

ICPL-ReID: Identity-Conditional Prompt Learning 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) brings a new perception perspective for smart city and intelligent transportation applications, effectively addressing challenges from complex illumination and adverse weather. However, complex modal differences between heterogeneous spectra pose challenges to efficiently utilizing complementary and discrepancy of spectra information. Most existing methods fuse spectral data through intricate modal interaction modules, lacking fine-grained semantic understanding of spectral information (*e.g.*, text descriptions, part masks, and object keypoints). To solve this challenge, we propose a novel Identity-Conditional text Prompt Learning framework (ICPL), which exploits the powerful cross-modal alignment capability of CLIP, to unify different spectral visual features from text semantics. Specifically, we first propose the online prompt learning using learnable text prompt as the identity-level semantic center to bridge the identity semantics of different spectra in online manner. Then, in lack of concrete text descriptions, we propose the multi-spectral identity-condition module to use identity prototype as spectral identity condition to constraint prompt learning. Meanwhile, we construct the alignment loop mutually optimizing the learnable text prompt and spectral visual encoder to avoid online prompt learning disrupting the pre-trained text-image alignment distribution. In addition, to adapt to small-scale multi-spectral data and mitigate style differences between spectra, we propose multi-spectral adapter that employs a low-rank adaption method to learn spectra-specific features. Comprehensive experiments on 5 benchmarks, including RGBNT201, Market-MM, MSVR310, RGNB300, and RGBNT100, demonstrate that the proposed method outperforms the state-of-the-art methods. The source code is publicly available at <https://github.com/lsh-ahu/ICPL-ReID>.

Index Terms—Multi-Spectral Object Re-Identification, Online Prompt Learning, Multi-Spectral Identity Condition, Low-Rank Adaption.

I. INTRODUCTION

E-identification (ReID) aims to match images with the same identity from the gallery repository based on specified query conditions [1]–[4]. With the development of com-

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, the Natural Science Foundation of Anhui Province under Grants 2308085Y40 and 2208085J18, and the Open Research Project of the Anhui Provincial Key Laboratory of Security Artificial Intelligence under Grant SAI202401. (*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).

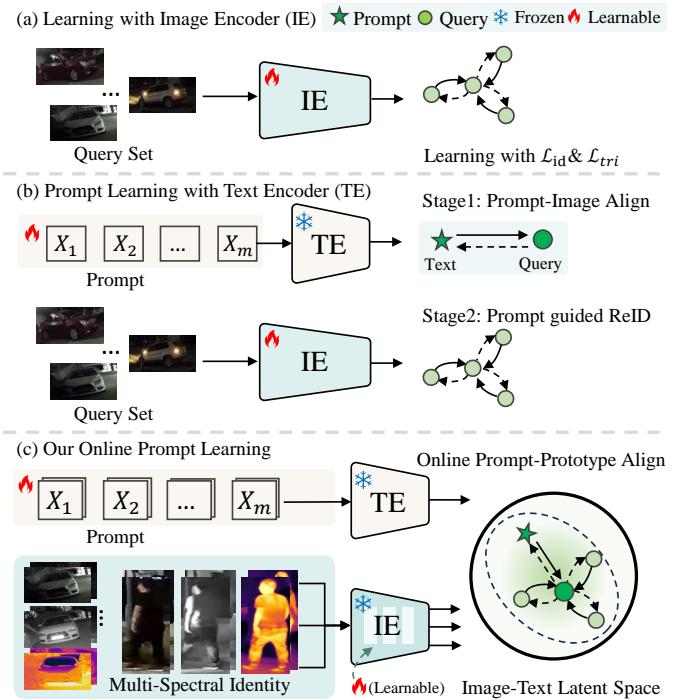


Fig. 1. (a) Classical pre-training models require ReID learning with id loss and triplet loss [3], [5]. (b) The existing research introduces a two-stage text prompt learning [14], which pre-aligned the text prompt for each identity and fine-tuned the ReID task with the text prompt separately. (c) Our method proposes an end-to-end text prompt learning framework, seamlessly integrates text prompt learning with multi-spectral ReID task, and alleviates the discrepancies between multi-spectral data.

puter vision technology, ReID is increasingly applied within intelligent video surveillance systems. However, challenges such as adverse weather conditions, illumination changes, camera viewpoint variations, and background occlusions still impede the practical application of ReID algorithms. As a result, academic and industrial communities have sparked a broad research fervor [3], [5]–[9], among which introducing auxiliary infrared spectra to complement visual data has drawn growing attention [10]–[13].

As infrared imaging devices become prevalent in practical production, their unique imaging principle gives people a new perception perspective. Particularly in low visibility environments such as nighttime, heavy fog, and rainy day [15]–[18], infrared spectra provides extra discriminative information for identifying challenging queries. In order to study this problem, Li *et al.* [10] and Zheng *et al.* [11], [19] first propose the multi-spectral vehicle and person ReID research tasks and established high-quality benchmarks, extensively promoting

research progress in this field. Their contributions underscore the significance of multi-spectral ReID and highlight the challenges encountered in harnessing multi-spectral data effectively.

To effectively utilize multi-spectral data, most existing methods favor a modal-fusion fashion [19]–[22]. However, these methods primarily use visual encoders with intricate model designs to learn modal interactions between spectra, overlooking the fine-grained semantics between infrared and RGB for the same identity, such as text descriptions [23], [24], part masks [25]–[27], and object keypoints [28], [29]. This prior information explicitly guides the model to further focus on the fine-grained semantic features that distinguish identities. Unlike traditional pre-trained models [30]–[32], CLIP [33] is trained on large-scale vision-language data, resulting in robust semantic alignment capability. The proposed prompt learning [34], [35] further demonstrates its robust generalization and semantic alignment capability across diverse downstream tasks. Inspired by these methods, Li *et al.* [14] and Chen *et al.* [36] propose learning text semantics through learnable text prompts combined with the frozen CLIP text branch. These methods enable the ReID task to effectively capture semantic information from images, even without concrete text labels.

Methods such as CLIP-ReID [14] and CCLNet [36] propose the two-stage text prompt learning method, exploring the application of image-prompt alignment paradigm in ReID tasks. However, as shown in Fig. 1 (b), the two-stage method requires additional overhead for text prompt pre-alignment, and the first-stage visual branch is frozen to extract multi-spectral features. This prevents text prompt from learning spectral-specific semantic features. Consequently, the text prompt can only provide fixed semantic constraints during the second-stage of spectra modalities training, preventing the model from aligning to spectra semantics and resulting in suboptimal solutions. To address this challenge, as illustrated in Fig. 1 (c), we propose an end-to-end joint optimization method for text prompt and multi-spectral learning. By mutually aligning learnable text prompt and optimizing spectral visual encoder in an online text prompt learning manner, this method mitigates the semantic shift between text prompt and spectra that occur with separate pre-alignment.

Online text prompt learning is not straightforward. Due to the lack of concrete text descriptions, the model cannot observe the real spectral and text alignment distribution, and still relies on the pre-trained alignment space of CLIP to align learnable text prompt to spectral modalities. However, direct online spectral alignment inevitably leads to image domain shift when adapting to infrared spectral data which have significant stylistic discrepancies with RGB image, resulting in image modality features deviating from the pre-trained image-text alignment distribution [37]–[40]. To address this problem, we propose the multi-spectral identity condition module to use the multi-spectral identity prototypes as the condition for prompt learning, replacing the spectra instances with these prototypes in online alignment to mitigate the semantic shift after adapting to spectral data, and providing robust multi-spectral identity constraints for text prompt. Additionally, we employ a momentum update method to dynamically aggregate

identity prototypes, enabling the learnable text prompt to collaboratively and progressively learn new spectral semantic features during training.

Although online prompt learning has mitigated the semantic shift issue between text prompt and spectral modalities, existing multi-spectral ReID methods still rely on full fine-tuning to learn spectra-specific features. However, due to the smaller scale and significant stylistic discrepancies in multi-spectral data compared to RGB image data, there is a risk of overfitting and dependency on certain spectral modalities [40]–[42]. To address these challenges, we propose the multi-spectral adapter, which uses a low-rank adaption method with the lightweight learnable adapter to adapt to different spectra modalities. This approach freezes the original pre-trained model and fine-tunes the few spectra-specific parameters, allowing us to learn spectra features without disturbing the pre-trained image-text alignment distribution.

In summary, we leverage the strong image-text alignment capability of the vision-language pre-training model, aligning multi-spectral modalities with learnable text prompt. By adopting our online prompt learning method, we effectively utilize the spectra data for object ReID tasks. The main contributions of this paper are summarized as follows:

- We propose a novel online prompt learning framework for the multi-spectral object ReID. To our best knowledge, this is the first work that fully leverages the image-text alignment capabilities of CLIP to enhance the multi-spectral object ReID task.
- To construct mutual alignment and optimization between text prompt and spectral visual encoder, we propose a multi-spectral identity condition module, which utilizes dynamically updated identity prototypes as constraint condition and constructs alignment loop for prompt learning in the lack of concrete text descriptions.
- To adapt to multi-spectral data, we propose the multi-spectral adapter, using a low-rank adaptation approach with the lightweight learnable adapter to learn spectral-specific parameters and maintain the pre-trained image-text alignment distribution.
- To validate the effectiveness of our method, we conduct extensive experiments on five multi-spectral benchmarks, including person and vehicle datasets. The results demonstrated that our method significantly outperformed the state-of-the-art approaches.

