DEEP: Decoupled Semantic Prompt Learning, Guiding and Embedding for Multi-Spectral Object Re-Identification

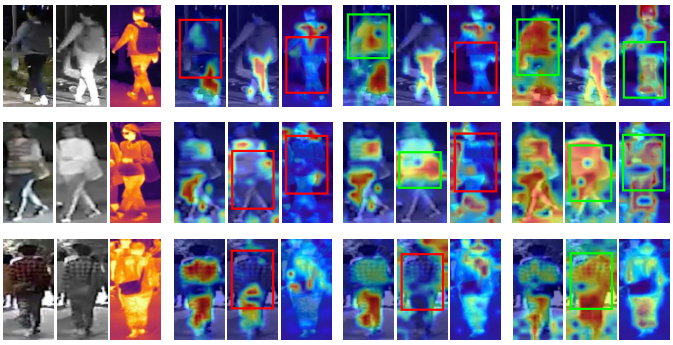
Shihao Li, Chenglong Li, Aihua Zheng*, Jin Tang, Bin Luo, *Senior Member, IEEE*

***Abstract*—Multi-spectral object re-identification (ReID) captures diverse object semantics to robustly recognize identity in complex environments. However, without explicit semantic guidance (*e.g.*, attributes, masks, and keypoints), existing modal fusion-based methods struggle to comprehensively capture person or vehicle semantics across spectra. Thanks to the large-scale vision-language pre-training, CLIP effectively aligns visual concepts across different image modalities to a unified semantic prompt. In this paper, we propose *DEEP*, a *DE*coupled s*E*mantic *P*rompt Learning, Guiding and Embedding framework for Multi-Spectral Object ReID. Specifically, to address the challenges posed by low-quality modality noise and spectral style discrepancies, we first propose a Decoupled Semantic Prompt (DSP) strategy, which explicitly decouples the semantic alignment into spectral-style learning with spectral-shared prompts and object content learning with instance-specific inversion token. Second, to lead the model focusing on semantically faithful regions, we propose a Semantic-Guided Spectral Fusion (SGSF) module that builds a semantic interaction bridge between spectra to explore complementary semantics across modalities. Finally, to further empower the spectral representation, we propose a Spectral Semantic Embedding (SSE) module constrained by semantic-aware structural consistency to refine the fine-grained identity semantics in each spectrum. Extensive experiments on five public benchmarks, RGBNT201, Market-MM, MSVR310, WMVEID863, and RGBNT100, demonstrate the proposed method outperforms the state-of-the-art methods. The source code is released at this link: https://github.com/lsh-ahu/DEEP-ReID.**

***Index Terms*—Multi-Spectral Object Re-Identification, Decoupled Semantic Prompt Learning, Vision-Language Foundation Model.**

I. Introduction

MULTI-spectral object re-identification (ReID) aims to recognize target identity in complex scenarios with adverse weather and illumination changes [2]–[5]. Compared to traditional object ReID [6]–[12], spectra with different imaging advantages contain rich identity semantics. However, significant modal disparity makes it challenging to effectively harness multiple spectra. To address this challenge, HAMNet [13] and CCNet [14] apply feature normalization to learn consistent spectral feature distributions. PFNet [15] and IEEE [16] design multi-modal fusion modules to fuse spectral features. TOP-ReID [17] and EDITOR [18] leverage cross-attention mechanisms to combine spectral features and select key tokens from each spectrum. DeMo [19] employs a mixture-of-experts to decouple spectral features, while MambaPro [20] leverages the selective state space model in Mamba [21] to construct a spectral modal feature aggregation model. PromptMA [22] proposes a visual prompt-based modality interaction strategy to aggregate complementary information across modalities



(a) IDs (b) CLIP-ReID (c) PromptSG (d) DEEP

Fig. 1. Semantic details comparison of prompt learning methods in Grad-CAM visualization [1]. Red marks denote the semantic missing regions. Green marks indicate regions where fine-grained semantic details are successfully captured.

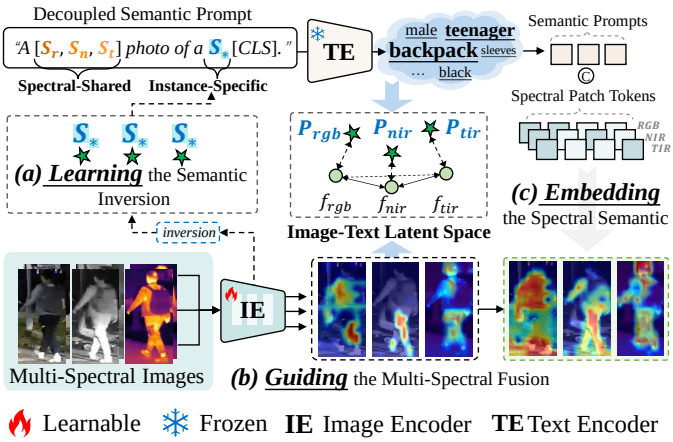


Fig. 2. Illustration of our proposed method. We aim to mine and harness spectral semantics by (a) ***Learning*** the decoupled semantic prompt, (b) ***Guiding*** the multi-spectral fusion, and (c) ***Embedding*** the spectral semantics to effectively extract the discriminative identity features across various spectra.

This research is supported in part by the National Natural Science Foundation of China under Grants 62372003 and 62376004, the University Synergy Innovation Program of Anhui Province under Grant GXXT-2022-036, and the Natural Science Foundation of Anhui Province under Grants 2308085Y40 and 2208085J18. (*The corresponding author is Aihua Zheng.)

C. Li, A. Zheng, and S. Li are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Security Artificial Intelligence, School of Artificial Intelligence, Anhui University, Hefei, 230601, China (e-mail: lcl1314@foxmail.com; ahzheng214@foxmail.com; shli0603@foxmail.com).

J. Tang and B. Lou are with Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China (e-mail: ahu_tj@163.com; ahu_lb@163.com).

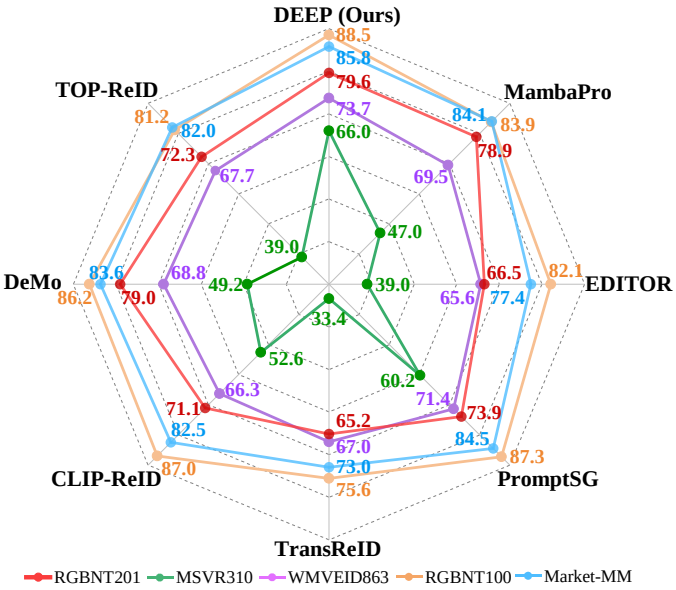


Fig. 3. The mAP metric comparison of SOTA methods on five multi-spectral ReID benchmarks. DEEP not only performs excellently on the person datasets (*i.e.*, RGBNT201, Market-MM) but also demonstrates outstanding performance on vehicle datasets (*i.e.*, MSVR310, WMVEID863, and RGBNT100).

and bridge the distribution gap. However, these methods generally lack explicit spectral semantic guidance, such as attributes [23]–[25], masks [26]–[29], and keypoints [30]–[32], making it challenging to accurately focus on the semantically relevant region of the person or vehicle. Unlike traditional pre-training models, based on large-scale vision-language pre-training, CLIP [33] effectively aligns visual concepts [34], [35] in images with the learnable semantic prompts [36], [37] without concrete text semantic annotation. Existing methods [38], [39] effectively leverage the prompt learning capability of CLIP [33] for single-spectral tasks. However, the multi-spectral object ReID suffers from low-quality spectral noise and spectral style discrepancies, which introduce semantic noise and incomplete semantics during prompt learning. These challenges significantly degrade the semantic alignment performance of CLIP-based methods [38], [39], making it difficult to capture fine-grained identity details of person or vehicle, as illustrated in Fig.1 (b) and (c). Therefore, we raise the question: *Can we further prompt spectral semantics and guide the model to comprehensively capture object identity features?*

Despite the success of prompt learning [34]–[39] in transferring CLIP [33] to standard ReID, research in the multi-spectral research field is still limited. The two-stage identity-level prompt strategy proposed in CLIP-ReID [38] struggles to adapt to multi-spectral data in the first stage and fails to further align the optimized spectral features in the second stage, leading to a significant semantic gap during training. On the other hand, PromptSG [39] aligns only single-modal data through instance-level inverted prompts, lacking the ability to adapt to unseen spectral styles and failing to capture complementary semantics across modalities. **To address this issue, we propose the Decoupled Semantic Prompt (DSP) strategy to explicitly decouple the semantic alignment into spectral-style learning with spectral-shared prompts and object content learning with instance-specific inversion tokens.** As shown in Fig. 2 (a), it encapsulates spectral styles into universal semantic prompts, avoiding interference of style information on instance content.

Leveraging the pre-trained image-text alignment space, the rough prompts guide the model focusing on identity-related regions within each spectra. However, as indicated by the red marks in Fig. 1, significant spectral discrepancies and low-quality noise lead to the loss of identity-related semantics within each spectrum, making it difficult for the model to fully capture visual semantics from a single spectrum. For example, as shown in Fig. 1 (b), CLIP-ReID [38] fails to focus on the person in the unseen thermal infrared modality, demonstrating the poor generalization of its two-stage alignment strategy. In Fig. 1 (c), although PromptSG [39] attends to more regions of the human body in the infrared spectrum, it still suffers from incomplete semantics in regions with heavy spectral noise and significant style discrepancy, indicating its limitations in multi-spectral data. **To solve this, we propose the Semantic-Guided Spectral Fusion (SGSF) module, which establishes the semantic fusion bridge between spectra, as illustrated in Fig. 2 (b).** This approach helps the spectral semantics focus on identity features from other spectra and extract complementary information between them.

Benefiting from visual-text multi-modal alignment, methods like PromptSG [39], MP-ReID [40] and TVI-LFM [41] are flexibly embedding the textual semantics in visual features to improve the retrieval performance. However, existing semantic embedding strategies often overlook the complementarity of semantically relevant regions across modalities. In addition, most TOP-K-based feature selection and merging methods [18], [22], [42] lack semantic guidance, and their simplistic token-dropping operations can disrupt the intra-modal structural consistency of the object identity. **Therefore, we propose the Spectral Semantic Embedding (SSE) module that enforces the structural consistency of identity representations.** The SSE unifies semantic attention across modalities and flexibly aligns semantically correlated local regions by a local feature rearrange strategy to preserve the integrity of object structure across different spectra, as shown in Fig. 2 (c).

Overall, as illustrated in Fig. 3, our method achieves state-of-the-art performance on multiple benchmarks. The main contributions of this paper are summarized as follows:

- **Learning.** We propose a novel prompt learning framework to learn instance semantics for multi-spectral object ReID, which is based on a Decoupled Semantic Prompt (DSP) strategy that adaptively captures semantics in different spectra during training and inference.
- **Guiding.** To address the semantic loss within spectra caused by spectral disparity and low-quality noise, we propose a Semantic-Guided Spectral Fusion (SGSF) module that constructs a semantic interaction bridge to extract complementary information between spectra.
- **Embedding.** To further enhance spectral representation, we propose a Spectral Semantic Embedding (SSE) module constrained by structural consistency to effectively

fusion semantic information into spectral representations.
- Extensive experiments are conducted on five benchmarks to validate the effectiveness of our method. The results demonstrate that the proposed method significantly outperformed the state-of-the-art approaches.

