individuals across different imaging modalities, potentially leading to more robust and accurate re-identification results.

To verify the discrimination of modality-invariance feature extracted from STAR-ReID, randomly sample ten IDs in the SYSU-MM01 test set for the qualitative visualization. The features of Baseline [57] and STAR-ReID are plotted with t-SNE [59] in Figure 7.

As shown in Figure 7, we visualize the feature distributions using t-SNE to illustrate the effectiveness of our proposed method. Subfigure (a) presents the original complex data distribution. In (b), many identity clusters from the Baseline model are closely intertwined, indicating difficulty in separating different identities across modalities. The circles highlight examples of this identity confusion.

Contrastingly, subfigure (c) demonstrates the feature distribution achieved by our STAR-ReID model (Baseline + VTR). Here, the feature clusters for each particular identity are ideally separated with clear boundaries. This separation indicates that STAR-ReID has effectively improved the discrimination of modality-invariant features through our proposed VIAP and VTR modules.

The comparison between (b) and (c) provides strong evidence for the superior performance of STAR-ReID in learning discriminative and modality-invariant features, addressing key challenges in cross-modality person re-identification.

Figure 9 showcases the effect of LJM on sample distance distributions in VI-ReID. The method effectively increases the separation between positive and negative samples, with the peak of negative samples (red curve) shifting from 0.218 to 0.271. This enhancement is achieved solely by optimizing similarity measures, without altering the feature space. The

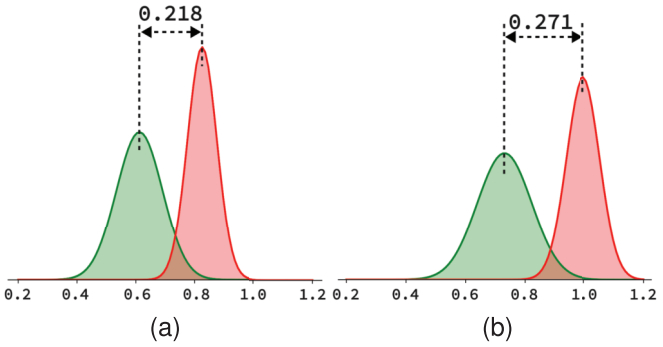

Fig. 9. Distribution of distances between samples before/after adding local features. The green curve represents positive samples, while the red curve represents negative samples. (a) common measurement. (b) the proposed LJM. Best viewed in color.

widened gap demonstrates that local fine-grained features guided by natural language are distinguishable, highlighting LJM's significance for VI-ReID. Such improved distribution suggests better accuracy in real-world scenarios, potentially reducing false positives and boosting system performance. This visual evidence underscores LJM's efficacy in advancing VI-ReID through refined similarity measurement techniques.

V. Conclusion

In this paper, we propose a novel STAR-ReID framework to learn discriminative representations for VI-ReID. It is the first one to introduce natural language specification to mine latent fine-grained cues between visible and infrared modalities. It can effectively conduct multi-modality learning through VIAP and VTR. Specifically, VIAP can enhance the visual-infrared interaction and global-local information fusion, which explicitly improves modality-invariant features. VTR is responsible for vision-language learning, which implicitly pays more attention to significant regions in vision according to natural language.

Furthermore, we introduce the LJM to enhance the metric by incorporating local feature distances, thereby improving the relevance of matching using fine-grained information. We also propose a human-LLM collaborative approach to incorporate textual descriptions into existing cross-modal person re-identification datasets, leveraging the power of human expertise and large language models.

Extensive experiments on SYSU-MM01 and RegDB datasets prove that STAR-ReID outperforms most state-of-the-art methods, with the mix-retrieve with text approach showing particular promise for real-world multi-gallery scenarios. The proposed framework shows significant potential in improving person identification under low-light conditions, and the introduction of natural language specifications provides possibilities for integration with large language models or AI agents, and aligns better with real-world interaction scenarios for locating target persons, making it highly significant for advancing intelligent video surveillance and cross-modal security analytics. And we believe that the multi-modality joint learning, combining visual, infrared, and textual information, will be the future trend for person re-identification, opening new avenues for improving the accuracy and robustness of VI-ReID systems.

