Multi-Granularity and Multi-Modal Prompt Learning for Person Re-Identification

Hao Tong
University of Science and Technology of China
Hefei, Anhui, China
haotong@mail.ustc.edu.cn

Jiawei Liu*
University of Science and Technology of China
Hefei, Anhui, China
jwliu6@ustc.edu.cn

Yong Wu
China Merchants Bank
Shenzhen, Guangdong, China
wuyong139@cmbchina.com

Guozhi Zhao
China Merchants Bank
Shenzhen, Guangdong, China
gzzhao@cmbchina.com

Fanrui Zhang
University of Science and Technology of China
Hefei, Anhui, China
zfr888@mail.ustc.edu.cn

Zheng-Jun Zha
University of Science and Technology of China
Hefei, Anhui, China
zhazj@ustc.edu.cn

Abstract

Pre-trained vision-language models, such as CLIP, are driving advancements in person re-identification by mining semantic information. Current approaches utilize globally learnable textual prompts to generate coarse-level, holistic yet ambiguous descriptions of individuals, which are then served as constraints for the fine-tuning of CLIP to learn visual representations of the pedestrians. However, relying exclusively on this type of prompt learning for the text encoder of CLIP overlooks the crucial fine-grained details of individuals and fails to model the necessary downstream adaptation capacity for the image encoder of CLIP. To address these limitations, we propose a novel Multi-Granularity and Multi-Modal Prompt Learning (MMPL) for person re-identification to fully unleash the substantial potential inherent in CLIP for acquiring discriminative representations. The MMPL encompasses a two-stage training procedure. In the first training stage, MMPL meticulously orchestrates the Hierarchical Prompt Learning (HPL) to refine crucial and distinctive information from hierarchical patch-level visual features. Aligning textual prompts with these subtle visual cues across diverse granularities, this process establishes patch-to-token level correspondences, ultimately yielding the creation of high-fidelity multi-granularity textual prompts. In the second training stage, MMPL integrates Collaborative Prompt Learning (CPL), generating supplementary visual prompts and fostering multi-modal interactive learning to aid CLIP's image encoder in narrowing the semantic gap between modalities, leveraging CLIP's extensive multi-modal knowledge to enhance feature representation. Comprehensive experimental evaluation across four widely recognized person re-identification benchmarks substantiates the effectiveness of our MMPL.

***Keywords:** Person re-identification, Prompt learning, Multi-granularity, Vision-language model.*

1. Introduction

Person re-identification (Person ReID) entails retrieving a particular individual of interest across diverse camera views within a vast gallery database, which involves addressing the issues like cluttered backgrounds [37], illumination variations [52], pose differences [30] and occlusions [21]. The field has garnered considerable interest from both academic and industrial sectors owing to its vital role in enhancing intelligent video surveillance systems.

The predominant approaches in person ReID methods hinge on the construction and training of convolutional neural networks (CNNs). Employing CNNs, these methods effectively translate pedestrian imagery into the embedding space under the guidance of typical metric learning loss functions [29]. The overarching goal is to diminish the distance among feature vectors corresponding to identical individuals and concurrently amplify the separation between

*Corresponding author.

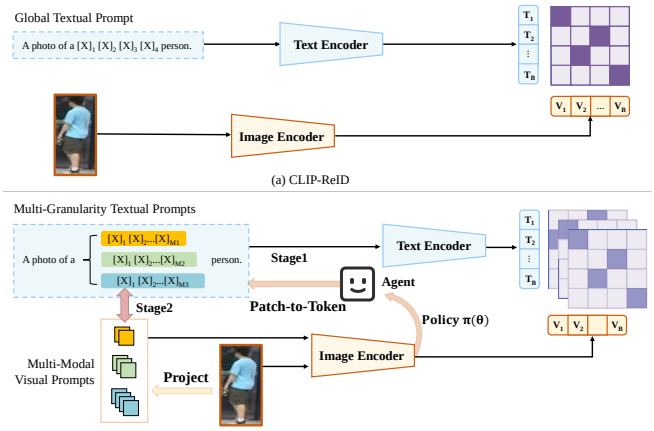


Figure 1. Contrasting CLIP-ReID with our MMPL. (a) CLIP-ReID focuses on fine-tuning a single global learnable textual prompt to guide the image encoder. (b) In stage 1, our MMPL meticulously fine-tunes multi-granularity textual prompts through the ingenious application of information selector agent and patch-to-token alignment, enabling accurate and comprehensive person descriptions. Stage 2 further enhances this process by integrating visual prompts, establishing a multi-modal joint prompt learning, and enhancing the image encoder in utilizing rich multi-modal semantic knowledge in CLIP to learn discriminative representations.

vectors of different identities. However, CNN-based methods often focus on less relevant regions due to the Gaussian distribution of effective receptive fields in CNN [10]. In contrast, vision transformers, such as ViT [6], have demonstrated superior performance in person ReID. Due to the integration of multi-head self-attention mechanisms, ViT is adept at capturing long-range dependencies and exhibits enhanced effectiveness in focusing on various segments of the human body, offering a distinct advantage over CNNs in person re-identification. Although CNN-based and ViT-based methods for person ReID have demonstrated promising results on established person re-identification datasets, their true potential remains constrained by the limited scope of these datasets. Furthermore, these methods commonly rely on pre-trained weights on the ImageNet dataset, which utilizes manually assigned one-hot labels. Consequently, these methods can lead to an oversight of visual content with rich semantics that falls outside the predefined category set, thereby hindering the capture and incorporation of valuable information from such visual elements and ultimately limiting their re-identification performance.