II. Related Work

A. Multi-Spectral Object ReID

Multi-spectral object ReID provides a robust recognition solution for complex real-world scenarios to address various lighting challenges. Compared to traditional single-modal methods, it offers advantages in strong robustness and multi-spectral cooperation, while also presenting challenges such as modality discrepancies and low-quality modality noise interference. HAMNet [13] and CCNet [14] reduce spectral differences by constraining feature distribution consistency, and propose decision-level feature alignment methods to fuse multiple spectra. PFNet [15] and IEEE [16] design complex multi-spectral feature interaction modules that utilize feature-level fusion methods to mine complementary features between spectra. FACENet [43] introduces glare light priors to restore low-quality modal features degraded by severe glare using other auxiliary spectra. HTT [44] proposes test-time training strategy to improve generalization performance in unseen multi-spectral data. TOP-ReID [17] proposes token permutation and reconstruction modules to align each spectral. EDITOR [18] selects object-centric tokens in the feature and frequency domain to reduce background interference and low-quality noise in multi-spectral fusion. DeMo [19] employs a multi-head attention mixture-of-experts framework to decouple features across spectral modalities. MambaPro [20] leverages adapter and visual prompt to adapt CLIP [33] for ReID tasks, while leveraging the selective state space model in Mamba [21] to aggregate both intra- and inter-modal features. PromptMA [22] proposes a visual prompt-based modality interaction strategy, and the prompt-based token selection and fusion method to aggregate complementary information across modalities and bridge the distribution gap. However, most of the above methods focus on the interaction and fusion of spectral features, ignoring the semantic complementary of object identity in different spectra. We propose instance-level spectral semantic prompt learning to build a semantic bridge between spectra, enabling a more robust capture of object identity features across different spectra.

B. Vision-Language Pre-training Model

The large-scale vision-language foundation models, such as CLIP [33], exhibit strong feature extraction and robust semantic understanding abilities. Thanks to their well-trained image-text aligned latent space, researchers effectively transfer these models to applications in classification, retrieval, and open-set tasks using prompt templates. CoOp [36] first proposes learnable prompts instead of fixed prompt templates to avoid the limitations of manual design. Since static prompts struggle to generalize to unseen categories, CoCoOp [37] proposes to adaptively learn instance level prompts through a lightweight meta network. Gal *et al.* [45] propose a text inversion method in image generation, which maps visual concepts to pseudo-words for personalized text image generation. Pic2Word [46] converts input images into language embeddings for flexible combination of image and text queries. PromptStyler [47] proposes using text prompts to represent different image domains and simulating the distribution changes in the image space via prompts. For ReID task, CLIP-ReID [38] first designs a two-stage prompt learning method to align the identity semantics without concrete textual description. PromptSG [39] inverts each image into personalized pseudo words, exploiting cross-modal text prompts to guide the model focus on semantic-related features. In contrast to the methods mentioned above, our approach focuses on multi-spectral tasks, learning the decoupled semantic prompt, guiding the fusion of spectral features, and end-to-end capturing identity semantics to recognize object identity.

III. Methodology

In this section, we provide a detailed introduction to our method. First, we define the problem and present some notations in Section III-A. Then, in Section III-B, we propose an instance-level spectral inversion method, which decouples the prompt template into spectral-shared style and instance-specific content. Next, Section III-C proposes a semantic-guided multi-spectral fusion module to link identity semantics between spectra. In Section III-D, we propose a semantic refinement and embedding module to enhance spectral representation. Finally, we explain the optimization and inference process in Section III-E.

A. Preliminary

First, we introduce the CLIP-based prompt learning training process and define some notations. As illustrated in Fig. 4, the diagram clarifies the core process of our framework. The visual encoder $\mathcal{I}(\cdot)$ and the text encoder $\mathcal{T}(\cdot)$ in the CLIP model effectively align object images with identity semantics within the cross-modal latent space. Without concrete textual labels, we introduce instance-level inversion text prompts $\mathcal{P}(\cdot)$ to capture the semantic features of spectral.

Specifically, we define each spectral sample as $x = \{x_{rgb}, x_{nir}, x_{tir}\}$, where x_m indicates the three heterogeneous modalities: visible light (RGB), near-infrared (NIR), and thermal infrared (TIR). Leveraging the spectral inversion method, we input the spectral images x into the visual branch $\mathcal{I}(\cdot)$, obtaining the raw spectral feature set $v = \{v_{rgb}, v_{nir}, v_{tir}\}$. To generate the initial rough description, we first project the spectral features v into spectral prompts p, which are then input into the frozen text encoder $\mathcal{T}(\cdot)$, to generate the spectral inversion token t.

Secondly, to capture the complementary identity semantics, we propose the spectral features v to interact with each spectral patch token v_p, to ensure semantics propagation between spectral modalities. Finally, we exploit the semantic-aware structural consistency between spectra to rearrange the spectral patch tokens, and embed the text prompts with spectral features to obtain the final predicted features.

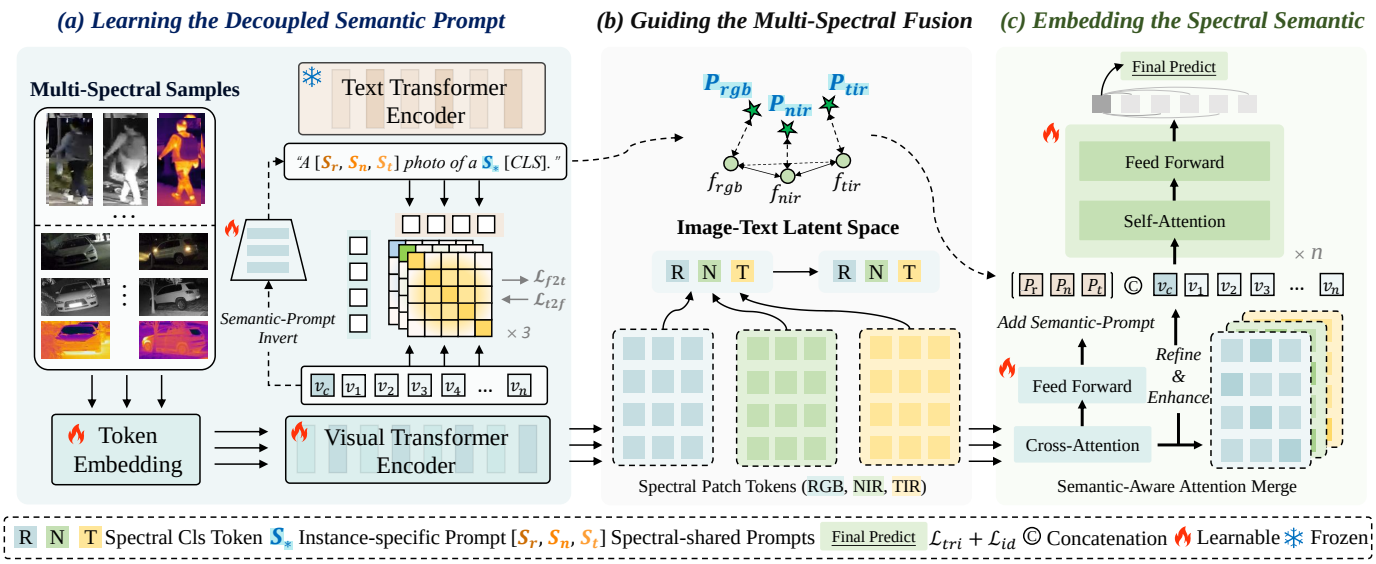


Fig. 4. Overview of our proposed framework. We invert the spectral features v into instance-specific semantic tokens $\boldsymbol{S}_*$, and combine them with the spectral-shared prompts $[\boldsymbol{S}_r, \boldsymbol{S}_n, \boldsymbol{S}_t]$ to form the spectral prompts $[\boldsymbol{P}_r, \boldsymbol{P}_n, \boldsymbol{P}_t]$, which guide the fusion of the spectral features $[\boldsymbol{R}, \boldsymbol{N}, \boldsymbol{T}]$. Finally, the prompts are refined via a unified attention mechanism between spectra, and embedding into the spectral representations to further improve final prediction.

B. Learning the Decoupled Semantic Prompt

Leveraging the pre-trained image-text alignment latent space of CLIP, we can roughly obtain text semantic vectors related to the object identity. However, the absence of real text semantics makes it challenging to extract identity semantics directly from different spectra. To address this issue, we propose using learnable text prompts $\mathcal{P}(\cdot)$ to capture the identity semantics within spectral modalities. Traditional approaches define static ID-specific prompts for each identity, overlooking the diversity of each spectral sample and making it difficult for a single prompt to capture instance-specific semantics. Additionally, the ID-specific semantic prompts are difficult to apply in inference. As shown in Fig. 4 (a), we compose the visual features v of each spectral singleton as part of the spectral prompt and project them to the text semantic space using the frozen text encoder $\mathcal{T}(\cdot)$.

We define the text prompt template "A photo of a $\boldsymbol{S}_*$ [*CLS*]" to generate a pseudo-description for the spectra, where spectral visual features are projected into pseudo-words $p_{content} = \{\boldsymbol{S}_*\}$ through the inversion network ϕ. Simple spectral pseudo-word inversion can provide vague object content semantics in the absence of concrete text. However, instance-level inversion struggles to capture underlying common semantic features within the spectra (*e.g.* temperature, stylistic). These stylistic semantics typically rely on large-scale concrete descriptions or computationally expensive modality prototypes, which cannot be directly applied in text prompt learning. We define learnable text prompts $p_{spectra} = [\boldsymbol{S}_r, \boldsymbol{S}_n, \boldsymbol{S}_t]$ for each spectral modality to obtain generalized spectral style semantics, where $\boldsymbol{S}_r$, $\boldsymbol{S}_n$, and $\boldsymbol{S}_t$ capture the semantics of all samples within RGB, NIR, and TIR modalities, respectively. The final prompt template consists of $p_{spectra}$ and $p_{content}$, forming p = "A $[\boldsymbol{S}_r, \boldsymbol{S}_n, \boldsymbol{S}_t]$ photo of a $\boldsymbol{S}_*$ [*CLS*]." This template is then fed into the text encoder $\mathcal{T}(\cdot)$, generating the rough text vector $p = \{p_{rgb}, p_{nir}, p_{tir}\}$. We

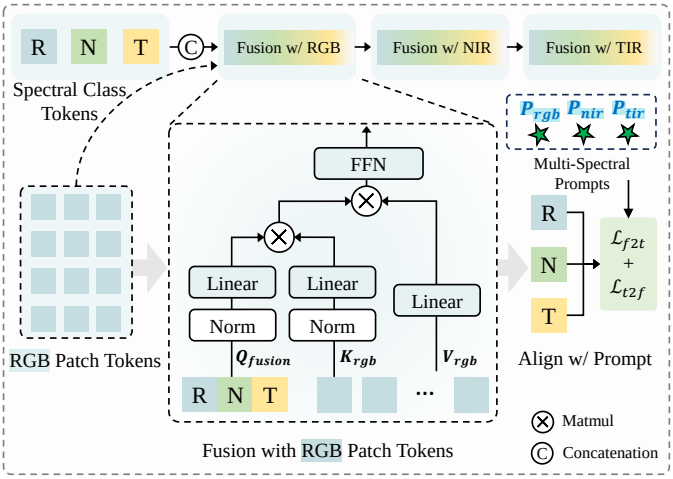


Fig. 5. Details of the semantic-guided spectral fusion module.

use the symmetric contrastive loss to ensure that the spectral prompts align with the spectral modalities in the text-image latent space, as follows:

$$\mathcal{L}_{i2t} = -\frac{1}{M}\log\frac{\exp(\langle v_m^{c,j}, p_m^c\rangle/\gamma)}{\sum_{k=1}^{N_m}\exp(\langle v_m^{c,j}, p_m^k\rangle/\gamma)}, \tag{1}$$

$$\mathcal{L}_{t2i} = -\frac{1}{M}\log\frac{\exp(\langle p_m^c, v_m^{c,j}\rangle/\gamma)}{\sum_{k=1}^{N_m}\exp(\langle p_m^c, v_m^{k,j}\rangle/\gamma)}, \tag{2}$$

where $v_m^{c,j}$ denotes the m-th spectra feature of the j-th sample in the c-th identity, and p_m^c is its positive text prompt, N_m represents the number of identities in the m-th spectra, M is the number of spectra, and γ is a temperature hyper-parameter.