II. RELATED WORK

A. Multi-Spectral Object ReID

Multi-spectral object ReID introduces near- and thermal-infrared modalities to enhance the robustness of the model in adverse environments, receiving increasing attention in recent years. Unlike single-model object ReID, which requires prior knowledge such as part segmentation, low-light enhancement, and defogging learning to deal with occlusion, nighttime, heavy fog, and domain discrepancy [15]–[18], [43], [44], multi-spectral object ReID naturally has the advantage of solving these challenges due to its diversity in imaging principle. Although cross-modal object ReID [45]–[47] achieves target

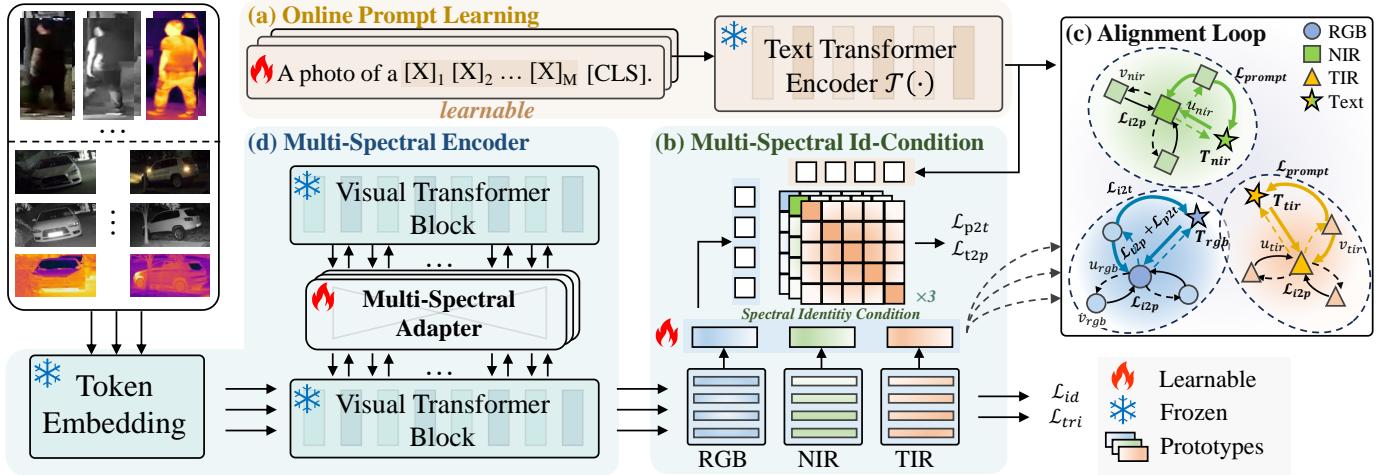


Fig. 2. Pipeline of our proposed framework. (a) For end-to-end training of multi-spectral ReID, online prompt learning leverages learnable text prompt as cross-modal constraints to jointly optimize ReID tasks. (b) The multi-spectral identity (id)-condition module first aggregates the RGB, NIR, and TIR spectral features into identity prototypes, and replaces instance features with prototypes to guide text prompt learning. The dynamically updated strategy enables the alignment of spectral-specific semantic features during training. (c) The alignment loop enables mutual optimization between the text prompts and spectral encoder. It consists of L_{prompt} and L_{i2p} , where $L_{t2p} + L_{p2t}$ within L_{prompt} help the text prompt T_m to learn the semantic information of the spectral identity prototype u_m . Meanwhile, the image-to-text alignment loss L_{i2t} , guides the spectral instance v_m by leveraging semantics without concrete text descriptions. (d) The multi-spectral encoder freezes and shares most parameters of the visual branch in CLIP for each spectral modality by adding a low-rank adapter to adapt spectra-specific data.

retrieval across time and scenes by introducing near-infrared or textual modality [48]–[51], it may weaken the unique characteristics of these modalities in the process of eliminating modality discrepancies [52], [53]. To address the above issues, Li *et al.* [10] pioneer the construction of a multi-spectral vehicle ReID benchmark with visible and infrared modalities as queries, and propose HAMNet to fuse spectra features in a heterogeneity-collaboration aware manner. Zheng *et al.* [19] propose a high-quality multi-spectral vehicle ReID dataset, dubbed MSVR310, covering a broader range of viewpoints, longer time spans, and more environmental complexities, to obtain more consistent multi-spectral feature distributions via cross-directional consistency networks, called CCNet. In the field of person ReID, Zheng *et al.* [11] propose the first multi-spectral person ReID dataset, named RGBNT201, and mine complementary features between modalities through progressive fusion with PFNet. Wang *et al.* [54] propose to mine modality-specific features through cross-modal interaction, relationship-based enhanced modality, and multi-modal margin loss. In contrast to CNN-based methods, the rise of Transformer has brought a new attention mechanism to multi-spectral object ReID. Wang *et al.* [21] propose TOP-ReID, which uses the token permutation module to perform cross-attention fusion of three-spectral features, and propose the complementary reconstruction module to reduce the gap in feature distribution between spectra by reconstructing token-level modal features. Zhang *et al.* [22] propose an object-centric selection method, called EDITOR, which uses a spatial-frequency token selection module to filter discriminative tokens in each spectra and aggregate token features from different spectra into a multi-spectral representation, providing higher interpretability for multi-spectra ReID task. Wang *et al.* [20] propose HTT, which improves the representation ability of multi-spectral descriptor by constraining the sample distri-

bution spacing between spectra based on ViT, and propose a multi-modal test-time training strategy to improve the model's generalization on unseen test data using self-supervised loss. However, most multi-modal methods do not consider the identity semantic consistency between spectral modalities. We propose to utilize the vision-language pre-training model CLIP [33] to mine more discriminative identity semantic features through the prompt learning approach.

B. Vision-Language Pre-training Model

In recent years, the rise of vision-language pre-training models represented by CLIP [33], [55], has increasingly attracted research attention due to their powerful representation and generalization capability on downstream tasks. Zhou *et al.* [34], [35] propose CoOp and CoCoOp, which achieve excellent transfer performance across various downstream tasks through text learnable prompt. Jia *et al.* [56] and Chen *et al.* [57] propose using a few learnable parameters to fine-tune the pre-trained model, achieving fewer resources and more efficient fine-tuning performance. However, these methods only explore the common classification tasks, which are distinctly different from ReID task settings. CLIP-ReID [14] proposes using learnable text prompt to guide the learning of ReID task, but the two-stage learning method lacks online alignment between image and text prompt. He *et al.* [26] combine component segmentation with learnable text prototypes and proposes an adaptive region generation and assessment method to address the challenge of person occlusion ReID, dubbed RGANet. However, this method focuses on the occlusion challenge and relies on additional segmentation supervision signal. CCLNet [36] employs learnable text prompt as soft ID labels to contend with the challenges of weak pseudo-label signals and substantial label noise in unsupervised cross-modal person ReID. However, limited by the lack of accurate ID

labels, it still cannot dynamically align identity text with image features. Different from existing prompt learning methods, we propose using online image-text alignment to transfer pre-trained models to multi-spectral object ReID task more effectively.

III. METHODOLOGY

In this section, we elaborate on the specific details of the proposed framework, which is trained in an end-to-end fine-tuning manner for multi-spectral object ReID. As shown in Fig. 2, it is comprised of the Online Prompt Learning Strategy (Sec.III.B), Multi-Spectral Identity Condition Module (Sec.III.C), and Multi-Spectral Adapter Module (Sec.III.D).

A. Overview

First, we introduce the basic pipeline for multi-spectral object ReID training based on CLIP and define relevant symbol definitions. Thanks to the image-text contrastive pre-training method and the large-scale training data, the vanilla CLIP model exhibits strong zero-shot capabilities with visual encoder $\mathcal{I}(\cdot)$ and text encoder $\mathcal{T}(\cdot)$. During training, we freeze them and adopt the parameter-efficient fine-tuning approach to better transfer their generalization ability across diverse visual modalities in multi-spectral object ReID.

In contrast to single- and cross-modal object ReID, multi-spectral object ReID introduces multiple spectra to provide additional auxiliary information. Each sample within the query and gallery sets is defined as $X^i = \{x_{rgb}^i, x_{nir}^i, x_{tir}^i\}$, where x^i are the i -th sample containing RGB, Near Infrared (NIR) and Thermal Infrared (TIR) heterogeneous visual modalities. Leveraging the Multi-Spectral Identity Condition module, we aggregate image features into identity prototypes $U^c = \{u_{rgb}^c, u_{nir}^c, u_{tir}^c\}$, where u^c are multi-spectral cluster centers of the c -th identity. The Online Prompt Learning strategy treats the learnable text modality as identity-level learnable vector $T^c = \{t_{rgb}^c, t_{nir}^c, t_{tir}^c\}$, where t^c are text prompt paired with the u^c for the c -th identity. Conditioned by identity prototypes U^c , we employ text prompt as cross-modality constraints to align image features $V^i = \{v_{rgb}^i, v_{nir}^i, v_{tir}^i\}$, where v^i are image features of the i -th sample.

By fully exploiting the image-text cross-modal alignment capabilities of large-scale pre-trained models, our method avoids designing complex spectral modality interaction modules. During the test inference, only the spectral features need to be concatenated as the final representation.

B. Online Prompt Learning

Detailed in Fig. 2, the basic multi-spectral ReID uses a triplet-stream visual encoder $\mathcal{I} = \{\mathcal{I}_{rgb}, \mathcal{I}_{nir}, \mathcal{I}_{tir}\}$ to extract spectral features, and uses metric learning methods to obtain a compact and separable feature distribution. However, the heterogeneity between spectra poses challenges for intra-spectra identity alignment. We propose to use learnable identity semantic prompt T to guide the identity alignment of multi-spectral ReID. The classical prompt learning method CoOp [34] defines the text prompt $T =$

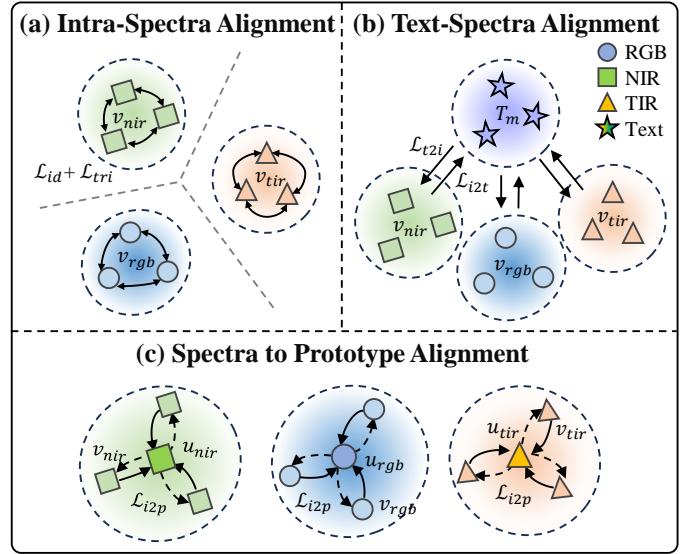


Fig. 3. Illustration of traditional instance feature learning strategy. (a) The classic ReID metric learning method employs \mathcal{L}_{id} and \mathcal{L}_{tri} to enhance intra-class compactness and inter-class separability within the spectra. (b) Cross-modal semantic alignment between text and spectra is typically achieved by constructing a latent text-image alignment space with symmetric \mathcal{L}_{i2t} and \mathcal{L}_{t2i} losses. (c) To bring spectral features closer to the prototype and enhance the perception of global sample features within each spectral instance.