References

[1] A. Wu, W. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. ICCV*, Oct. 2017, pp. 5380–5389.

[2] W. Li et al., "DC-former: Diverse and compact transformer for person re-identification," in *Proc. AAAI*, vol. 37, Jun. 2023, pp. 1415–1423.

[3] H. Rao and C. Miao, "TranSG: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22118–22128.

[4] M. Kim, S. Kim, J. Park, S. Park, and K. Sohn, "PartMix: Regularization strategy to learn part discovery for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18621–18632.

[5] Y. Zhang and H. Wang, "Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2153–2162.

[6] S. Yan, N. Dong, L. Zhang, and J. Tang, "CLIP-driven fine-grained text-image person re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 6032–6046, 2023.

[7] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2787–2797.

[8] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with K-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1318–1327.

[9] N. Jiang, Y. Qiu, W. Song, J. Liu, Z. Shi, and L. Wang, "Joint visual-textual reasoning and visible-infrared modality alignment for person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2024, pp. 1–6.

[10] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 1528–1535.

[11] Y. Shi, Z. Wei, H. Ling, Z. Wang, J. Shen, and P. Li, "Person retrieval in surveillance videos via deep attribute mining and reasoning," *IEEE Trans. Multimedia*, vol. 23, pp. 4376–4387, 2021.

[12] Y. Shi et al., "Adaptive and robust partition learning for person retrieval with policy gradient," *IEEE Trans. Multimedia*, vol. 23, pp. 3264–3277, 2021.

[13] H. Liu, S. Ma, D. Xia, and S. Li, "SFANet: A spectrum-aware feature augmentation network for visible-infrared person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 4, pp. 1958–1971, Apr. 2023.

[14] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin, "Structure-aware positional transformer for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 2352–2364, 2022.

[15] Z. Cui, J. Zhou, and Y. Peng, "DMA: Dual modality-aware alignment for visible-infrared person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2696–2708, 2024.

[16] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 4610–4617.

[17] Y. Wu, L. Meng, Y. Zichao, S. Chan, and H. Wang, "Wrim-net: Wide-ranging information mining network for visible-infrared person re-identification," in *Proc. 18th Eur. Conf. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, vol. 15111, Milan, Italy, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham, Switzerland: Springer, 2024, pp. 55–72, doi: 10.1007/978-3-031-73668-1.

[18] L. Tan, "RLE: A unified perspective of data augmentation for cross-spectral re-identification," in *Proc. 38th Annu. Conf. Neural Inf. Process. Syst.*, 2024, pp. 1–12. [Online]. Available: https://openreview.net/forum?id=Ok6jSSxzfj

[19] R. Sun, L. Chen, L. Zhang, R. Xie, and J. Gao, "Robust visible-infrared person re-identification based on polymorphic mask and wavelet graph convolutional network," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2800–2813, 2024.

[20] J. Feng, A. Wu, and W.-S. Zheng, "Shape-erased feature learning for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22752–22761.

[21] J. Liu, Y. Sun, F. Zhu, H. Pei, Y. Yang, and W. Li, "Learning memory-augmented unidirectional metrics for cross-modality person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19366–19375.

[22] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng, "Learning with twin noisy labels for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Aug. 2022, pp. 14308–14317.

[23] K. Ren and L. Zhang, "Implicit discriminative knowledge learning for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 393–402, doi: 10.1109/cvpr52733.2024.00045.

[24] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Nov. 2019, pp. 3623–3632.

[25] Z. Zhang, S. Jiang, C. Huang, Y. Li, and R. Y. D. Xu, "RGB-IR cross-modality person ReID based on teacher–student GAN model," *Pattern Recognit. Lett.*, vol. 150, pp. 155–161, Oct. 2021.