C. Guiding the Multi-Spectral Fusion

Simple spectral inversion offers rough identity semantics within individual spectral modalities. However, to construct a

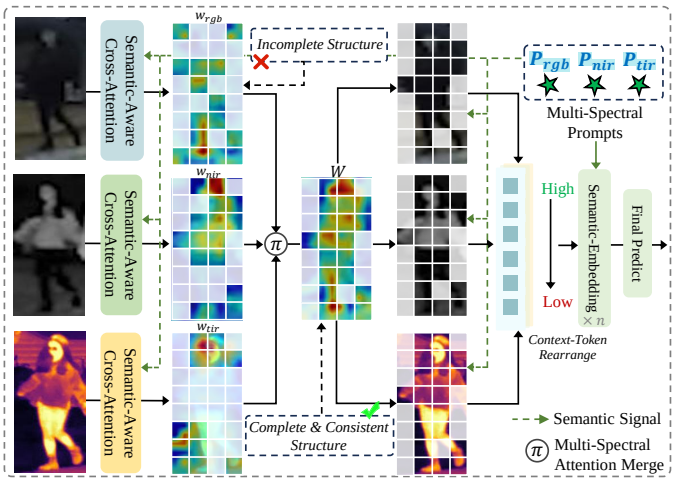


Fig. 6. Architecture of our semantic embedding module constrained by structure consistency.

comprehensive identity semantic set, it is essential to extract complementary features across multiple spectral modalities. To this end, we propose a semantic-guided spectral fusion strategy that serves as a bridge, linking spectral semantic prompts for a comprehensive representation. As illustrated in Fig. 5, the process and details of prompt guidance and spectral fusion are comprehensively presented.

To enable effective spectral semantic alignment, we aggregate the global identity representations from all modalities into a raw spectral identity feature f. Meanwhile, patch-level features are extracted from each spectrum, forming a multi-spectral context $v_p = \{v_{p,rgb}, v_{p,nir}, v_{p,tir}\}$, where each $v_{p,m}$ contains fine-grained embeddings for modality m. The cross-attention mechanism is employed to guide the identity representation f to focus on spectrum-specific semantics in a progressive manner. Formally, the raw identity feature f serves as query Q, and the multi-spectral context v_p as the key K and value V. The specific formula is as follows:

$$f_1 = \text{Concat}\left(v_{rgb}, v_{nir}, v_{tir}\right), \tag{3}$$

$$f_m = \text{FFN}\left(\text{LN}\left(\text{CA}\left(f_{m-1}, v_{p,m}\right)\right)\right), \tag{4}$$

where f_m denotes the feature output of the identity semantic set in the m-th spectra, $v_{p,m}$ is the m-th spectra in the spectral context, CA represents the Cross-Attention mechanism [48], LN is the LayerNorm [49], and FFN refers to the Feed-Forward Network [48]. This allows the model to selectively focus on the complementary semantics across different spectra.

As the spectra fusion, the identity semantics in f_m become comprehensive gradually, and the final fused feature is $\hat{f}$. Finally, we replace the spectral feature v with $\hat{f}$ in Eq. (1) and (2), and align the fusion identity feature with the spectral prompt p to produce the comprehensive semantic prompt set, as follows:

$$\mathcal{L}_{SupCon} = \mathcal{L}_{f2t} + \mathcal{L}_{t2f}, \tag{5}$$

$$\mathcal{L}_{f2t} = -\frac{1}{M}\log\frac{\exp(\langle \hat{f}_m^{c,j}, p_m^c\rangle/\gamma)}{\sum_{k=1}^{N_m}\exp(\langle \hat{f}_m^{c,j}, p_m^k\rangle/\gamma)}, \tag{6}$$

$$\mathcal{L}_{t2f} = -\frac{1}{M}\log\frac{\exp(\langle p_m^c, \hat{f}_m^{c,j}\rangle/\gamma)}{\sum_{k=1}^{N_m}\exp(\langle p_m^c, \hat{f}_m^{k,j}\rangle/\gamma)}, \tag{7}$$

where $\hat{f}_m^{c,j}$ denotes the m-th spectral fusion feature of the j-th sample in the c-th identity.

D. Embedding the Spectral Semantic

Although the fusion mechanism transfers modal knowledge across different spectra, localized low-quality noise within a spectrum disrupts the intrinsic integrity of the object structure. As shown in Fig. 6, the modal semantic region under different spectra exhibits significant structural differences.

To thoroughly capture the fine-grained local details of the identity within each spectrum, we propose a semantic embedding module constrained by structure consistency. Specifically, we leverage semantic-aware attention to represent the identity structure and merge fine-grained structure information across different spectra. To maintain generality, we present the TIR modality as an example.

First, we employ a cross-attention to combine the semantic prompt p_{tir} with spectrum $v_{p,tir}$, extracting fine-grained semantic attention within the spectrum as w_{tir},

$$w_{tir} = \text{softmax}(p_{tir} \cdot (v_{p,tir})^\top/\tau), \tag{8}$$

where τ is the temperature hyper-parameter. The semantic attentions across different spectrum are denoted as $w = \{w_{rgb}, w_{nir}, w_{tir}\}$.

We unify the semantic intensity of different spectra within each local feature and fuse it as the structural information of the foreground identity and background noise to obtain complete structure attention W. The formula is as follows:

$$W = \sum_m^M \frac{\exp(w_m/\tau)}{\sum_n^M \exp(w_n/\tau)} \cdot w_m, m, n \in \{\text{rgb.nir}, \text{tir}\} \tag{9}$$

Based on the complete spectral structural attention, we realign patches from different spectral through a rearrange strategy to obtain structurally consistent spectral features, as shown below:

$$\begin{aligned} \hat{v}_{p,tir} &= \left[v_{p,tir}^{\pi(1)}, v_{p,tir}^{\pi(2)}, \ldots, v_{p,tir}^{\pi(N)}\right]^\top, \\ \pi &= \underset{j\in\{1,\ldots,N\}}{\text{argsort}}\ (W_j), \end{aligned} \tag{10}$$

where π sorts each patch based on attention weight W and reassigns their index positions $\pi(N)$ from highest to lowest, ensuring the structural alignment of foreground identity within each spectral. This leads to structurally consistent image features, as illustrated in Fig. 6. Trivially, we do not discard low-weight background tokens but retain them to maintain complete contextual information.

For semantic prompts, we refine the fine-grained information within the semantic feature p based on the shared complete structural attention. The specific formulation is as follows:

$$\hat{p}_{tir} = W \cdot v_{p,tir} + p_{tir}. \tag{11}$$

TABLE I
COMPARISON PERFORMANCES WITH THE STATE-OF-THE-ART METHODS ON RGBNT201 [15]. THE BEST AND SECOND BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE, RESPECTIVELY. THE SUPERSCRIPT † REPRESENTS CLIP-BASED METHODS, ∗ DENOTES VIT-BASED METHODS AND OTHERS ARE CNN-BASED METHODS.

	Methods	Venue	RGBNT201			
			mAP	R-1	R-5	R-10
Single	MUDeep [51]	ICCV17	23.8	19.7	33.1	44.3
	MLFN [52]	CVPR18	24.7	23.7	38.5	49.5
	PCB [3]	ECCV18	32.8	28.1	37.4	46.9
	HACNN [53]	CVPR18	19.3	14.7	25.5	32.8
	OSNet [11]	ICCV19	22.1	22.9	37.2	45.9
	CAL [54]	ICCV21	27.6	24.3	36.5	45.7
	TransReID* [8]	ICCV21	65.2	66.9	78.5	84.7
	CLIP-ReID† [38]	AAAI23	71.1	71.8	80.3	85.6
	PromptSG† [39]	CVPR24	73.9	75.6	82.7	88.2
Multi	HAMNet [13]	AAAI20	27.7	26.3	41.5	51.7
	PFNet [15]	AAAI21	38.5	38.9	52.0	58.4
	IEEE [16]	AAAI22	46.4	47.1	58.5	64.2
	UniCat* [55]	NIPSW23	57.0	55.7	-	-
	HTT* [44]	AAAI24	71.1	73.4	83.1	87.3
	TOP-ReID* [17]	AAAI24	72.3	76.6	84.7	89.4
	EDITOR* [18]	CVPR24	66.5	68.3	81.1	88.2
	RSCNet* [42]	TCSVT24	68.2	72.5	-	-
	PromptMA† [22]	TIP25	78.4	80.9	87.0	88.9
	DeMo† [19]	AAAI25	79.0	82.3	88.8	**92.0**
	MambaPro† [20]	AAAI25	78.9	83.4	**89.8**	91.9
	DEEP†	Ours	**79.6**	**84.2**	89.4	91.5

TABLE II
COMPARISON PERFORMANCES WITH THE STATE-OF-THE-ART METHODS ON MARKET-MM [16].

	Methods	Venue	Market-MM			
			mAP	R-1	R-5	R-10
Single	MLFN [52]	CVPR18	42.7	68.1	87.1	92.0
	HACNN [53]	CVPR18	42.9	69.1	86.6	92.2
	OSNet [11]	ICCV19	39.7	69.3	86.7	91.3
	TransReID* [8]	ICCV21	73.0	88.9	95.8	97.6
	CLIP-ReID† [38]	AAAI23	82.5	93.7	97.9	98.8
	PromptSG† [39]	CVPR24	84.5	94.2	**97.9**	**98.9**
Multi	HAMNet [13]	AAAI20	60.0	82.8	92.5	95.0
	PFNet [15]	AAAI21	60.9	83.6	92.8	95.5
	IEEE [16]	AAAI22	64.3	83.9	93.0	95.7
	HTT* [44]	AAAI24	67.2	81.5	95.8	97.8
	TOP-ReID* [17]	AAAI24	82.0	92.4	97.6	98.6
	EDITOR* [18]	CVPR24	77.4	90.8	96.8	98.3
	PromptMA† [22]	TIP25	83.6	93.3	-	-
	DeMo† [19]	AAAI25	83.6	93.1	97.5	98.7
	MambaPro† [20]	AAAI25	84.1	92.8	97.7	98.7
	DEEP†	Ours	**85.8**	**94.7**	97.8	98.7

TABLE III
STATISTICAL ANALYSIS OF DATASETS USED IN OUR EXPERIMENTS.

	Datasets	# Samples	# IDs	# Cams
Person	RGBNT201 [15]	4,787	201	4
	Market-MM [16]	32,668	1,501	6
Vehicle	MSVR310 [14]	2,087	310	8
	WMVEID863 [43]	4,709	863	8
	RGBNT100 [13]	17,250	100	8

Finally, we concatenate the optimized spectral identity features $\hat{f}_{tir}$, spectral semantic features $\hat{p}_{tir}$, and spectral patches $\hat{v}_{p,tir}$, then encode them using a self-attention mechanism,

$$\tilde{f}_{tir} = \text{FFN}(\text{LN}(\text{MSA}(\text{Concat}(\hat{f}_{tir}, \hat{p}_{tir}, \hat{v}_{p,tir})))), \quad (12)$$

where MSA is the Multi-head Self-Attention [48] mechanism. We use the representations of each spectrum $\tilde{f}$=$\{\tilde{f}_{rgb}, \tilde{f}_{nir}, \tilde{f}_{tir}\}$ for the final prediction.