$“[X]_1, [X]_2, \dots, [X]_M, [CLS]”$ as a learnable feature, and optimizes prompt with the frozen visual branch. However, the training process of ReID is distinctly different from this training strategy. Typically, we need to unfreeze the visual encoder \mathcal{I} and learn the fine-grained features of query samples to obtain the final fine-tuned encoder \mathcal{I}' . During training, frequently changing image features V can make it challenging to align the learnable text prompt T . One solution is to use a two-stage method to separate the training of text prompt and visual encoder, such as CLIP-ReID [14]. In this method, the frozen visual encoder \mathcal{I} is used to pre-align the prompt T , and then the frozen prompt T is used as an unlearnable classifier to optimize the visual encoder \mathcal{I}' . However, this approach does not account for the visual encoder \mathcal{I} changing the original visual feature distribution after adapting to spectral modalities with significant stylistic discrepancies, which broadens the distribution gap between the text prompt T learned in the first stage and the optimized spectral feature V in the second stage.

To this end, we propose an online prompt learning training method that collaboratively trains text prompt learning with multi-spectral ReID task. Specifically, we define learnable text prompt of each identity described as “a photo of a $[X]_m^1, [X]_m^2, \dots, [X]_m^M, [CLS]$ ”, $m \in [rgb, nir, tir]$, as illustrated in Fig. 2 (a). Each learnable token $[X]_m$ is randomly initialized, and $[CLS]$ is object class, (e.g., person or vehicle). Meanwhile, we use the learnable visual encoder \mathcal{I} to learn the features V of the ReID task. We use the loss \mathcal{L}_{i2t} to pull the visual features closer to the text prompt, and the loss \mathcal{L}_{t2i} to bring text prompt closer to the visual features. The specific formula is as follows:

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

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

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

C. Multi-Spectral Identity Condition

To avoid online prompt learning disrupting the pre-trained text-image alignment distribution, we propose the Multi-Spectral Identity Condition module. It consists of two components: the Spectral Identity Condition to generate the identity condition by dynamically aggregating instance prototypes, and the Alignment Loop to mutually optimize text prompt and spectral encoder via the identity condition.

Spectral Identity Condition. Online multi-spectral alignment is challenging due to the lack of concrete text descriptions, making it not a genuine image-text cross-modal task. Solely fine-tuning the visual encoder \mathcal{I} inevitably disrupts the pre-trained image-text alignment distribution, and the significant stylistic discrepancies between different spectra intensify this issue. To constrain the learning of text prompt, we employ an identity-conditional method during the training process to cluster samples of identical identities into prototypes. As shown in Fig. 2 .(b), it gradually aligns the learnable prompt to the multi-spectral modalities and collaboratively pulls the spectral modalities to the same identity semantic center. Specifically, before each training epoch, we aggregate image features belonging to the same identity to yield multi-spectral prototypes $U_m = \{u_m^1, u_m^2, \dots, u_m^j\}$:

$$u_m^c = \frac{1}{N_m} \sum_{j=1}^{N_m} v_m^{c,j}, N_m = |v_m^c|, \quad (3)$$

where $v_m^{c,j}$ are the j -th instance feature of image sample in the c -th identity, u_m^c are prototypes of the c -th identity, and m is a spectral modality in $\{rgb, nir, tir\}$.

During the training phase, we randomly select P identities and N samples for each identity. Each sample contains M different spectra modalities, accumulating to a total of $P \times N \times M$ images for a mini-batch. Subsequently, we apply a momentum update mechanism to dynamically refresh the prototypes of each identity within the memory bank, ensuring that they evolve in sync with the training process.

$$u_m^{c,l+1} = \alpha \cdot u_m^{c,l} + (1 - \alpha) \cdot v_m^{c,j}, \quad (4)$$

where l and $l + 1$ are the index of the current and next iteration, and α is the updating factor that controls the feature propagation impacts.

To enhance the robustness of identity prototype, we employ an image-to-prototype alignment loss, denoted as \mathcal{L}_{i2p} , to constrain samples that share the same identity across different spectra:

$$\mathcal{L}_{i2p} = -\frac{1}{M} \log \frac{\exp(\langle v_m^j, u_m^+ \rangle / \gamma)}{\sum_{i=1}^{N_m} \exp(\langle v_m^j, u_m^i \rangle / \gamma)}, \quad (5)$$

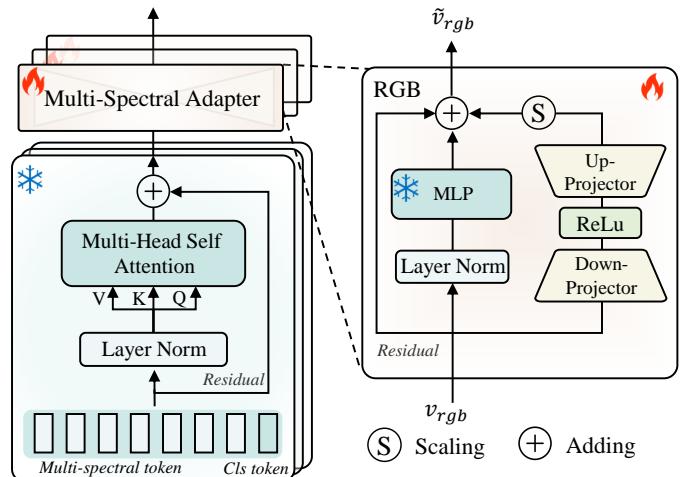


Fig. 4. Architecture of our multi-spectral adapter.

where u_m^+ is the positive feature of v_m^j , and γ is a temperature hyper-parameter.

Alignment Loop. As shown in Fig. 2 (c), the individual image feature V^i is replaced by identity prototype set $U^c = \{u_{rgb}^c, u_{nir}^c, u_{tir}^c\}$. Owing to the dynamically updated identity prototypes, the learnable text prompt are able to focus on various samples of the same identity. This enables the text prompt to continually observe the latest optimized multi-spectral features during the training process. Specifically, we employ the contrastive loss \mathcal{L}_{t2p} and \mathcal{L}_{p2t} to align the text prompt with the prototypes:

$$\mathcal{L}_{t2p} = -\frac{1}{M} \log \frac{\exp(\langle t_m^c, u_m^c \rangle / \gamma)}{\sum_{k=1}^{N_m} \exp(\langle t_m^c, u_m^k \rangle / \gamma)}, \quad (6)$$

$$\mathcal{L}_{p2t} = -\frac{1}{M} \log \frac{\exp(\langle u_m^c, t_m^c \rangle / \gamma)}{\sum_{k=1}^{N_m} \exp(\langle u_m^c, t_m^k \rangle / \gamma)}, \quad (7)$$

where u_m^c, t_m^c are the positive pair of the c -th identity, and γ is a temperature hyper-parameter.

Intuitively, as depicted in Fig. 3 (a) and Fig. 3 (b), both the object ReID and the text-image alignment tasks require many-to-many feature learning. This makes it challenging to effectively learn text prompts and guide the model to recognize identities, particularly in multi-spectral datasets that exhibit low-quality noise and style discrepancies. However, as shown in Fig. 3 (c), identity prototypes simplify this to a many-to-one problem, effectively alleviating the complexity of prompt optimization.

$$\mathcal{L}_{prompt} = \lambda_1 \cdot \mathcal{L}_{i2t} + \lambda_2 \cdot (\mathcal{L}_{t2p} + \mathcal{L}_{p2t}), \quad (8)$$

as shown above Eq. (8), we construct the complete alignment loop for online text prompt learning by replacing Eq. (2) with Eq. (6) and (7). The hyper-parameters λ_1 and λ_2 are simply set to further smooth the learning of text prompts and their alignment with spectral features during training. The detailed training pseudo-code is shown in Algorithm 1.

D. Multi-Spectral Adapter

Existing pre-trained models primarily focus on RGB images. Although the same object has similar semantics in

different spectra, there are still significant discrepancies in stylistic. This issue poses challenges for semantic alignment between spectra. Additionally, due to the lack of large-scale multi-spectral data, the model may rely on certain spectra when fully fine-tuning on small-scale datasets, and suffer from catastrophic forgetting problems. To address this challenge, we propose a simple yet effective multi-spectral adaption module that utilizes a low-rank adaption approach to adapt spectra-specific features, as shown in Fig. 2 (d).

In detail, we first freeze most parameters of the visual encoder \mathcal{I} , preserving only the classification layer, and the last batch normalization layer for training. As shown in Fig. 4, we introduce a lightweight learnable adapter in the feed-forward network of each transformer block. For each adapter, the input features are compressed to the \tilde{d} -dimensional through the channel down-projection layer, denoted as $W_{down} \in R^{(d \times \tilde{d})}$. Subsequently, the features are re-expanded to the original d -dimension through the channel up-projection layer, denoted as $W_{up} \in R^{(\tilde{d} \times d)}$. The \tilde{d} -dimension is the intermediate dimension in the bottleneck layer that is smaller than the d -dimension. A non-linear activation layer ReLU is used to introduce a non-linear transformation for the bottleneck layer between the two linear projection layers. Finally, this bottleneck network is connected to the original feed-forward network through residual connections with a scaling factor s . Formal description is as follows:

$$\begin{aligned} \tilde{v}_m = & s \cdot \text{ReLU}(LN(v_m) \cdot W_{down}) \cdot W_{up} \\ & + FFN(v_m) + v_m, \end{aligned} \quad (9)$$

where \tilde{v}_m is the optimized spectra feature, which is used as the input for the next block.