[26] X. Fan, W. Jiang, H. Luo, and W. Mao, "Modality-transfer generative adversarial network and dual-level unified latent representation for visible thermal person re-identification," *Vis. Comput.*, vol. 38, no. 1, pp. 279–294, Jan. 2022.

[27] Z. Wei, X. Yang, N. Wang, and X. Gao, "Dual-adversarial representation disentanglement for visible infrared person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2186–2200, 2024.

[28] Q. Zhang, C. Lai, J. Liu, N. Huang, and J. Han, "FMCNet: Feature-level modality compensation for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, May 2022, pp. 7349–7358.

[29] X. Mao, Q. Li, and H. Xie, "AlignGAN: Learning to align cross-domain images with conditional generative adversarial networks," 2017, *arXiv:1707.01400*.

[30] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 677–683.

[31] D. Liu et al., "Generative metric learning for adversarially robust open-world person re-identification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1, pp. 1–19, Jan. 2023, doi: 10.1145/3522714.

[32] J. Achiam et al., "Gpt-4 technical report," 2023, *arXiv:2303.08774*.

[33] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1970–1979.

[34] Z.-J. Zha, M. Wang, J. Shen, and T.-S. Chua, *Text Mining in Multimedia*. Boston, MA, USA: Springer, 2012, pp. 361–384, doi: 10.1007/978-1-4614-3223-4_11.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[39] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–23, May 2020.

[40] S. Li, M. Cao, and M. Zhang, "Learning semantic-aligned feature representation for text-based person search," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2724–2728.

[41] C. Gao et al., "Contextual non-local alignment over full-scale representation for text-based person search," 2021, *arXiv:2101.03036*.

[42] Y. Wu, Z. Yan, X. Han, G. Li, C. Zou, and S. Cui, "LapsCore: Language-guided person search via color reasoning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Los Alamitos, CA, USA, Oct. 2021, pp. 1604–1613.

[43] P. Zeng et al., "Relation-aware aggregation network with auxiliary guidance for text-based person search," *World Wide Web*, vol. 25, no. 4, pp. 1565–1582, Jul. 2022.

[44] X. Han, S. He, L. Zhang, and T. Xiang, "Text-based person search with limited data," 2021, *arXiv:2110.10807*.

[45] X. Shu et al., "See finer, see more: Implicit modality alignment for text-based person retrieval," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2022, pp. 624–641.

[46] S. Yan, N. Dong, L. Zhang, and J. Tang, "Clip-driven fine-grained text-image person re-identification," 2022, *arXiv:2210.10276*.

[47] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[48] Y. Bai et al., "RaSa: Relation and sensitivity aware representation learning for text-based person search," 2023, *arXiv:2305.13653*.

[49] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–21.

[50] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[51] K. Jiang, T. Zhang, X. Liu, B. Qian, Y. Zhang, and F. Wu, "Cross-modality transformer for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 480–496.

[52] X. Li et al., "Counterfactual intervention feature transfer for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 381–398.

[53] Y. Zhang, S. Zhao, Y. Kang, and J. Shen, "Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification," in *Proc. ECCV*, 2022, pp. 462–479.

[54] M. Ye, Z. Wu, C. Chen, and B. Du, "Channel augmentation for visible-infrared re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2299–2315, Apr. 2024.

[55] G. Zhang, Y. Zhang, and Z. Tan, "ProtoHPE: Prototype-guided high-frequency patch enhancement for visible-infrared person re-identification," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 944–954.

[56] X. Yang, W. Dong, M. Li, Z. Wei, N. Wang, and X. Gao, "Cooperative separation of modality shared-specific features for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 26, pp. 8172–8183, 2024.

[57] H. Lu, X. Zou, and P. Zhang, "Learning progressive modality-shared transformers for effective visible-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 1835–1843.

[58] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2019.

[59] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.