E. Optimization and Inference

During the training phase, we continue to use standard identity classification loss $\mathcal{L}_{id}$ [10] and triplet loss $\mathcal{L}_{tri}$ [50] for optimization.

$$\mathcal{L}_{id} = \sum_{i=1}^{N} -q_i \log(p_i), q_i = \begin{cases} 0, y \neq i \\ 1, y = i, \end{cases} \quad (13)$$

$$\mathcal{L}_{tri} = \text{Max}(d_p - d_n + \delta, 0), \quad (14)$$

where p_i is the prediction probability for the i-th class, δ is the margin of triplet loss.

The objective function used in our framework is defined as follows:

$$\mathcal{L}_{final} = \mathcal{L}_{id} + \mathcal{L}_{tri} + \lambda \mathcal{L}_{SupCon}, \quad (15)$$

where λ is a hyper-parameter set to a fixed value.

At the testing phase, the text prompts generated from spectral inversion are utilized as part of the prediction features to directly describe the identity semantics.

IV. EXPERIMENT

In this section, we conduct detailed experiments on the proposed framework. First, we introduce the datasets, evaluation protocol, and implementation details of the proposed framework (Sec.IV.A-B). Second, we conduct comparative experiments with state-of-the-art methods on person and vehicle datasets (Sec.IV.C). Third, we perform ablation studies and thoroughly analyze the effectiveness of each component (Sec.IV.D). Then, we analyze computational efficiency for the overall framework (Sec.IV.E). Finally, we provide further analysis for missing-spectral scenarios and extensive visualization experiments to offer deeper insight into our framework (Sec.IV.F).

A. Datasets and Evaluation Protocols

Datasets. We conduct experiments on five public multi-spectral datasets, including two person datasets and three vehicle datasets. RGBNT201 [15] is a real-world person dataset covering four non-overlapping camera views, and each sample consists of three modalities: visible light (RGB), near-infrared (NIR), and thermal-infrared (TIR), while Market-MM [16] is a synthetic pedestrian dataset generated with CycleGAN [56] from a single-modal dataset [57]. For vehicle ReID, RGBNT100 [13] captures a large number of tri-spectral vehicle pairs, MSVR310 [14] collects eight non-repetitive views of each vehicle over a long timespan, and

TABLE IV
COMPARISON PERFORMANCES WITH THE STATE-OF-THE-ART METHODS ON MSVR310 [14] AND RGBNT100 [13].

	Methods	Venue	MSVR310		RGBNT100	
			mAP	R-1	mAP	R-1
Single	PCB [3]	ECCV18	23.2	42.9	57.2	83.5
	BoT [10]	CVPRW19	23.5	38.4	78.0	95.1
	OSNet [11]	ICCV19	28.7	44.8	75.0	95.6
	AGW [60]	TPAMI21	28.9	46.9	73.1	92.7
	PFD [30]	AAAI22	23.0	39.9	67.5	92.6
	FED [61]	CVPR22	21.7	37.4	65.8	91.7
	TransReID* [8]	ICCV21	33.4	48.9	75.6	92.9
	CLIP-ReID† [38]	AAAI23	52.6	71.1	87.0	96.9
	PromptSG† [39]	CVPR24	60.2	77.2	87.3	96.8
Multi	HAMNet [13]	AAAI20	27.1	42.3	74.5	93.3
	PFNet [15]	AAAI21	23.5	37.4	68.1	94.1
	IEEE [16]	AAAI22	21.0	41.0	61.3	87.8
	CCNet [14]	INFFUS23	36.4	55.2	77.2	96.3
	TOP-ReID* [17]	AAAI24	35.9	44.6	81.2	96.4
	FACENet* [43]	INFFUS25	36.2	54.1	81.5	96.9
	EDITOR* [18]	CVPR24	39.0	49.3	82.1	96.4
	RSCNet* [42]	TCSVT24	39.5	49.6	82.3	96.6
	DeMo† [19]	AAAI25	49.2	59.8	86.2	97.6
	MambaPro† [20]	AAAI25	47.0	56.5	83.9	94.7
	PromptMA† [22]	TIP25	55.2	64.5	85.3	97.4
	DEEP†	Ours	**66.0**	**82.1**	**88.5**	**97.6**

TABLE V
COMPARISON PERFORMANCES WITH THE STATE-OF-THE-ART METHODS ON WMVEID863 [43].

	Methods	Venue	WMVEID863			
			mAP	R-1	R-5	R-10
Single	DenseNet [62]	CVPR17	42.9	47.9	61.9	68.7
	HACNN [53]	CVPR18	46.9	48.9	66.9	73.8
	BoT [10]	CVPR19	51.1	55.7	69.8	74.7
	OSNet [11]	ICCV19	42.9	46.8	61.9	69.4
	AGW [60]	TPAMI21	30.3	35.3	43.3	46.5
	PFD [30]	AAAI22	50.2	55.3	69.8	75.3
	TransReID* [8]	ICCV21	67.0	74.7	79.5	82.4
	CLIP-ReID† [38]	AAAI23	66.3	73.5	79.5	84.2
	PromptSG† [39]	CVPR24	71.4	78.6	85.4	86.8
Multi	HAMNet [13]	AAAI20	45.6	48.5	63.1	68.8
	PFNet [15]	AAAI21	50.1	55.9	68.7	75.1
	IEEE [16]	AAAI22	45.9	48.6	64.3	67.9
	CCNet [14]	INFFUS23	50.3	52.7	69.6	75.1
	TOP-ReID* [17]	AAAI24	67.7	75.3	80.8	83.5
	EDITOR* [18]	CVPR24	65.6	73.8	80.0	82.3
	FACENet* [43]	INFFUS25	69.8	77.0	81.0	84.2
	DeMo† [19]	AAAI25	68.8	77.2	81.5	83.8
	MambaPro† [20]	AAAI25	69.5	76.9	80.6	83.8
	DEEP†	Ours	**73.7**	**81.0**	**86.3**	**87.4**

WMVEID863 [43] is the largest multi-spectral vehicle dataset, designed to address the challenge of severe light glare pollution. As shown in Table III, we provide a statistical analysis of the number of unique samples, identities, and cameras in each dataset.

Evaluation Protocols. In line with the convention of the ReID community [8], [58], we use the Cumulative Matching Characteristics (CMC) and the Mean Average Precision (mAP) as evaluation metrics. As in previous works [15], [43], we adopt the common evaluation protocol for RGBNT201 [15], Market-MM [16], WMVEID863 [43] and RGBNT100 [13]. For MSVR310 [14], we enforce a strict protocol [14] that filters out samples with the same identity and time span based on time labels to avoid easy matching.

B. Implementation Details

We resize each spectral image to 256×128 (128×256 for vehicle images to maintain the aspect ratio). During training, we use data augmentation strategies such as random horizontal flipping, padding with 10 pixels, random erasing [59], and random cropping. We select the ViT-B/16 vision encoder of CLIP as the backbone, freezing all parameters of the text branch for text semantic projection, while keeping the spectral-shared prompts $[\boldsymbol{S}_r, \boldsymbol{S}_n, \boldsymbol{S}_t]$ and the visual branch fully trainable. We employ the Adam optimizer with a learning rate of $5e$-6, a weight decay of 0.0001, and a momentum set to 0.9. Training lasts for 60 epochs, with a linear decay of the learning rate by 0.1 at the 20-th and 40-th epochs, reducing it to $5e$-7 and $5e$-8. All experiments are conducted on one NVIDIA RTX 3090 Ti GPU using the PyTorch framework.

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

As shown in Tables I, II, IV, V, our method not only surpasses state-of-the-art methods [8], [17], [18], [43] but also shows a clear performance improvement compared to existing CLIP-based methods [19], [20], [38], [39]. For a fair comparison, we extend CLIP-ReID [38] and PromptSG [39] into a three-branch network to comply with multi-spectral data.

Comparison on RGBNT201 and Market-MM. As shown in Table I, traditional CNN- and ViT-based methods struggle to achieve superior performance when addressing the real-world challenges of multi-spectral data in RGBNT201 [15]. CLIP-based methods perform well on this dataset due to their strong representation capabilities. However, the lack of multi-spectral information interaction limits the performance of methods like CLIP-ReID [38] and PromptSG [39] on multi-spectral datasets. DeMo [19] leverages the mixture-of-experts learning to decouple modality-specific and shared features, achieving impressive performance. MambaPro [20] utilizes state space in mamba modeling for intra-modal feature extraction and cross-modal feature interaction for the multi-spectral ReID task. In contrast, these approaches fail to consider the identity semantics of the object across spectra and lack identity semantic exploration. Our DEEP outperforms CLIP-ReID [38] by **8.5%/12.4%** in mAP/Rank-1 and surpasses DeMo [19] by **0.6%/1.9%**, on the RGBNT201 [15] dataset. On the larger-scale synthetic dataset Market-MM [16], as shown in Table II, DEEP demonstrates a significant advantage in semantic awareness. Compared with DeMo [19] and MambaPro [20], which are also based on the CLIP [33] backbone, they fail to maintain their leading performance. DEEP further improves mAP/Rank-1 by **2.2%/1.6%** over DeMo [19] and by **1.3%/0.5%** over PromptSG [39]. These experiments evidence our proposed DEEP employs a simple yet effective semantic prompt learning method to effectively guide the model in cap-

TABLE VI
ABLATION OF DIFFERENT COMPONENTS ON RGBNT201 [15], MSVR310 [14], AND WMVEID863 [43]. WE USE A TRIPLET-STEAM CLIP VISUAL ENCODER AS THE BASELINE.

	Components			RGBNT201				MSVR310				WMVEID863			
	DSP	SGSF	SSE	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
(a)	×	×	×	71.0	71.5	81.3	86.4	49.1	65.5	82.7	85.6	66.4	72.4	79.4	82.9
(b)	✓	×	×	73.3	76.3	85.3	90.2	55.2	74.5	86.1	89.3	67.1	74.0	81.7	85.4
(c)	✓	×	✓	73.2	76.6	87.3	91.7	63.0	78.5	89.0	92.6	72.8	80.4	85.9	86.3
(d)	×	✓	×	75.0	75.4	85.9	88.5	51.8	69.2	84.4	89.2	62.2	69.2	81.7	85.4
(e)	✓	✓	×	78.7	82.1	88.3	**91.9**	55.4	75.0	84.6	89.0	67.5	75.6	83.3	86.3
(f)	✓	✓	✓	**79.6**	**84.2**	**89.4**	91.5	**66.0**	**82.1**	**90.7**	**93.9**	**73.7**	**81.0**	**86.3**	**87.4**

TABLE VII
EFFECTS OF THE DECOUPLED SEMANTIC PROMPT LEARNING. "LER." DENOTES LEARNABLE PROMPT, AND "INV." DENOTES INVERTED PROMPT.

	Prompt Type		RGBNT201		MSVR310	
	$p_{spectra}$	$p_{content}$	mAP	R-1	mAP	R-1
(a)	×	×	77.8	78.9	63.2	79.4
(b)	✓	×	79.1	80.9	64.7	81.9
(c)	×	✓	78.8	82.4	65.3	81.4
(d)	✓	✓	**79.6**	**84.2**	**66.0**	**82.1**
(e)	Inv.	Inv.	78.0	80.4	61.7	78.8
(f)	Ler.	Ler.	76.1	80.6	61.4	80.0
(g)	Inv.	Ler.	78.0	84.7	64.5	81.0
(h)	**Ler.**	**Inv.**	**79.6**	**84.2**	**66.0**	**82.1**

TABLE VIII
EFFECTS OF THE SEMANTIC-GUIDED SPECTRAL FUSION.