E. Optimization

As in vanilla ReID task setting [5], we use identity classification loss \mathcal{L}_{id} to each sample for id constraints, and triplet loss \mathcal{L}_{tri} to pull together samples sharing the same identity.

$$\mathcal{L}_{id} = \sum_{i=1}^N -q_i \log(p_i) \begin{cases} q_i = 0, y \neq i \\ q_i = 1, y = i \end{cases} \quad (10)$$

$$\mathcal{L}_{tri} = \max(d_p - d_n + \delta, 0), \quad (11)$$

where y as truth ID label and p_i as ID prediction logits of class i . d_p and d_n are feature distances of positive pair and negative pair. δ is the margin of triplet loss.

The complete loss function \mathcal{L}_{final} is defined as follows:

$$\mathcal{L}_{final} = \mathcal{L}_{id} + \mathcal{L}_{tri} + \lambda_3 \cdot \mathcal{L}_{i2p} + \mathcal{L}_{prompt}, \quad (12)$$

where λ_3 serves as a hyper-parameter to balance the training process.

IV. EXPERIMENT

In this section, we conduct detailed experiments on the proposed framework. First, we introduce the datasets, evaluation protocols, and implementation details (Sec.IV.A-B). Second, we conduct comparative experiments with the latest methods on person and vehicle datasets (Sec.IV.C). Then, we perform ablation experiments and visual analysis of the proposed method (Sec.IV.D). Finally, we further analyze the components of the framework (Sec.IV.E).

Algorithm 1 Identity-conditional prompt learning process.

Input: Multi-spectral training data X_{rgb} , X_{nir} , X_{tir} .
Parameter: Learnable text tokens $[X]_{rgb}$, $[X]_{nir}$, $[X]_{tir}$, an image encoder \mathcal{I} , a text encoder \mathcal{T} and update momentum α .
Output: The final multi-spectral loss \mathcal{L}_{final} .

```

1: Initialize  $\mathcal{I}$ ,  $\mathcal{T}$  from the pre-trained CLIP.
2: for  $n$  in [1, epochs] do
3:   // Extract identity prompt and prototype.
4:    $T_{rgb}, T_{nir}, T_{tir} = \mathcal{T}([X]_{rgb}, [X]_{nir}, [X]_{tir})$ 
5:    $U_{rgb}, U_{nir}, U_{tir} = \text{average}(\mathcal{I}(X_{rgb}, X_{nir}, X_{tir}))$ 
6:   for  $i$  in [1, iterations] do
7:     // Sample a batch samples from  $X_{rgb}$ ,  $X_{nir}$ ,  $X_{tir}$ .
8:      $v_{rgb}, v_{nir}, v_{tir} = \mathcal{I}(x_{rgb}, x_{nir}, x_{tir})$ 
9:     // Get prompt from  $[X]_{rgb}$ ,  $[X]_{nir}$ ,  $[X]_{tir}$ .
10:     $t_{rgb}, t_{nir}, t_{tir} = \mathcal{T}([x]_{rgb}, [x]_{nir}, [x]_{tir})$ 
11:    Optimize  $[X]_1, [X]_2, \dots, [X]_m$  via Eq. (6) and Eq. (7).
12:    Optimize visual branch via Eq. (1) and Eq. (5).
13:    // Update text prompt and prototype.
14:     $T^{i+1} = t^i; U^{i+1} = \alpha \cdot U^i + (1 - \alpha) \cdot v^i$ 
15:    Calculate id and triplet loss via Eq. (10) and Eq. (11).
16:  end for
17: end for

```

A. Datasets and Evaluation Protocols

We conduct experiments on five publicly available multi-spectral datasets, including two multi-spectral person ReID datasets RGBNT201 [11] and Market-MM [54], and the multi-spectral vehicle ReID datasets MSVR310 [19], RGBNT100, and RGBN300 [10].

RGBNT201 [11] contains 14,361 person images, totaling 4,787 samples, each consisting of 3 spectra modalities, for 201 persons. Within the dataset, 141 identities are divided into the training set, 30 identities into the validation set, and another 30 into the testing set. These samples cover four non-overlapping perspectives. The entire test set is also utilized as a gallery and query set during the testing phase.

Market-MM [54] is a synthetic dataset generated based on the single-modality Market1501 [4], with a total of 1501 identities and 32,668 sets of multi-spectral samples. The training set comprises 751 identities and 12,936 triples, and the rest 750 identities compose the gallery set with 19,732 triples, while the query set contains 750 identities and 3,368 triples. To synthesize multi-spectral data, thermal-infrared spectra are generated from RGB images using CycleGAN, near-infrared spectra are created by converting RGB images to grayscale, and RGB images are reduced by 60% brightness to simulate night scenes.

MSVR310 [19] contains 6,261 high-quality vehicle images, divided into 310 different vehicles, with 2,087 samples, each consisting of 3 spectra modalities. The training set includes 155 vehicles and a total of 1,032 samples. The gallery set contains 1,055 samples of the remaining 155 vehicles, while the query set consists of 52 randomly selected vehicles and 591 samples from the gallery set. These samples are captured at long time spans, covering 8 viewpoints around the vehicle

TABLE I

COMPARISON PERFORMANCES WITH THE STATE-OF-THE-ART METHODS ON RGBNT201. THE BEST AND SECOND-BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE, RESPECTIVELY.

	Methods	Venue	mAP	R-1	R-5	R-10
Single	MUDDeep [58]	ICCV17	23.8	19.7	33.1	44.3
	MLFN [59]	CVPR18	24.7	23.7	38.5	49.5
	PCB [16]	ECCV18	32.8	28.1	37.4	46.9
	HACNN [60]	CVPR18	19.3	14.7	25.5	32.8
	OSNet [6]	ICCV19	22.1	22.9	37.2	45.9
	CAL [61]	ICCV21	27.6	24.3	36.5	45.7
Multi	HAMNet [10]	AAAI20	27.7	26.3	41.5	51.7
	PFNet [11]	AAAI21	38.5	38.9	52.0	58.4
	IEEE [54]	AAAI22	46.4	47.1	58.5	64.2
	UniCat [62]	NIPS23	57.0	55.7	-	-
	HTT [20]	AAAI24	71.1	73.4	83.1	87.3
	TOP-ReID [21]	AAAI24	72.3	76.6	84.7	89.4
	EDITOR [22]	CVPR24	66.5	68.3	81.1	88.2
	CLIP-ReID [14]	AAAI23	71.1	71.8	80.3	85.6
	ICPL	Ours	75.1	77.4	84.2	87.9

TABLE II

COMPARISON PERFORMANCES WITH THE STATE-OF-THE-ART METHODS ON MARKET-MM. THE BEST AND SECOND BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE, RESPECTIVELY. HERE THE SUPERSCRIPT * REPRESENTS THE RESULTS ARE REPRODUCED BY US.

	Methods	Venue	mAP	R-1	R-5	R-10
Single	MLFN [59]	CVPR18	42.7	68.1	87.1	92.0
	HACNN [60]	CVPR18	42.9	69.1	86.6	92.2
	OSNet [6]	ICCV19	39.7	69.3	86.7	91.3
Multi	HAMNet [10]	AAAI20	60.0	82.8	92.5	95.0
	PFNet [11]	AAAI21	60.9	83.6	92.8	95.5
	IEEE [54]	AAAI22	64.3	83.9	93.0	95.7
	HTT [20]	AAAI24	67.2	81.5	95.8	97.8
	TOP-ReID* [21]	AAAI24	82.0	92.4	97.6	98.6
	EDITOR* [22]	CVPR24	77.4	90.8	96.8	98.3
	CLIP-ReID [14]	AAAI23	<u>82.5</u>	<u>93.7</u>	<u>97.9</u>	<u>98.8</u>
	ICPL	Ours	85.1	94.7	98.4	99.1

and various challenges such as illumination change, shadow, reflection, and color distortion.

RGBN300 and RGBNT100 [10] RGBN300 contains 50,125 sample pairs of 300 different vehicles, each pair containing both RGB and near-infrared modality. Each vehicle is collected by 2 to 8 camera views, with 50 to 200 image pairs. The training set randomly selects 150 vehicles with 25,200 image pairs, the rest 150 vehicles with 24,925 image pairs as the gallery set. From these, 4,985 image pairs are used as the query set. On this basis, RGBNT100 selected 100 vehicles and added 17,250 additional thermal-infrared images to form a three-spectra dataset. This dataset includes 8,675 image triples from 50 vehicles for the training set and 8,575 triples from the other 50 vehicles for the test gallery set. From the test gallery, 1,715 samples are selected to form the query set.

Evaluation Protocols. We use the Cumulative Matching Characteristic (CMC) curve and mean Average Precision (mAP) as evaluation metrics. In MSVR310 [19], as in previous work, we adopt a strict evaluation protocol, which filters out samples with the same identity and time span in the matching results using time labels to avoid easy match-

TABLE III

COMPARISON PERFORMANCES WITH THE STATE-OF-THE-ART METHODS ON MSVR310. THE BEST AND SECOND-BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE, RESPECTIVELY.

	Methods	Venue	mAP	R-1	R-5	R-10
Single	DMML [63]	ICCV19	19.1	31.1	48.7	57.2
	Circle Loss [64]	CVPR20	22.7	34.2	52.1	57.2
	PCB [16]	ECCV18	23.2	42.9	58.0	64.6
	BoT [5]	CVPRW19	23.5	38.4	56.8	64.8
	MGN [15]	MM18	26.2	44.3	59.0	66.8
	HRCN [65]	ICCV21	23.4	44.2	66.0	73.9
Multi	OSNet [6]	ICCV19	28.7	44.8	66.2	73.1
	AGW [66]	TPAMI21	28.9	46.9	64.3	72.3
	TransReID [3]	ICCV21	26.9	43.5	62.4	70.7
	HAMNet [10]	AAAI20	27.1	42.3	61.6	69.5
	PFNet [11]	AAAI21	23.5	37.4	57.0	67.3
	PFD [28]	AAAI22	23.0	39.9	56.3	64.0
Multi	FED [67]	CVPR22	21.7	37.4	58.9	67.3
	IEEE [54]	AAAI22	21.0	41.0	57.7	65.0
	CCNet [19]	INFS23	36.4	55.2	72.4	79.7
	TOP-ReID [21]	AAAI24	35.9	44.6	-	-
	EDITOR [22]	CVPR24	39.0	49.3	-	-
	CLIP-ReID [14]	AAAI23	<u>52.6</u>	<u>71.1</u>	<u>85.1</u>	<u>89.0</u>
	ICPL	Ours	56.9	77.7	87.6	91.5

TABLE IV

COMPARISON PERFORMANCES ON RGBNT100 AND RGBN300. THE BEST AND SECOND-BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE, RESPECTIVELY. HERE THE SUPERSCRIPT * REPRESENTS THE RESULTS ARE REPRODUCED BY US.

Methods	Venue	RGBNT100		RGBN300		
		mAP	R-1	mAP	R-1	
Single	PCB [16]	ECCV18	57.2	83.5	57.7	82.0
	MGN [15]	MM18	58.1	83.1	60.5	83.7
	ADB [68]	ICCV19	60.4	85.1	58.9	83.1
	OSNet [6]	ICCV19	75.0	95.6	-	-
	TransReID [3]	ICCV21	75.6	92.9	79.0*	92.5*
	HAMNet [10]	AAAI20	64.1	84.7	61.9	84.0
Multi	DANet [69]	ICPR22	-	-	71.0	89.9
	GAFNet [70]	ICSP22	74.4	93.4	72.7	91.9
	GraFT [71]	ARXIV23	76.6	94.3	75.1	92.1
	GPFNet [72]	TITS23	75.0	94.5	73.3	90.0
	PHT [73]	SENSORS23	79.9	92.7	79.3	93.7
	UniCat [62]	NIPSW23	81.3	97.5	80.2	92.9
	TOP-ReID [21]	AAAI24	81.2	96.4	77.7*	91.9*
	EDITOR [22]	CVPR24	82.1	96.4	75.2*	90.0*
	CLIP-ReID [14]	AAAI23	<u>87.0</u>	<u>96.9</u>	<u>85.5</u>	<u>94.9</u>
	ICPL	Ours	87.0	98.6	87.0	96.3

ing. In RGBNT201 [11], Market-MM [54], RGBNT100 and RGBN300 [10], we follow the commonly used evaluation protocol as in previous works.

B. Implementation Details

We resize the images of each spectra to 256x128 (128x256 to maintain the aspect ratio of vehicle) and use random horizontal flipping, padding with 10 pixels, random cropping, and random erasing [74] as feature enhancement strategies. It is worth noting that data augmentation is not used during the prototype aggregation stage. We select ViT-B/16 as our visual backbone, freezing most parameters in the visual and text branches, while keeping a few learnable parameters, includ-

TABLE V

ABLATION OF DIFFERENT COMPONENTS ON RGBNT201, MSVR310, AND RGBNT100. WE USE A TRIPLET-STEAM CLIP VISUAL ENCODER AS THE BASELINE. OUR COMPONENT SPLITS INCLUDE THE SPECTRAL IDENTITY CONDITION (SIC) AND THE ALIGNMENT LOOP (AL) MODULES IN THE MULTI-SPECTRAL IDENTITY CONDITION MODULE (MS-IC), AS WELL AS THE MULTI-SPECTRAL ADAPTER MODULE (MS-A).

	MS-IC		MS-A	RGBNT201				MSVR310				RGBNT100			
	SIC	AL		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	85.3	96.6	97.2	97.6
(b)	✓	✗	✗	72.0	73.4	80.3	85.6	52.3	71.1	84.8	89.2	86.6	96.7	97.2	97.6
(c)	✓	✓	✗	73.0	75.1	82.5	86.8	55.9	76.0	86.8	90.5	86.0	97.0	97.7	98.1
(d)	✗	✗	✓	72.9	73.7	81.9	87.9	55.6	77.0	87.5	91.5	85.9	98.2	98.7	98.8
(e)	✓	✗	✓	72.7	75.1	82.8	87.1	56.7	77.0	86.5	90.4	85.8	98.3	98.8	98.9
(f)	✓	✓	✓	75.1	77.4	84.2	87.9	56.9	77.7	87.6	91.5	87.0	98.6	99.0	99.0

ing the learnable text prompt “[X]₁, [X]₂, …, [X]_M, [CLS]”, multi-spectral adapter, classifier, the last batch normalization layer and image-text projection layer in the visual encoder. The batch size is set to 64, where 16 identities are randomly selected from each small batch, 4 samples are randomly selected from each identity, and each sample includes 3 images with different modalities. We employ the Adam optimizer with a weight decay of 0.0005, momentum of 0.9, and an initial learning rate of $3.5e-4$. The training lasts for 120 epochs, and a warmup strategy is used in the first 10 epochs. Linear decay of 0.1 is applied at 30 and 50 epochs, with decay rates of $3.5e-5$ and $3.5e-6$. All our experiments are conducted on one NVIDIA RTX 4090 using Pytorch.

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

Comparison on RGBNT201 and Market-MM. Table I and Table II report our performance on RGBNT201 [11] and Market-MM [54] datasets. Clearly, our method has significant advantages over traditional methods and achieves the best performance. We extend single-modal CLIP-ReID [14] to triplet-stream one for fair comparison by replicating the visual encoder three times for three spectra. Specifically, the triplet-stream CLIP-ReID has achieved comparable performance to existing Transformer-based models, such as TOP-ReID [21], EDITOR [22], HTT [20], etc. However, the two-stage alignment method does not consider the collaborative alignment of spectral and learnable text prompt, preventing text prompt from learning the new spectral data. Therefore, when we adopt the online identity-conditional prompt learning (ICPL), the model can build a mutual alignment loop between learnable text prompt and spectral visual encoder, which enables our model to achieve **4.0%/5.6%** and **2.6%/1.0%** mAP/Rank-1 performance improvement on both datasets compared with CLIP-ReID [14].

Comparison on MSVR310. As shown in Table III, most methods encounter a performance drop when facing the viewpoint variation and long time span challenges in MSVR310 [19] dataset. The robust semantic generalization capability of CLIP enables the triplet-stream CLIP-ReID to outperform methods using multi-spectral feature fusion by a large margin. However, our approach further improves performance, achieving a **4.3%/6.6%** mAP/Rank-1 enhancement over CLIP-ReID [14]. This demonstrates that our model can effectively focus

on viewpoint-invariant identity semantic features through collaborative training with identity semantic prompt.

Comparison on RGBNT100 and RGBN300. As shown in Table IV, our method achieves the best performance on both datasets. Notably, on the RGBNT100 [10] dataset, the mAP of ICPL is comparable to that of CLIP-ReID [14], with a **1.7%** improvement in Rank-1. This could be explained by the large number of repeated samples from the same viewpoints in the RGBNT100 [10] dataset, which dilutes the model performance in terms of mAP. However, the more challenging Rank-1 metric reflects the better matching ability of ICPL. On the more complex RGBN300 [10] dataset, while the single-modal TransReID [3] model has shown relative effectiveness, both TOP-ReID [21] and EDITOR [22] fail to achieve the expected performance. In contrast, our method, with end-to-end prompt learning of vehicle textual semantics, further boosts **1.5%/1.4%** in mAP/Rank-1 performance over CLIP-ReID [14], highlighting the applicability of our approach to vehicle tasks.

D. Ablation Study

In this section, we conduct a series of ablation experiments on RGBNT201 [11], MSVR310 [19], and RGBNT100 [10] to verify the effectiveness of each component in our proposed framework. This includes the Spectral Identity Condition (SIC) and the Alignment Loop (AL) within the Multi-Spectral Identity Condition module (MS-IC), and the Multi-Spectral Adapter module (MS-A).

Ablation Study of Individual Components. In order to achieve three spectra inputs, we replicate the CLIP visual encoder three times to form a triplet-stream network, and use it as our baseline in Table V(a). We maintain the data augmentation strategy during training and use \mathcal{L}_{id} and \mathcal{L}_{tri} as the loss functions. The baseline performance on two datasets highlights the strong representation capability of CLIP. When the visual branch is optimized using only the SIC module, the model achieves improvements of **1.0%/1.9%**, **3.2%/5.6%**, and **1.3%/0.1%** in mAP/Rank-1 performance across three datasets, as shown in Table V(b). Prototypes promote spectral feature aggregation using identity anchors, effectively representing sample features within identities. As shown in Table V(c), introducing a learnable text prompt with the AL module leads to **2.0%/3.6%**, **6.8%/10.5%**,

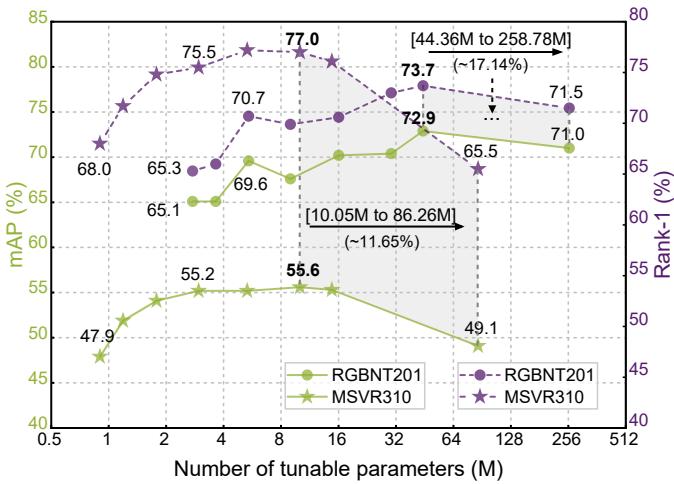


Fig. 5. The performance trend on mAP and Rank-1 as the number of tunable parameters grows.

and **0.7%/0.4%** mAP/Rank-1 performance gains across three datasets. This improvement is attributed to the robust identity prototype significantly reducing the difficulty of text prompt learning in the absence of real text descriptions, compared with randomly aligned prompts with instance samples. As shown in Table V(d), using the **MS-A** module results in mAP/Rank-1 performance gains of **1.9%/2.2%, 6.5%/11.5%,** and **0.6%/1.6%** cross three datasets. Notably, the performance improvement on MSVR310 dataset is more significant than on the person dataset. This is due to the MSVR310 dataset encompassing more diverse data in viewpoints and time spans, which benefited our adapter learning with diverse data. As shown in Table V(e), a simple combination of the SIC and MS-A modules leads to a slight performance fluctuation. This phenomenon may be due to the lightweight adapter only relying on the aggregated prototypes, which disrupts the pre-trained feature space distribution, thereby reducing the CLIP generalization performance on unseen test data. Finally, as shown in Table V(f), using all components leads to optimal performance through the mutual optimization of the text prompt and visual encoder, with mAP/Rank-1 performance improvements of **4.1%/5.9%, 7.8%/12.2%,** and **1.7%/2.0%** cross three datasets, respectively.