Methods	RGBNT201		MSVR310	
	mAP	R-1	mAP	R-1
(a) No Align	75.0	78.3	62.1	78.7
(b) Align-Spec	77.0	81.1	62.6	79.4
(c) Align-Fusion	**79.6**	**84.2**	**66.0**	**82.1**

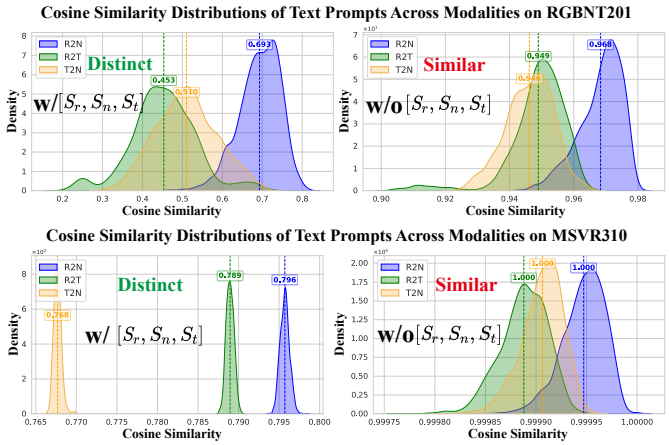


Fig. 7. Cosine similarity distributions of the decoupled semantic prompt on RGBNT201 [15] and MSVR310 [14].

turing instance-level semantic features across multi-spectral data, obtaining the improvement of person ReID performance.

Comparison on MSVR310 and RGBNT100. Table IV shows that our method outperforms other existing methods on the MSVR310 [14] and RGBNT100 [13] datasets. The more significant improvement on MSVR310 [14] can be attributed to that most existing methods overlook the semantic consistency of vehicles across different illuminations and viewpoints. Semantic learning-based methods, such as ours, leverage semantic prompts to guide the model to focus on vehicle foreground. This enables methods like CLIP-ReID [38] and PromptSG [39] to achieve notable gains. Moreover, DEEP benefits from semantic-guided spectral fusion, effectively integrating illumination-invariant features from the TIR and NIR spectra. This strengthens retrieval performance and allows the model to address the key challenges of MSVR310 [14], including large viewpoint variations, temporal changes, and background clutter. For the large-scale RGBNT100 [13] dataset, which contains a large number of duplicate vehicle same viewpoint samples, most methods achieve strong performance, DEEP still demonstrates a **1.2%/0.8%** improvement over PromptSG [39] as shown in Table IV. These performance improvements verify our DEEP leverages the learning and embedding of multi-spectral semantics to boost the model for better mining vehicle appearance knowledge from various viewpoints while effectively integrating vehicle features across spectra, leading to superior performance.

Comparison on WMVEID863. As shown in Table V, when dealing with challenges in WMVEID863 [43], the low-quality spectral and severe light glare pollution, most existing methods struggle to extract robust features. For CLIP-ReID [38], the two-stage prompt learning strategy limits its performance, making it highly susceptible to severe light glare pollution. Although DeMo [19] and MambaPro [20] have to some extent leveraged the capabilities of CLIP [33], they do not exhibit significant performance advantages over FACENet [43], which employs ViT and the glare light priors. Meanwhile, the instance-level semantic priors introduced by PromptSG [39] are notably more effective than handcrafted priors. Compared to these methods, DEEP effectively decouples style semantics across different spectra, mitigating the impact of strong illumination on feature representation. Its prompt-guided modality fusion leverages semantic awareness to preserve identity consistency across spectra. As a result, DEEP achieves further improvements over PromptSG [39], with gains of **2.3%/2.4%** in mAP/Rank-1, and outperforms

TABLE IX
EFFECTS OF THE MODAL COMBINATIONS IN THE SPECTRAL FUSION.

Modal Combinations	Modals			RGBNT201		MSVR310	
	RGB	NIR	TIR	mAP	R-1	mAP	R-1
(a) Only RGB	✓	×	×	69.5	70.6	61.8	78.3
(b) Only NIR	×	✓	×	65.5	64.7	59.3	75.5
(c) Only TIR	×	×	✓	73.9	76.1	58.5	74.6
(d) RGB + NIR	✓	✓	×	69.8	69.7	62.7	79.0
(e) RGB + TIR	✓	×	✓	77.1	79.9	62.2	78.8
(f) NIR + TIR	×	✓	✓	71.7	72.5	60.0	76.5
(g) All Modals	✓	✓	✓	**79.6**	**84.2**	**66.0**	**82.1**

TABLE X
EFFICIENCY COMPARISON OF DIFFERENT SPECTRAL FUSION STRATEGIES.

Methods	RGBNT201		MSVR310		Params	FLOPs
	mAP	R-1	mAP	R-1	M	G
(a) No Fusion	73.2	76.6	63.0	78.5	96.4	45.9
(b) Linear Weight	77.3	80.0	64.4	80.0	98.8	45.9
(c) Conv 1×1	77.5	82.1	64.2	80.0	97.2	46.1
(d) Conv 3×3	78.0	81.3	62.7	78.7	103.5	46.8
(e) **SGSF (Ours)**	**79.6**	**84.2**	**66.0**	**82.1**	105.9	46.2

DeMo [19] by **4.9%/3.8%**.

D. Ablation Study

We leverage the pre-trained ViT-B/16 visual encoder from CLIP as the backbone and extend it into a three-branch network as our baseline for analysis on the RGBNT201 [15] person dataset and the MSVR310 [14], WMVEID863 [43] vehicle datasets. Our proposed framework is divided into three main components: the Decoupled Semantic Prompt (DSP), the Semantic-Guided Spectral Fusion (SGSF), and the Spectral Semantic Embedding (SSE).

Ablation Study of Individual Components. The component analysis of each module is shown in Table VI. Compared with Table VI (a), adding the DSP module in Table VI (b) consistently improves performance across all datasets. The improvement is particularly significant on MSVR310 [14], as the dataset includes high-resolution images with complex backgrounds and significant discrepancies in viewpoint. DSP helps the model focus on semantically relevant vehicle regions and suppresses background noise. In comparison to Table VI (c), the semantic embedding module notably enhances performance on the vehicle dataset, emphasizing the vital role of semantic embedding constrained by structure consistency. As SGSF relies on semantic prompts learned from DSP, we simplify it by removing text-image contrastive learning. As shown in Table VI (d), the simplified SGSF can still effectively mine complementary information between spectra on RGBNT201 [15] and MSVR310 [14]. Performance improves significantly on pedestrian datasets because their noise mainly stems from foreground objects and spectral variations. The SGSF module effectively captures complementary features across modalities. In contrast, improvements on vehicle datasets are more limited due to complex background information in MSVR310 [14] and strong light interference in WMVEID863 [43], which introduces noise during fusion and hinders model performance. In Table VI (e), the combination of DSP and SGSF significantly improves performance across all datasets. SGSF helps DSP capture complementary semantics across spectra, while DSP mitigates the impact of low-quality noise (*e.g.*, background clutter and flare light pollution) for SGSF fusion. Finally, comparing Table VI (e) and (f), the fused spectral modalities are able to more effectively embed spectral semantics. Based on the experiments above, our method significantly improves the mAP/Rank-1 by **8.6%/12.7%**, **16.9%/16.6%**, and **7.3%/8.6%** on the three datasets, respectively, compared with the baseline. These findings highlight the crucial role our approach plays in capturing spectral instance semantics, enabling spectral collaborative fusion, and embedding semantic instances.

TABLE XI
EFFICIENCY COMPARISON OF DIFFERENT SPECTRAL SEMANTIC EMBEDDING STRATEGIES. ♯ MEANS HIGH-RESOLUTION INPUTS OF 384×192 FOR PERSON OR 192×384 FOR VEHICLE.

Methods	RGBNT201		MSVR310		Params	FLOPs
	mAP	R-1	mAP	R-1	M	K
(a) No Refine	77.4	77.2	65.4	81.0	105.9	0
(b) No Rearrange	78.3	79.7	63.0	80.9	105.9	15.5
(c) Mean Pooling	77.0	78.1	64.8	80.9	105.9	213.6
(d) Mean Pooling$^\sharp$	77.9	81.0	65.4	81.9	105.9	478.6
(e) SSE	79.6	84.2	66.0	82.1	105.9	413.1
(f) SSE$^\sharp$	80.4	84.4	66.6	83.4	106.0	925.4

Effects of the Decoupled Semantic Prompt learning. We decompose the spectral prompt template to analyze the impact of different prompt components on model accuracy. We conduct a comprehensive ablation study in Table VII to validate the effectiveness of the proposed DSP strategy. The analysis is performed from two perspectives: (1) the effectiveness of different prompt components, and (2) the effectiveness of different prompt types. For the first group in Table VII (a)–(d), we analyze the contribution of spectral-shared prompts $p_{spectra}$ and instance-specific content prompts $p_{content}$. In Table VII (a), we apply a fixed prompt template, "a photo of a person/vehicle," for each sample. Unlike the learnable prompt, the fixed template struggles to adapt to the task-specific semantics required in downstream tasks, resulting in a performance decline. Table VII (b) demonstrates that a shared learnable spectral prompt improves **1.3%/2.0%** and **1.5%/2.5%** in mAP/Rank-1, it suggests that shared learnable prompts function as unified semantic templates during training, supporting the semantic transfer of spectral across the entire dataset. In Table VII (c), instance-level prompts improve performance by **1.0%/3.5%** and **2.1%/2.0%** in mAP/Rank-1, suggesting that instance-specific inversion enables the model to capture unique identity semantics for each sample. Finally, combining spectral-shared prompts and instance-specific prompts more effectively captures spectral identity features, resulting in mAP/Rank-1 gains of **1.8%/5.3%** and **2.8%/2.7%**, significantly enhancing the ability of DEEP to capture spectral identity features.

For the second group in Table VII (e)–(h), we investigate

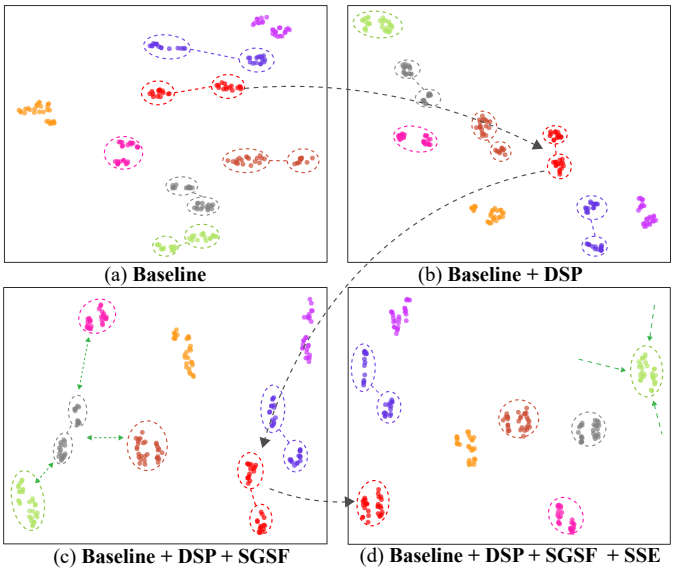


Fig. 8. T-SNE [63] visualization of the feature distribution (a) Baseline, (b) Baseline + DSP, (c) Baseline + DSP + SGSF, and (d) Baseline + DSP + SGSF + SSE on RGBNT201 [15].