E. Further Analysis

Different Variants of Multi-Spectral Identity Condition. To verify the effectiveness of the MS-IC module in online prompt learning, we design different versions of it for comparison. As shown in Table VI (a), we replicate the CLIP visual encoder three times as the triplet-stream baseline, which can achieve performance comparable to SOTA methods without the text encoder. As shown in Table VI (b), We first apply image-to-text loss \mathcal{L}_{i2t} to the baseline and randomly initialize a text prompt for each identity. However, the randomly initialized text prompt lack actual semantics, they cannot effectively assist the model learning, which leads to performance degradation.

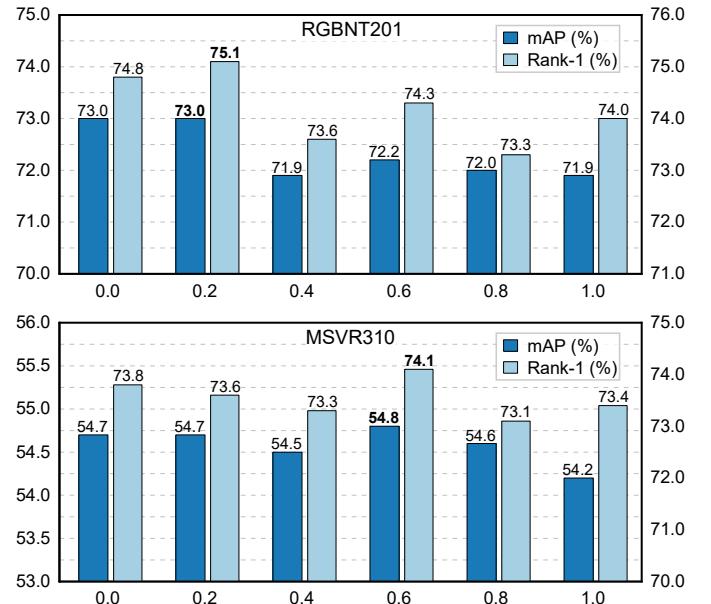


Fig. 6. Hyper-parameter analysis on prototype factor α .

TABLE VI
DIFFERENT VARIANTS OF MULTI-SPECTRAL IDENTITY CONDITIONAL ON RGBNT201 AND MSVR310.

Method	RGBNT201		MSVR310	
	mAP	R-1	mAP	R-1
(a) Baseline	71.0	71.5	49.1	65.5
(b) + \mathcal{L}_{i2t}	64.3	65.7	46.2	66.2
(c) + $\mathcal{L}_{i2t} + \mathcal{L}_{t2i}$	68.0	65.6	52.2	70.4
(d) + MS-IC ($\mathcal{L}_{i2p} + \mathcal{L}_{prompt}$)	73.0	75.1	55.9	76.0

For the second variant in Table VI (c), we use image-to-text loss \mathcal{L}_{i2t} and text-to-image loss \mathcal{L}_{t2i} to align the randomly initialized text prompt with spectral samples. Observing the experimental results, after aligning the text prompt with the spectral modalities, the model has a significant performance improvement on the MSVR310 dataset [19]. This indicates that identity-related semantics can be learned through online alignment of text prompt, and this identity semantics is effective in multi-spectral ReID task. However, on the RGBNT201 dataset [11], the model is still lower than the baseline, indicating that simple online alignment is relatively suboptimal. Finally, by replacing the above alignment loss with image-to-prototype loss \mathcal{L}_{i2p} and prompt loss \mathcal{L}_{prompt} in Table VI (d), our method achieves superior performance improvements on both datasets. This proves that online identity-conditional prompt learning can effectively transfer the image-text alignment capability of CLIP to multi-spectral ReID task.

Effectiveness on Multi-Spectral Adapter. The low-rank adaption method significantly reduces model training parameters, while effectively addressing discrepancies between the pre-trained model and multi-spectral ReID task. By changing the intermediate hidden dimension of adapter, choose from {16, 32, 64, 128, 256, 512, 768}, and compare with the full fine-tuning method. Notably, to further reduce learnable parameters, we validate that different spectra share one

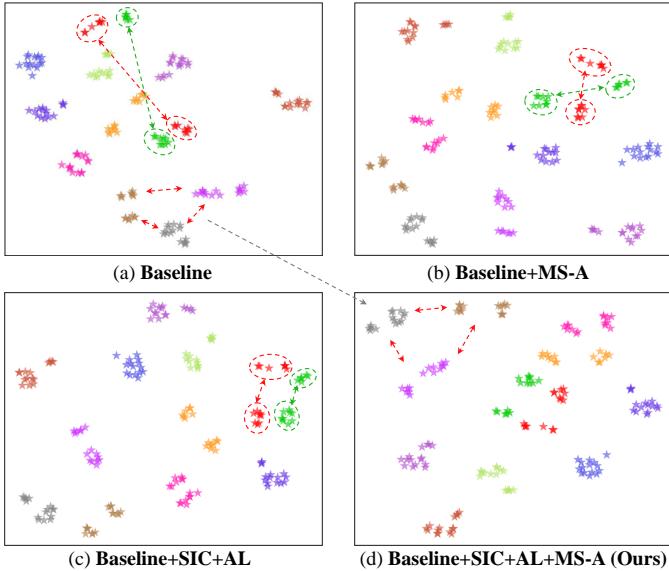


Fig. 7. Visualization results of the (a) Baseline, (b) Baseline + MS-A, (c) Baseline + SIC + AL, and (d) Baseline + SIC + AL + MS-A (Ours). Better view with colors and zooming in.

learnable adapter on the MSVR310 dataset [19], as shown in Fig. 5, which also achieves optimal performance with only 10.05M parameters ($\tilde{d}=512$), about 11.65% of the full fine-tuning parameters (86.26M). Even when each spectral modality has an individual adapter on RGBNT201, we achieve notable performance with only 44.36M parameters ($\tilde{d}=768$), about 17.14% of the full fine-tuning parameters (258.78M). The above results demonstrate the effectiveness of our multi-spectral adapter in alleviating the discrepancies between multi-spectral data and pre-training data. Notably, our results indicate that even with the intermediate dimension reduced to 128, the model still performs well on the vehicle dataset. This suggests that the MSVR310 dataset [19] is a curated dataset to provide diverse vehicle samples and rich camera viewpoints while omitting the most redundant vehicle samples, making the dataset sufficiently refined.

Hyper-parameters Analysis. During the identity-conditional alignment process, the prototype always plays a pivotal role in maintaining and propagating global sample features to text prompt. As the updating factor increases, the prototypes gradually coagulate from dynamic to static anchor. As shown in Fig. 6, when the factor is less than 1.0, the features of freshly optimized samples are always updated synchronously with the prototypes, enabling the learnable text prompt to adapt to the current training task promptly. In contrast, when the factor is fixed at 1.0, the text prompt cannot dynamically align immediately, resulting in the model falling into sub-optimal solutions. This result indicates the significance of learnable text prompt in dynamically aligning to the learned spectra-specific features during the training process.

As shown in Table VII, we further explore the influence of the multi-spectral adapter on the frozen visual encoder. We fuse the newly learned spectral features of each layer with the original visual features by adding them after adjusting the scale

TABLE VII
HYPER-PARAMETER ANALYSIS ON SCALING FACTOR s OF ADAPTER.

factor	RGBNT201		MSVR310	
	mAP	R-1	mAP	R-1
0.1	63.3	62.0	52.3	73.6
0.2	66.7	67.2	55.3	76.8
0.3	68.6	68.3	55.6	77.0
0.4	70.2	69.5	55.3	77.2
0.5	72.9	73.7	54.4	73.9
0.6	70.6	70.1	55.1	75.8
0.7	70.5	72.2	55.0	76.3
0.8	71.5	73.4	54.1	73.4
0.9	70.3	72.1	54.0	74.6
1.0	71.2	74.2	53.1	73.3

TABLE VIII
COMPARISON OF DIFFERENT LEARNABLE PROMPT NUMBER M.

num	RGBNT201		MSVR310	
	mAP	R-1	mAP	R-1
1	72.7	76.6	56.2	75.0
2	74.0	77.2	55.3	75.8
4	75.1	77.4	56.9	77.7
8	74.1	77.5	56.6	77.0
16	73.8	74.9	55.1	75.3
32	73.0	74.6	55.2	74.1

factor. When the scale factor is in the middle, the model can achieve a balanced combination of the newly learned spectral features and the original features. However, when the scale factor is too small (≤ 0.2), the model encounters a significant performance decline due to the difficulty in effectively learning new features. On the contrary, when the scaling factor is too large (≥ 0.9), excessive loss of original features leads to performance fluctuation.

Comparison of Different Learnable Prompt Number M.

We analyze the number of learnable prompt tokens, on the RGBNT201 [11] and MSVR310 [19] datasets. As shown in Table VIII, an appropriate token number M helps the model achieve optimal performance. If M is too small, the limited capacity for semantic learning is insufficient to capture the semantic information from the spectra, leading to a performance decline. On the other hand, when M is too large, the redundant prompts are hard to optimize, causing the model to experience incorrect semantic guidance that hinders its performance.

Balancing and Trade-offs of Loss Weight Factors. To further insight into the mutual optimization process between text prompt and visual encoder, we analyze the loss function weight factors on the RGBNT201 [11] and MSVR310 [19] datasets. As shown in Table IX, $\mathcal{L}_{t2p} + \mathcal{L}_{p2t}$ and \mathcal{L}_{i2p} perform best with the default weight factor of 1.0 or slightly reduced to 0.9. However, the \mathcal{L}_{i2t} requires a lower weight factor of 0.1, which aligns with our expectations. For the randomly initialized semantic prompt, CLIP-ReID employs a pre-alignment stage to ensure the prompt captures identity semantics. In contrast, ICPL optimizes both the semantic prompt and spectral encoder together. A lower λ_1 factor prevents semantic noise from disrupting the spectral encoder during the early alignment process. Meanwhile, setting each weight to 0 degrades performance, indicating their necessity in optimization.

TABLE IX

ANALYSIS OF LOSS WEIGHT FACTORS. λ_1 FOR THE IMAGE-TO-TEXT ALIGNMENT LOSS \mathcal{L}_{i2t} . λ_2 FOR THE SYMMETRICAL PROTOTYPE-TEXT ALIGNMENT LOSS $\mathcal{L}_{t2p} + \mathcal{L}_{p2t}$. AND λ_3 FOR THE IMAGE-TO-PROTOTYPE ALIGNMENT LOSS \mathcal{L}_{i2p} . THE **BOLDED** WEIGHTS REPRESENT THE DEFAULT SETTINGS.

RGBNT201								MSVR310									
λ_1	mAP	R-1	λ_2	mAP	R-1	λ_3	mAP	R-1	λ_1	mAP	R-1	λ_2	mAP	R-1	λ_3	mAP	R-1
0.0	73.8	76.7	0.0	72.8	75.6	0.0	71.2	73.1	0.0	55.0	73.6	0.0	56.2	76.1	0.0	55.0	75.0
0.1	75.1	77.4	0.1	73.7	76.6	0.1	72.3	74.5	0.1	56.9	77.7	0.1	56.3	75.5	0.1	55.5	75.0
0.2	73.8	75.2	0.2	73.3	76.3	0.2	73.0	74.8	0.2	55.9	74.3	0.2	57.0	76.5	0.2	55.4	75.0
0.3	73.7	76.6	0.3	74.2	77.4	0.3	73.0	75.2	0.3	55.3	74.6	0.3	56.6	76.6	0.3	56.4	76.0
0.4	73.0	75.2	0.4	74.1	77.0	0.4	73.0	74.9	0.4	55.6	75.5	0.4	56.9	76.1	0.4	56.1	76.3
0.5	72.9	75.0	0.5	73.5	76.6	0.5	73.5	75.8	0.5	55.7	75.5	0.5	56.7	75.6	0.5	56.7	76.0
0.6	72.8	75.0	0.6	73.8	76.0	0.6	73.3	76.3	0.6	56.2	75.0	0.6	56.1	75.1	0.6	56.2	77.2
0.7	72.8	74.3	0.7	73.8	75.5	0.7	73.8	76.0	0.7	55.7	75.3	0.7	56.1	76.0	0.7	56.5	75.1
0.8	72.6	73.9	0.8	73.5	75.8	0.8	74.4	76.4	0.8	56.6	75.6	0.8	56.4	76.1	0.8	56.9	76.6
0.9	72.8	73.3	0.9	73.7	76.4	0.9	75.1	77.4	0.9	56.0	75.8	0.9	56.7	75.5	0.9	56.9	77.7
1.0	73.0	74.0	1.0	75.1	77.4	1.0	74.2	76.1	1.0	56.2	76.3	1.0	56.9	77.7	1.0	56.8	75.8
2.0	69.2	70.9	2.0	73.3	75.4	2.0	71.0	73.7	2.0	54.5	75.0	2.0	56.5	76.8	2.0	56.3	76.8
5.0	61.3	62.4	5.0	72.6	75.1	5.0	69.9	71.3	5.0	52.0	72.4	5.0	56.2	76.6	5.0	54.0	74.1

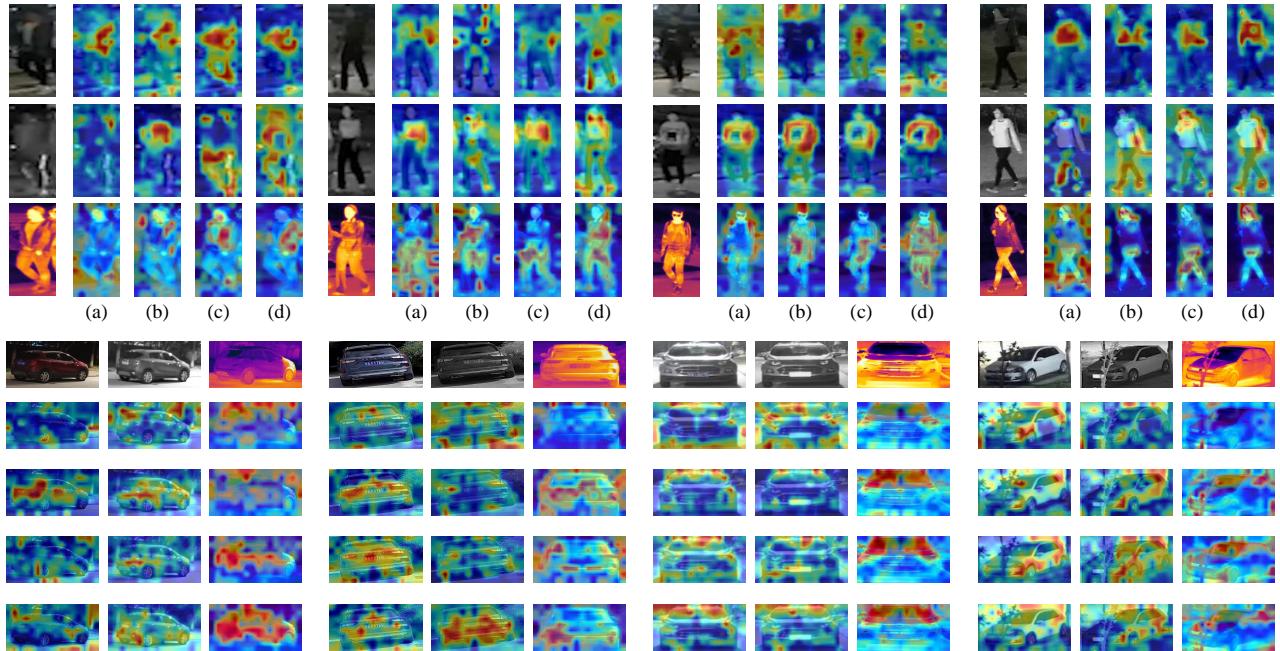


Fig. 8. Visualization results of the (a) Baseline, (b) Baseline + MS-A, (c) Baseline + SIC + AL, and (d) Baseline + SIC + AL + MS-A (Ours), drawn by Grad-CAM [75]. Better view with colors and zooming in.

However, when these weights are increased beyond 2.0 to 5.0, we observe a significant performance drop. This suggests that excessively large weights skew the optimization process, potentially leading to optimization imbalances and limiting model generalization. Based on the above experimental results, setting weights within a moderate range (e.g., around 0.1 to 1.0) effectively balances the optimization objectives within the alignment loop, achieving optimal performance. In future work, we will explore more advanced learnable loss tuning methods [7] to more flexibly and cleverly optimize different training objectives in multi-spectral ReID tasks.

Computational Efficiency Analysis. As shown in Table X, to evaluate the computational efficiency of our proposed framework, we compare ICPL with CLIP-ReID [14] across three datasets of varying scales. In most scenarios, ICPL exhibits faster computational efficiency across different training settings. While the training time for ICPL is higher on

the RGBNT201 dataset, it still achieves acceptable training efficiency due to the faster convergence in shorter periods. For example, in the 20-epoch setting, the training time is only 1.08 times that of CLIP-ReID [14]. On the larger scale RGBNT100 dataset, ICPL shows a clear advantage. In the 20-epoch short training setting, the training time is only 0.78 times that of CLIP-ReID [14], which is crucial for large-scale datasets.

F. Visualization

Feature Distribution. To evaluate the impact of different components in ICPL on model performance, we utilize the T-SNE [76] to visualize the sample distribution, providing in-depth insights into the approach. As shown in Fig. 7 (b), the lightweight MS-A module reduces the distance between challenging samples, enhancing the pre-trained model to better adapt to multi-spectral data. In Fig. 7 (c), the semantic

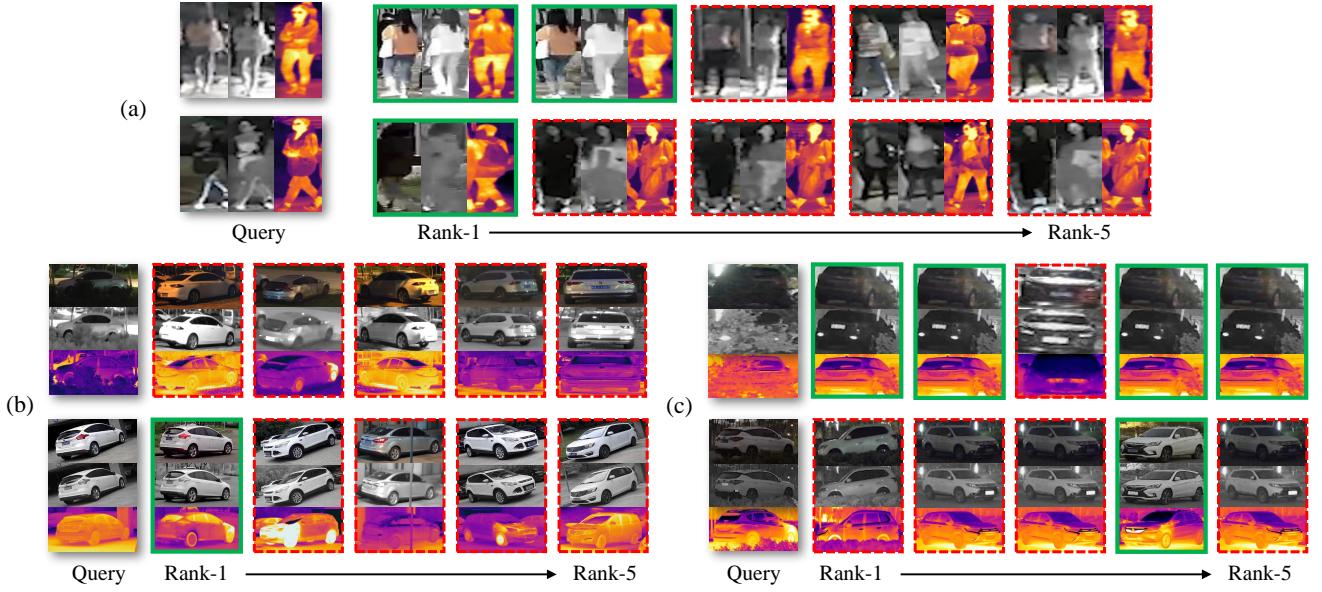


Fig. 9. Visualization of top-5 failure cases on (a) RGBNT201 [11], (b) MSVR310 [19], and (c) RGBNT100 [10] datasets.