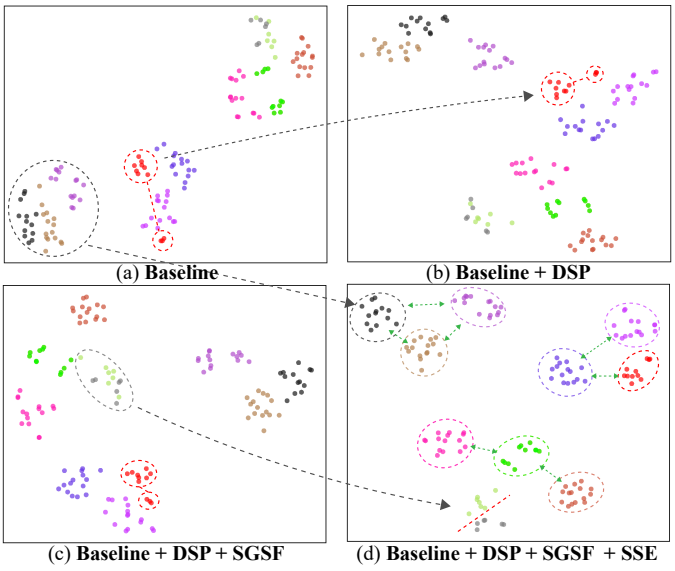


Fig. 9. T-SNE [63] visualization of the feature distribution (a) Baseline, (b) Baseline + DSP, (c) Baseline + DSP + SGSF, and (d) Baseline + DSP + SGSF + SSE on MSVR310 [14].

the impact of learnable prompt type and inverted prompt type. In Table VII (e) and (f), the fully inverted prompts and fully learnable prompts result in degraded performance across both datasets, suggesting that these prompt types either lack sufficient semantic specificity or suffer from overfitting. For instance, the Table VII (f) yields the lowest mAP of **76.1%** on RGBNT201 [15] and **61.4%** on MSVR310 [14]. The hybrid strategy Table VII (g), with inverted spectral and learnable content prompts, improves results but remains inferior to Table VII (h). Our final design Table VII (h), which adopts learnable spectral prompts and inverted content prompts, reaches the optimal mAP/Rank-1 scores of **79.6%/84.2%** and **66.0%/82.1%**, respectively, proving the effectiveness of decoupling prompt learning to better generalize across modalities.

Moreover, to further validate the effectiveness of spectral-shared prompts in DSP, we measure the cosine similarity between prompts from different modalities. As illustrated in Fig. 7, the use of spectral-shared prompts significantly reduces semantic similarity across modalities, indicating that the learned prompts effectively capture modality-specific style semantics and achieve decoupled learning.

Effects of the Semantic-Guided Spectral Fusion. To analyze how different guiding features influence the propagation of identity information, we modify the loss $\mathcal{L}_{f2t}$ and the loss $\mathcal{L}_{t2f}$ to compare various prompt alignment strategies. As shown in Table VIII (c), aligning with the fusion features significantly enhances the model performance. In contrast, as shown in Table VIII (b), aligning with the original spectral features results in semantic loss within the spectra, leading to incomplete identity information and a decrease in model performance. Additionally, in Table VIII (a), without alignment with spectral prompts, degrades the mAP/Rank-1 performance by **4.6%/5.9%** and **3.9%/3.4%**. This drop is primarily due to confusing semantic prompts that incorrectly guide the model to overfit features unrelated to the object. Overall, accurate and comprehensive semantic prompts are key for the model to effectively learn identity discriminative features.

Effects of the Modal Combinations in the Spectral Fusion. As shown in Table IX, we conduct an ablation study to investigate the impact of different modal combinations in the fusion module. When using only a single modality in the Table IX (a)-(c), the TIR branch achieves the superior performance on RGBNT201 [15], with **73.9%/76.1%** in mAP/Rank-1. On MSVR310 [14], the RGB branch performs well, reaching **61.8%/78.3%** in mAP/Rank-1. Compared to single modalities, combining RGB with TIR in Table IX (e) brings a notable performance gain, especially on RGBNT201 [15], where mAP improves by **7.6%** over RGB only. This demonstrates the strong complementarity between visible and thermal modalities. In contrast, the Table IX (d) and (f) combining RGB and NIR or NIR and TIR provide less significant improvements. Finally, fusing all three modalities in Table IX (g) consistently achieves the optimal results on both datasets, validating that comprehensive multi-modal information enables more robust and discriminative representation learning.

E. Efficiency Analysis

Efficiency Comparison of the Different Spectral Fusion Strategies. As shown in Table X, we evaluate various spectral fusion strategies on RGBNT201 [15] and MSVR310 [14]. In Table X (b), a simple linear weighting scheme leads to a significant improvement over the no-fusion baseline in Table X (a), demonstrating the inherent complementarity among spectral modalities and the benefit of even basic fusion. In Table X (c), Conv 1×1 provides slightly better performance than linear fusion, indicating its ability to introduce lightweight nonlinearity, though its modeling capacity remains limited. In Table X (d), Conv 3×3 does not achieve further gains and even causes performance degradation on MSVR310 [14], likely due

TABLE XII
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE NUMBER OF LEARNABLE PARAMETERS AND FLOPS ON RGBNT201 [15] AND MSVR310 [14]. ⋆ DENOTES IMPLEMENTATION WITH A SHARED VISUAL ENCODER FOR THREE MODALITIES. ♯ MEANS HIGH-RESOLUTION INPUTS OF 384×192 FOR PERSON OR 192×384 FOR VEHICLE.

	Methods	Venue	RGBNT201					MSVR310				
			Params (M)	FLOPs (G)	FPS	mAP	R-1	Params (M)	FLOPs (G)	FPS	mAP	R-1
Single	CLIP-ReID⋆ [38]	AAAI23	86.4	34.8	214.4	63.9	66.4	86.3	34.3	218.1	53.0	72.8
	CLIP-ReID [38]	AAAI23	259.1	34.3	217.7	71.1	71.8	260.1	34.3	216.5	52.6	71.1
	PromptSG⋆ [39]	CVPR24	94.6	45.9	151.7	66.6	69.0	94.6	45.9	151.7	60.3	75.5
	PromptSG [39]	CVPR24	267.5	45.9	152.4	73.9	75.6	267.5	45.9	151.6	60.2	77.2
Multi	TOP-ReID [17]	AAAI24	324.5	35.5	206.9	72.3	76.6	324.4	35.5	206.6	35.9	44.6
	EDITOR [18]	CVPR24	119.3	40.8	170.4	66.5	68.3	118.9	40.8	168.5	39.0	49.3
	PromptMA [22]	TIP25	107.9	36.2	191.7	78.4	80.9	129.2	37.4	191.5	55.2	64.5
	DeMo [19]	AAAI25	98.8	35.1	207.3	79.0	82.3	98.5	34.2	213.5	49.2	59.8
	DeMo♯ [19]	AAAI25	98.9	76.5	90.4	80.1	82.9	98.6	76.5	92.7	55.2	71.6
	MambaPro [20]	AAAI25	74.8	52.4	134.5	78.9	83.4	74.7	52.4	135.5	47.0	56.5
	MambaPro♯ [20]	AAAI25	74.8	114.6	63.5	79.3	83.0	74.7	114.6	63.9	50.1	63.1
	DEEP	Ours	105.9	46.2	155.6	79.6	84.2	104.1	44.9	157.3	66.0	82.1
	DEEP♯	Ours	106.0	91.8	79.7	80.4	84.4	104.2	90.6	79.1	66.6	83.4

TABLE XIII
PERFORMANCE OF MISSING-SPECTRAL SETTINGS ON MSVR310 [14]. "M (X)" MEANS MISSING THE X SPECTRAL MODALITY. BEST IN BOLD, SECOND BEST UNDERLINED.

Methods	M (RGB)		M (NIR)		M (TIR)		M (RGB+NIR)		M (RGB+TIR)		M (NIR+TIR)		Avg.	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
CCNet [14]	25.5	43.7	27.2	42.3	26.4	41.6	10.8	30.1	20.6	34.2	24.1	37.2	22.4	38.2
TOP-ReID [17]	23.5	41.1	28.6	41.8	31.1	45.9	10.8	23.5	22.3	40.6	26.5	36.5	23.8	38.2
DeMo [19]	36.9	55.3	43.1	56.5	46.1	60.9	10.5	24.2	34.1	53.5	40.8	53.6	35.3	50.7
PromptMA [22]	42.3	59.1	49.5	66.7	51.5	70.2	19.1	34.2	39.8	58.2	48.0	66.2	41.7	59.1
DEEP(Ours)	**53.8**	**72.3**	**59.8**	**75.8**	**62.9**	**78.0**	**26.8**	**44.2**	**49.8**	**66.7**	**57.2**	**74.8**	**51.7**	**68.6**

TABLE XIV
PERFORMANCE OF THE MISSING-SPECTRAL SETTINGS ON RGBNT201 [15]. "M (X)" MEANS MISSING THE X SPECTRAL MODALITY. BEST IN BOLD, SECOND BEST UNDERLINED.

Methods	M (RGB)		M (NIR)		M (TIR)		M (RGB+NIR)		M (RGB+TIR)		M (NIR+TIR)		Avg.	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
PCB [3]	23.6	24.2	24.4	25.1	19.9	14.7	20.6	23.6	11.0	6.8	18.6	14.4	19.7	18.1
TOP-ReID [17]	54.4	57.5	64.3	67.6	51.9	54.5	35.3	35.4	26.2	26.0	34.1	31.7	44.4	45.4
DeMo [19]	63.3	65.3	72.6	75.7	56.2	54.1	45.6	46.5	26.3	24.9	40.3	38.5	50.7	50.8
PromptMA [22]	**67.4**	**68.4**	72.5	**75.7**	58.9	57.3	**51.5**	**53.0**	**33.3**	30.4	43.9	42.2	**54.6**	**54.5**
DEEP(Ours)	65.2	66.0	**73.2**	73.4	**59.4**	**58.6**	45.8	49.6	30.9	**34.1**	**44.1**	**44.7**	53.1	54.4

to excessive local modeling that introduces noise or overfitting. In contrast, our proposed Spectral-Guided Semantic Fusion (SGSF) module in Table X (e) achieves superior performance on both datasets. This result highlights the effectiveness of the SGSF module in capturing complementary semantics across modalities and semantic prompts alignment, while maintaining relatively high computational efficiency.

Efficiency Comparison of the Different Spectral Semantic Embedding Strategies. As shown in Table XI, we compare several methods for achieving structural consistency in semantic embedding. Compared to the No Refine method in Table XI (a), the No Rearrange method in Table XI (b) uses semantic attention fusion to improve model performance on RGBNT201 [15] but causes a significant drop on MSVR310 [14]. This indicates that in vehicle scenarios, relying solely on attention-level fusion does not guarantee structural consistency and may introduce noise that degrades performance. Table XI (c) applies a simple mean pooling strategy to maintain structural consistency, but despite the increased computational cost, the model performance shows no significant improvement. In contrast, the SSE rearrange strategy in Table XI (e) consumes only 1.93 times the computation of mean pooling and significantly improves performance. Moreover, Tables XI (d) and (f) demonstrate that increasing input resolution leads to higher computation but also improved performance. Overall, the rearrange strategy effectively maintains identity structural consistency, shows well generalization ability on diverse datasets, and significantly improves model performance while computational complexity scales linearly with input resolution.