TABLE X
COMPARISON OF THE COMPUTATION COST ON MSVR310, RGBNT201,
AND RGBNT100 DATASETS. ALL MODELS ARE EVALUATED ON A SINGLE
4090 GPU.

Method	Training Times ↓					
	20 epoch	40 epoch	60 epoch			
(1) MSVR310 (Images: 2678)						
CLIP-ReID [14]	506s	1x	861s	1x	1212s	1x
ICPL(Ours)	419s	0.83x	820s	0.95x	1220s	1.00x
(2) RGBNT201 (Images: 5623)						
CLIP-ReID [14]	754s	1x	1273s	1x	1769s	1x
ICPL(Ours)	818s	1.08x	1625s	1.28x	2439s	1.38x
(3) RGBNT100 (Images: 18965)						
CLIP-ReID [14]	2389s	1x	4193s	1x	5893s	1x
ICPL(Ours)	1856s	0.78x	3725s	0.89x	5585s	0.95x

guidance from SIC and AL further compacts the sample distribution within each identity. Finally, as depicted in Fig. 7 (d), the complete ICPL expands the separation between samples from different identities, improving inter-class separability.

Discriminative Attention Maps. To further validate the effectiveness of our proposed components, we use Grad-CAM [75] to visualize the features of each spectral modality. As shown in Fig. 8 (a), the simply fine-tuned visual encoder cannot effectively focus on the discriminative regions of the object. In scenarios such as nighttime, the model only focuses on the background region and struggles to generate effective feature responses on infrared spectra. This indicates that simple fully fine-tuning cannot effectively transfer the pre-trained visual encoder to multi-spectral datasets with significant stylistic discrepancies. As shown in Fig. 8 (b) to (d), introducing the MS-A module and online text prompt learning with the SIC and AL modules significantly reduces feature response on background areas, while greatly enhancing response on objects in the infrared-spectral modality. Based on the above observations, our identity-conditional prompt learning method

effectively promotes the model to focus on regions with rich identity semantics for the same identity in different spectral modalities through our mutual optimization online alignment strategy.

Failure Cases Analysis. To further analyze the retrieval performance of ICPL in real-world scenarios, we visualize the failure cases across three datasets: RGBNT201 [11], MSVR310 [19], and RGBNT100 [10]. Thanks to the image-text alignment capability of prompt learning, ICPL significantly improves the model performance. However, as shown in Fig. 9, extreme lighting degradation, background occlusion, and low-quality noise within the spectral still pose challenges for ICPL in focusing on object semantics in such scenarios. In the future, we plan to explore more robust and fine-grained multi-spectral prompt learning methods to address these challenges.

V. CONCLUSION

In this paper, we introduce a novel prompt learning framework that harnesses the cross-modal alignment capabilities of the vision-language pre-training model for the multi-spectral ReID task. First, our framework enables online prompt learning for multi-spectral ReID, using learnable text prompt as identity-level spectral semantic center to bridge the identity semantics of different spectra. Second, we propose the multi-spectral identity condition module, which establishes a mutual alignment loop between the text prompt and spectral visual encoder, making the text prompt well-aligned even without concrete spectral text descriptions. Finally, we propose the multi-spectral adapter module, utilizing a lightweight adapter to optimize the frozen visual encoder, enabling adaptation to new multi-spectral data while preserving the pre-trained image-text alignment distribution of CLIP. Extensive experiments on person and vehicle datasets demonstrate the effectiveness of our method. In future work, we will explore the fine-grained text prompt to fully exploit the cross-modal alignment capability of the visual-language pre-trained model.

REFERENCES

- [1] 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.
- [2] 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.
- [3] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transeid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 993–15 002.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] X. Yuan, X. Xu, Z. Wang, K. Zhang, W. Liu, and R. Hu, "Searching parameterized retrieval & verification loss for re-identification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 3, pp. 560–574, 2023.
- [8] A. Lu, C. Qian, C. Li, J. Tang, and L. Wang, "Duality-gated mutual condition network for RGBT tracking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 4118–4131, 2025.
- [9] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, "RGBT tracking via multi-adaptor network with hierarchical divergence loss," *IEEE Transactions on Image Processing*, vol. 30, pp. 5613–5625, 2021.
- [10] 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.
- [11] 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.
- [12] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia, and C. Lin, "Grayscale enhancement colorization network for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1418–1430, 2022.
- [13] A. Lu, C. Li, J. Zhao, J. Tang, and B. Luo, "Modality-missing rgbt tracking: Invertible prompt learning and high-quality benchmarks," *International Journal of Computer Vision*, pp. 1–21, 2024.
- [14] 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.
- [15] 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.
- [16] 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*, vol. 11208, 2018, pp. 501–518.
- [17] 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*, vol. 26, pp. 406–419, 2024.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] Z. Hu, B. Yang, and M. Ye, "Empowering visible-infrared person re-identification with large foundation models," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [24] 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.
- [25] C. Cui, S. Huang, W. Song, P. Ding, M. Zhang, and D. Wang, "Profod: Prompt-guided feature disentangling for occluded person re-identification," in *Proceedings of the ACM International Conference on Multimedia*, 2024, pp. 1583–1592.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, 2021.
- [32] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 966–11 976.
- [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] 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.
- [35] 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. 16 795–16 804.
- [36] 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.
- [37] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko, "Visda: The visual domain adaptation challenge," *arXiv preprint arXiv:1710.06924*, 2017.
- [38] J. A. Samadhi, H. Gani, N. Hussein, M. U. Khattak, M. Naseer, F. S. Khan, and S. H. Khan, "Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization," in *Proceedings of the Conference on Neural Information Processing Systems*, 2023.
- [39] S. Song, Z. Miao, H. Yu, J. Fang, K. Zheng, C. Ma, and S. Wang, "Deep domain adaptation based multi-spectral salient object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 128–140, 2022.
- [40] H. Li, Y. Li, M. Yang, P. Hu, D. Peng, and X. Peng, "Incomplete multi-view clustering via prototype-based imputation," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2023, pp. 3911–3919.
- [41] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 692–12 702.

- [42] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8228–8237.
- [43] J. Liu, Z. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7202–7211.
- [44] 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.
- [45] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 5390–5399.
- [46] 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.
- [47] T. Liang, Y. Jin, W. Liu, and Y. Li, "Cross-modality transformer with modality mining for visible-infrared person re-identification," *IEEE Transactions on Multimedia*, vol. 25, pp. 8432–8444, 2023.
- [48] Z. Huang, J. Liu, L. Li, K. Zheng, and Z. Zha, "Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2022, pp. 1034–1042.
- [49] J. Liu, Z. Zha, R. Hong, M. Wang, and Y. Zhang, "Deep adversarial graph attention convolution network for text-based person search," in *Proceedings of the ACM International Conference on Multimedia*, 2019, pp. 665–673.
- [50] Z. Miao, H. Liu, W. Shi, W. Xu, and H. Ye, "Modality-aware style adaptation for rgb-infrared person re-identification," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021, pp. 916–922.
- [51] Y. Feng, J. Yu, F. Chen, Y. Ji, F. Wu, S. Liu, and X. Jing, "Visible-infrared person re-identification via cross-modality interaction transformer," *IEEE Transactions on Multimedia*, vol. 25, pp. 7647–7659, 2023.
- [52] Y. Zhang and H. Wang, "Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2153–2162.
- [53] Y. Ling, Z. Luo, Y. Lin, and S. Li, "A multi-constraint similarity learning with adaptive weighting for visible-thermal person re-identification," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021, pp. 845–851.
- [54] 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.
- [55] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the International conference on machine learning*, vol. 162, 2022, pp. 12888–12900.
- [56] M. Jia, L. Tang, B. Chen, C. Cardie, S. J. Belongie, B. Hariharan, and S. Lim, "Visual prompt tuning," in *Proceedings of the European Conference on Computer Vision*, vol. 13693, 2022, pp. 709–727.
- [57] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adapt-former: Adapting vision transformers for scalable visual recognition," in *Proceedings of the Conference on Neural Information Processing Systems*, vol. 35, 2022, pp. 16664–16678.
- [58] 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. 5409–5418.
- [59] 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.
- [60] 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.
- [61] 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.
- [62] 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.
- [63] G. Chen, T. Zhang, J. Lu, and J. Zhou, "Deep meta metric learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9546–9555.
- [64] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6397–6406.
- [65] J. Zhao, Y. Zhao, J. Li, K. Yan, and Y. Tian, "Heterogeneous relational complement for vehicle re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 205–214.
- [66] 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.
- [67] 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.
- [68] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8350–8360.
- [69] E. Kamenou, J. M. del Rincón, P. Miller, and P. Devlin-Hill, "Closing the domain gap for cross-modal visible-infrared vehicle re-identification," in *Proceedings of the International Conference on Pattern Recognition*. IEEE, 2022, pp. 2728–2734.
- [70] J. Guo, X. Zhang, Z. Liu, and Y. Wang, "Generative and attentive fusion for multi-spectral vehicle re-identification," in *Proceedings of the International Conference on Intelligent Computing and Signal Processing*, 2022, pp. 1565–1572.
- [71] H. Yin, J. Li, E. Schiller, L. McDermott, and D. Cummings, "Graft: Gradual fusion transformer for multimodal re-identification," *arXiv preprint arXiv:2310.16856*, 2023.
- [72] Q. He, Z. Lu, Z. Wang, and H. Hu, "Graph-based progressive fusion network for multi-modality vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 12431–12447, 2023.
- [73] W. Pan, L. Huang, J. Liang, L. Hong, and J. Zhu, "Progressively hybrid transformer for multi-modal vehicle re-identification," *Sensors*, vol. 23, no. 9, p. 4206, 2023.
- [74] 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. 13001–13008.
- [75] 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.
- [76] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, 2008.



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.



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*.



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-Doctor 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.