Training and Inference Efficiency Analysis. As shown in Table XII, following the analysis strategy introduced in

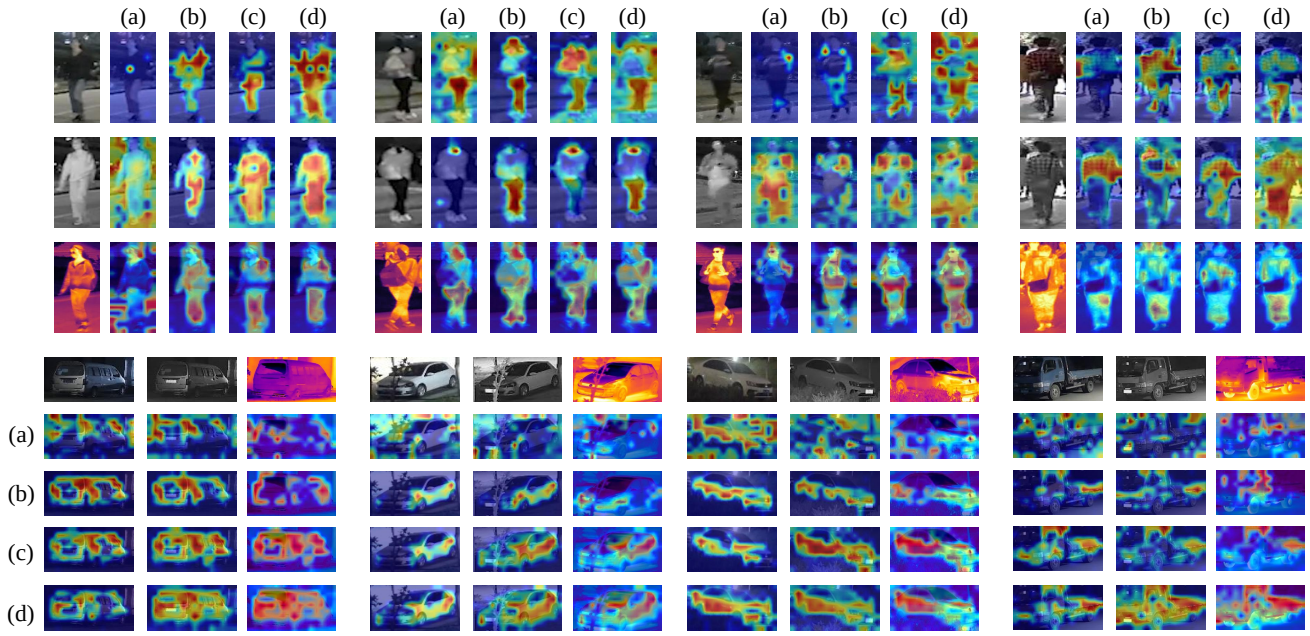


Fig. 10. Visualization results of the (a) Baseline, (b) Baseline + DSP, (c) Baseline + DSP + SGSF, and (d) Baseline + DSP + SGSF + SSE, drawn by Grad-CAM [1], for RGBNT201 [15] and MSVR310 [14].

DeMo [19], we examine the number of learnable parameters, flops of several state-of-the-art methods. As shown in Table XII, DEEP achieves superior performance with a balanced parameter count and computational cost. Although DEEP uses the text encoder of CLIP, it reduces learnable parameters by 2M compared to PromptMA [22] and 13.4M compared to EDITOR [18]. In terms of computational complexity, DEEP is comparable to PromptSG [39] and lower MambaPro [22] by 6.2G FLOPs. For inference time, DEEP is slightly slower than EDITOR [18] by 14.8 FPS but faster than MambaPro [20] by 21.1 FPS. Compared with single-modal methods, we default to replicating the visual encoder three times for three modalities. Additionally, we also evaluate shared-backbone variants of CLIP-ReID [38] and PromptSG [39] with fewer parameters (marked with *), but they exhibit inferior performance on RGBNT201. Note that CLIP-ReID [38] disables the text branch during inference and yields significantly inferior performance compared to DEEP and other state-of-the-art models. Furthermore, when we increase the image resolution to 384×192 (192×384 for vehicle), DEEP$^\sharp$ shows further improvement in performance. The FLOPs only increase by 15.3G compared to DeMo$^\sharp$ [19] and remain lower than MambaPro$^\sharp$ [20] by 22.8G, while the inference speed is just slightly slower than DeMo$^\sharp$ [19] by 10.7 FPS and still higher than MambaPro$^\sharp$ [20] by 16.2 FPS. These results highlight that the performance gain of our method is attributed to architectural advantages rather than the increase in model size.

F. Further Analysis and Visualization

Evaluation of the Missing-Spectral Scenarios. As shown in Table XIII and Table XIV, we investigate the practical challenge of missing spectral on RGBNT201 [15] and MSVR310 [14]. Following the missing spectral setting of DeMo [19], we replace the input spectral with zero vectors to simulate the absence of specific modalities. As shown in Table XIII, benefiting from DEEP advanced semantic learning capability that enables effective focus on semantically relevant regions under significant viewpoint variations and complex background interference, it outperforms PromptMA [22] by **10.0%/9.5%** in mAP/Rank-1 on the MSVR310 [14] dataset. However, as shown in Table XIV, since DEEP is not specifically designed for missing-spectral scenarios, it performs slightly worse than PromptMA [22] on RGBNT201 [15]. The latter is specifically designed for missing spectral, employing a visual prompt-based strategy to address this issue. The results demonstrate that our approach is robust and generalizable in the presence of incomplete spectral data. In future work, we plan to further explore the potential of semantic learning to address real-world missing-spectral problems.

Feature Distribution. To conduct a more generalized analysis, we utilize T-SNE [63] to visualize feature distributions of our proposed modules across both pedestrian and vehicle scenarios. As shown in Fig. 8 (b) and Fig. 9 (b), DSP effectively enhances the model to narrow down different samples within the same identity. After adding the SGSF, as shown in Fig. 8 (c) and Fig. 9 (c), we can observe it improves separability and enlarges the distribution gaps among distinct identities. Finally, the SSE module further enables the model to distinguish challenging samples, as depicted in Fig. 8 (d) and Fig. 9 (d), resulting in a compact intra-class and well-separated inter-class feature distribution.

Discriminative Attention Maps. To further investigate the contribution of each component in DEEP, we employ Grad-CAM [1] for feature visualization. As shown in Fig.10 (a) and (b), without prompt constraints, vanilla CLIP often fails to localize the target object, whereas DSP effectively

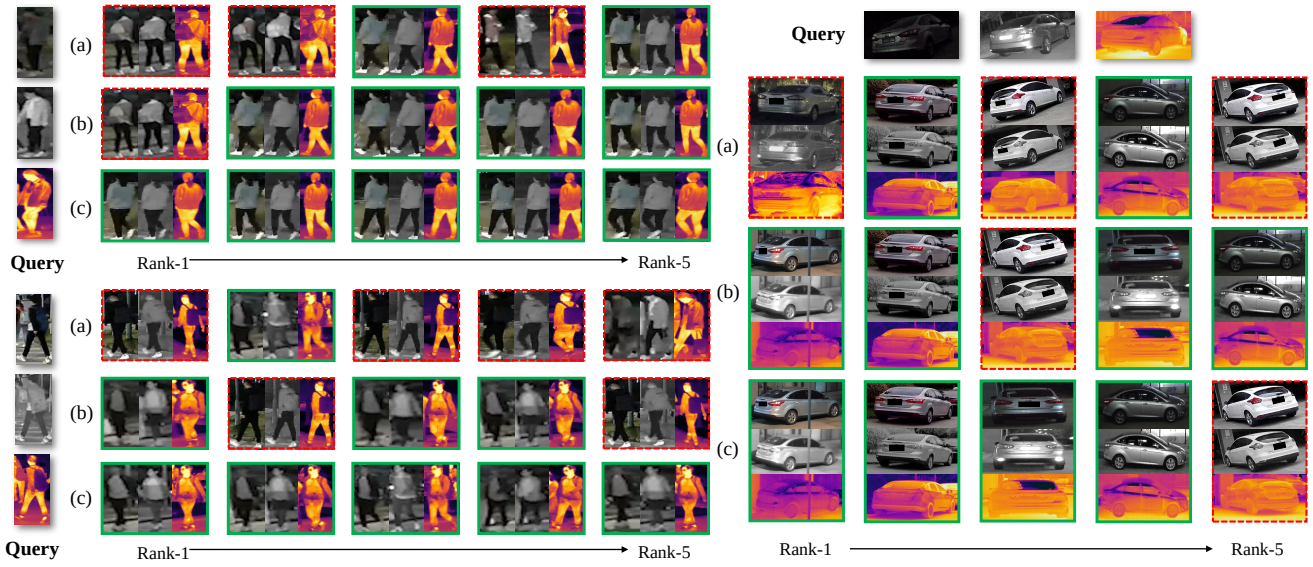


Fig. 11. Visualization of top-5 retrieval results on RGBNT201 [15] and MSVR310 [14], (a) CLIP-ReID [38], (b) PromptSG [39], (c) DEEP (Ours). The query image is randomly selected from the test dataset.

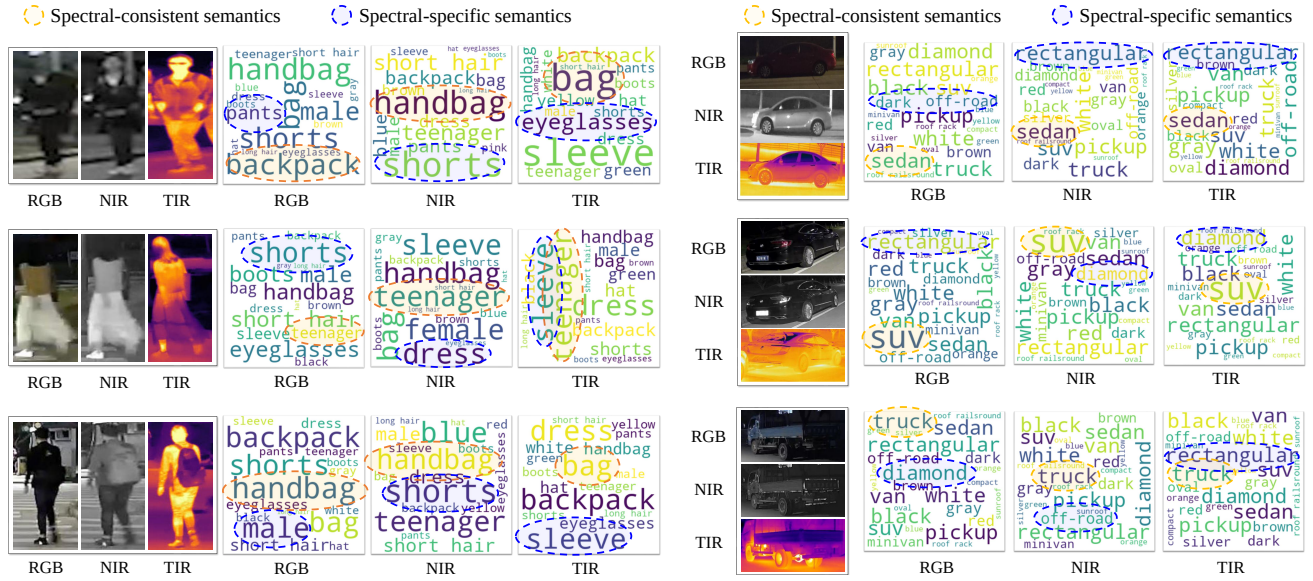


Fig. 12. Word cloud analysis of the spectral-specific prompts on RGBNT201 [15] and MSVR310 [14]. Larger words indicate higher semantic relevance.

directs attention to the object body. Similarly, Fig.10 (b) and (c) demonstrate that SGSF leverages complementary spectral semantics to perceive appearance features from low-quality modalities. In Fig. 10 (d), SSE further refines and embeds semantic features, enabling the model to capture fine-grained details of vehicles and pedestrians while maintaining precise focus on the target. Taking MSVR310 [14] as an example, DEEP accurately isolates the vehicle from background. This robust property is attributed to the alignment of semantic prompts, which suppresses background distractions and enables the extraction of discriminative, illumination-invariant, and viewpoint-invariant features, thereby substantially enhancing the retrieval performance.

Retrieval Results. We randomly select some samples from person and vehicle datasets and analyze their retrieval results. As depicted in Fig. 11 (a), CLIP-ReID [38] is prone to confusion by images with similar camera views, particularly under adverse conditions like low light or blurry noises. However, as shown in Fig. 11 (b), PromptSG [39] identifies more samples, indicating that learning instance-level semantic information significantly improves the model discriminative performance. Finally, as shown in Fig. 11 (c), our method outperforms others by retrieving more gallery samples in the challenging scenarios mentioned above, indicating that fused spectral semantic information significantly enhances the capability of the model to address practical challenges.

Word Cloud Analysis. To better interpret the information embedded in the learnable prompts, we adopt Euclidean distance to retrieve the word vectors that are closest to the object in the attribute vocabulary. To construct the person vocabulary,

we follow previous works [39], [64] by carefully curating a set of person attributes from the Market-1501 [65]. For the vehicle vocabulary, we build it based on captions from a text-based vehicle retrieval dataset [66]. We generate the word cloud visualization to analyze the distance between spectral prompts and attribute labels. As shown in Fig. 12, the three word clouds represent the RGB, NIR, and TIR modalities, respectively. Larger words indicate higher semantic relevance. The words highlighted by the yellow dashed circles illustrate spectral-consistent semantics captured by DEEP, such as 'backpack', 'teenager', 'handbag', and 'sedan.' Conversely, the words marked by the blue dashed circles reveal modality-specific semantics, such as 'eyeglasses', 'dress', 'shorts', and 'rectangular.' The semantic differences across different samples underscore the strong generalization ability of DEEP under diverse scenarios. Although the word cloud further reveals the semantic content learned by DEEP, the semantics obtained via prompt learning inevitably exhibit bias due to the absence of real semantic label guidance. For instance, in nighttime and low-quality modalities, the prompts capture incorrect semantic information. This limitation motivates us to explore multi-modal large language models in the future to enhance multi-spectral semantic learning.

V. Conclusion

In this paper, we propose a novel multi-spectral prompt learning framework, DEEP, which leverages the powerful visual-language foundational model CLIP for the multi-spectral object ReID task. We first propose a decoupled semantic prompt learning method, which decomposes the prompt template into spectral-shared prompts and instance-specific inversion tokens. This enables DEEP to effectively capture object semantics within different spectra. Second, we propose a semantic-guided spectral fusion module that builds a semantic bridge between spectra. By mining the complementary semantics across multiple spectra, this method focuses on the semantic information of the object foreground for multi-spectral fusion. Finally, we propose a spectral semantic embedding module that refines spectral semantic prompts through semantic-aware structural consistency across spectra and enhances the representation capability of the spectral features. Extensive experiments show that DEEP significantly outperforms existing methods on person and vehicle benchmarks. In future work, we plan to investigate fine-grained semantics with the visual-language foundation model, to further mine the semantic learning potential on multi-spectral object re-identification.

References

[1] 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," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 618–626.

[2] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the ACM International Conference on Multimedia*, 2018, pp. 274–282.

[3] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 480–496.

[4] A. Lu, Z. Zhang, Y. Huang, Y. Zhang, C. Li, J. Tang, and L. Wang, "Illumination distillation framework for nighttime person re-identification and a new benchmark," *IEEE Transactions on Multimedia*, pp. 1–14, 2023.

[5] W. Chen, I. Chen, C. Yeh, H. Yang, J. Ding, and S. Kuo, "Sjdl-vehicle: Semi-supervised joint defogging learning for foggy vehicle re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 347–355.

[6] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1092–1108, 2020.

[7] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2022.

[8] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 993–15 002.

[9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015, pp. 1116–1124.

[10] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1487–1495.

[11] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3702–3712.

[12] H. Li, A. Zheng, L. Sun, and Y. Luo, "Camera topology graph guided vehicle re-identification," *IEEE Transactions on Multimedia*, vol. 26, pp. 1565–1577, 2024.

[13] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo, "Multi-spectral vehicle re-identification: A challenge," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 11 345–11 353.

[14] A. Zheng, X. Zhu, Z. Ma, C. Li, J. Tang, and J. Ma, "Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark," *Information Fusion*, vol. 100, p. 101901, 2023.

[15] A. Zheng, Z. Wang, Z. Chen, C. Li, and J. Tang, "Robust multi-modality person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 3529–3537.

[16] Z. Wang, C. Li, A. Zheng, R. He, and J. Tang, "Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2633–2641.

[17] Y. Wang, X. Liu, P. Zhang, H. Lu, Z. Tu, and H. Lu, "Top-reid: Multi-spectral object re-identification with token permutation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 5758–5766.

[18] P. Zhang, Y. Wang, Y. Liu, Z. Tu, and H. Lu, "Magic tokens: Select diverse tokens for multi-modal object re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 117–17 126.

[19] Y. Wang, Y. Liu, A. Zheng, and P. Zhang, "Demo: Decoupled feature-based mixture of experts for multi-modal object re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[20] Y. Wang, X. Liu, T. Yan, Y. Liu, A. Zheng, P. Zhang, and H. Lu, "Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[21] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[22] S. Zhang, W. Luo, D. Cheng, Y. Xing, G. Liang, P. Wang, and Y. Zhang, "Prompt-based modality alignment for effective multi-modal object re-identification," *IEEE Transactions on Image Processing*, pp. 1–1, 2025.

[23] H. Li, C. Li, A. Zheng, J. Tang, and B. Luo, "Attribute and state guided structural embedding network for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 31, pp. 5949–5962, 2022.

[24] A. Zheng, P. Pan, H. Li, C. Li, B. Luo, C. Tan, and R. Jia, "Progressive attribute embedding for accurate cross-modality person re-id," in *Proceedings of the ACM International Conference on Multimedia*, 2022, pp. 4309–4317.

[25] J. Wang, A. Zheng, Y. Yan, R. He, and J. Tang, "Attribute-guided cross-modal interaction and enhancement for audio-visual matching," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 4986–4998, 2024.

[26] C. Cui, S. Huang, W. Song, P. Ding, M. Zhang, and D. Wang, "Profd: Prompt-guided feature disentangling for occluded person re-identification," in *Proceedings of the ACM International Conference on Multimedia*, 2024, pp. 1583–1592.

[27] S. He, W. Chen, K. Wang, H. Luo, F. Wang, W. Jiang, and H. Ding, "Region generation and assessment network for occluded person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 120–132, 2024.

[28] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, R. Jin, and X. Sun, "Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 050–15 061.

[29] M. Liu, Z. Zhang, Y. Bian, X. Wang, Y. Sun, B. Zhang, and Y. Wang, "Cross-modality semantic consistency learning for visible-infrared person re-identification," *IEEE Transactions on Multimedia*, vol. 27, pp. 568–580, 2025.

[30] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2540–2549.

[31] T. Liang, Y. Jin, W. Liu, S. Feng, T. Wang, and Y. Li, "Keypoint-guided modality-invariant discriminative learning for visible-infrared person re-identification," in *Proceedings of the ACM International Conference on Multimedia*, 2022, pp. 3965–3973.

[32] X. Liu, K. Liu, J. Guo, P. Zhao, Y. Quan, and Q. Miao, "Pose-guided attention learning for cloth-changing person re-identification," *IEEE Transactions on Multimedia*, vol. 26, pp. 5490–5498, 2024.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International conference on machine learning*, vol. 139, 2021, pp. 8748–8763.

[34] Y. Zhang, K. Yu, S. Wu, and Z. He, "Conceptual codebook learning for vision-language models," in *Proceedings of the European Conference on Computer Vision*, vol. 15135, 2024, pp. 235–251.

[35] Y. Zhang, C. Zhang, K. Yu, Y. Tang, and Z. He, "Concept-guided prompt learning for generalization in vision-language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 7377–7386.

[36] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[37] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 816–16 825.

[38] S. Li, L. Sun, and Q. Li, "Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 1405–1413.

[39] Z. Yang, D. Wu, C. Wu, Z. Lin, J. Gu, and W. Wang, "A pedestrian is worth one prompt: Towards language guidance person re- identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 343–17 353.

[40] Y. Zhai, Y. Zeng, Z. Huang, Z. Qin, X. Jin, and D. Cao, "Multi-prompts learning with cross-modal alignment for attribute-based person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 6979–6987.

[41] Z. Hu, B. Yang, and M. Ye, "Empowering visible-infrared person re-identification with large foundation models," in *Advances in Neural Information Processing Systems*, 2024.

[42] Z. Yu, Z. Huang, M. Hou, J. Pei, Y. Yan, Y. Liu, and D. Sun, "Representation selective coupling via token sparsification for multi-spectral object re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[43] A. Zheng, Z. Ma, Y. Sun, Z. Wang, C. Li, and J. Tang, "Flare-aware cross-modal enhancement network for multi-spectral vehicle re-identification," *Information Fusion*, vol. 116, p. 102800, 2025.

[44] Z. Wang, H. Huang, A. Zheng, and R. He, "Heterogeneous test-time training for multi-modal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 5850–5858.

[45] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: personalizing text-to-image generation using textual inversion," in *Proceedings of the International Conference on Learning Representations*, 2023.

[46] K. Saito, K. Sohn, X. Zhang, C. Li, C. Lee, K. Saenko, and T. Pfister, "Pic2word: Mapping pictures to words for zero-shot composed image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 305–19 314.

[47] J. Cho, G. Nam, S. Kim, H. Yang, and S. Kwak, "Promptstyler: Prompt-driven style generation for source-free domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 656–15 666.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proceedings of the Conference on Neural Information Processing Systems*, vol. 30, 2017.

[49] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[50] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[51] X. Qian, Y. Fu, Y. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 5399–5408.

[52] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2109–2118.

[53] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2285–2294.

[54] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1005–1014.

[55] J. Crawford, H. Yin, L. McDermott, and D. Cummings, "Unicat: Crafting a stronger fusion baseline for multimodal re-identification," *arXiv preprint arXiv:2310.18812*, 2023.

[56] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 2223–2232.

[57] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015, pp. 1116–1124.

[58] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "Fastreid: A pytorch toolbox for general instance re-identification," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 9664–9667.

[59] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 13 001–13 008.

[60] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[61] Z. Wang, F. Zhu, S. Tang, R. Zhao, L. He, and J. Song, "Feature erasing and diffusion network for occluded person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4744–4753.

[62] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

[63] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[64] Z. Chen, Z. Zhang, X. Tan, Y. Qu, and Y. Xie, "Unveiling the power of CLIP in unsupervised visible-infrared person re-identification," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 3667–3675.

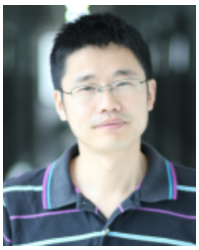
[65] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern recognition*, vol. 95, pp. 151–161, 2019.

[66] L. Ding, L. Liu, Y. Huang, C. Li, C. Zhang, W. Wang, and L. Wang, "Text-to-image vehicle re-identification: Multi-scale multi-view cross-modal alignment network and a unified benchmark," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 7673–7686, 2024.

Shihao Li is currently a Ph.D. student in the School of Artificial Intelligence, Anhui University, Hefei, China. He received the B.Eng. degree from Tongling University, Tongling, China, in 2017. His current research interests include computer vision, artificial intelligence and object re-identification.

Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he was a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Postdoctoral Research Fellow with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently a Professor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision and deep learning. Dr. Li was the recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

Aihua Zheng received B.Eng. degrees and finished Master-Docter combined program in Computer Science and Technology from Anhui University of China in 2006 and 2008, respectively. And received Ph.D. degree in computer science from University of Greenwich of UK in 2012. She visited University of Stirling and Texas State University during June to September in 2013 and during September 2019 to August 2020 respectively. She is currently a Professor and PhD supervisor at the School of Artificial Intelligence, Anhui University. Her main research interests include vision based artificial intelligence and pattern recognition. Especially on person/vehicle re-identification, audio visual computing, and multimodal intelligence.



Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in Computer Science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor and PhD supervisor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision, pattern recognition, and machine learning.

Bin Luo received the B.Eng. degree in electronics and the M.Eng. degree in computer science from the Anhui University of China, Hefei, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, York, U.K., in 2002. He has authored or coauthored more than 200 papers in journals and refereed conferences. He is currently a Professor with the Anhui University of China. His research interests include random graphbased pattern recognition, image and graph matching, and spectral analysis. He is currently the Chair of the IEEE Hefei Subsection. He was a peer Reviewer of international academic journals such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition*, *Pattern Recognition Letters*.