Recently, pre-trained vision-language models have made substantial progress in capturing semantically rich visual concepts, attributed to the incorporation of natural language supervision. The Contrastive Language-Image Pre-training (CLIP) [33] represents a significant milestone of these advancements, effectively bridging the gap between visual content and their corresponding high-level textual descriptions, and aligning the two modalities in a harmonious manner. CLIP-ReID [20], a pioneer exploration inspired by recent advancements in vision-language models, stands out for its innovative approach to utilizing textual information to describe visual concepts that extend beyond mere appearance. As depicted in Fig. 1, CLIP-ReID provides a broader scope of supervision for the image encoder by introducing and fine-tuning global textual prompts with a robust text encoder, thereby enhancing the extraction of discriminative features from pedestrian images. Nevertheless, this type of prompt learning for the text encoder of CLIP has inherent limitations in capturing fine-grained details that are crucial for person ReID, and it is also susceptible to the influence of noises and occlusions. Furthermore, relying exclusively on single-modality textual prompts as constraints is rudimentary and inefficient, as it does not adequately model the necessary downstream adaptation of CLIP's image encoder.

To surmount these challenges, we propose a novel **M**ulti-Granularity and **M**ulti-Modal **P**rompt **L**earning framework (MMPL) for person re-identification, to learn robust and discriminative representations of pedestrians by fully exploiting the capabilities of pre-trained vision-language models like CLIP. The MMPL framework is constructed around a two-stage training procedure. Specifically, 1) in the first training stage, MMPL implements Hierarchical Prompt Learning (HPL) to distill information from hierarchical patch-level visual features and synchronize them with textual prompts across various granularities at the patch-to-token level. HPL is meticulously designed to generate comprehensive and nuanced textual prompts, facilitating a detailed and thorough depiction of pedestrian characteristics. Specifically, our approach begins by segmenting the image feature map into a multitude of patches of equal length, each representing a different granularity level. Subsequently, we train an information selector agent to identify patches that significantly contribute to the person ReID selectively. The agent operates within an environment that offers feedback in the form of rewards and updates its state based on the patch features and the similarity matrix it perceives. Its overarching goal is to optimize the cumulative expected reward by strategically selecting a fixed number of patches that contain discriminative cues. As a result, this process yields refined patch-level visual features across various granularities. Thereafter, we devise an optimal transport strategy to align these nuanced visual features with the learnable, multi-granularity textual prompts at the patch-to-token level. Such alignment augments the textual prompts' comprehensiveness, facilitating accurate and intricate representations of pedestrians and substantially diminishing the effects of noise and occlusions. 2) In the second training stage, we construct supplementary visual prompts and establish explicit constraints based on the text features extracted from the fine-tuned textual prompts through the

text encoder, facilitating the gradual assimilation of abundant information from the fine-tuned textual prompts into the image encoder of CLIP. Leveraging both visual and textual prompts, we construct Collaborative Prompt Learning (CPL), which effectively bridges the modality gap through multi-modal joint interactive learning, enabling the image encoder to harness CLIP's extensive multi-modal semantic knowledge. This results in the acquisition of more discriminative representations of pedestrians.

In summary, the principal contributions of this work are as follows: (1) We propose a novel MMPL framework, which employs a two-stage training process to fully harness the substantial capacity of CLIP for acquiring discriminative representations of pedestrians. (2) We architect the Hierarchical Prompt Learning to meticulously refine and synchronize essential and discernible information extracted from patch-level visual features, aligning them with corresponding textual prompts at the patch-to-token level. This methodically calibrated process across diverse granularities culminates in the production of high-quality, multi-granularity textual prompts. (3) We develop Collaborative Prompt Learning that integrates visual and textual prompts to facilitate multi-modal interactive learning, effectively aiding the CLIP image encoder in bridging the modal gap. This enables the image encoder to fully exploit the rich multi-modal semantic knowledge inherent in CLIP for learning more discriminative representations.

2. Related Work

2.1. Person Re-Identification

In the realm of computer vision, person ReID is a critical task that aims to identify and match individuals across various non-overlapping cameras. Previous research focuses on designing sophisticated hand-crafted descriptors to extract low-level features for pedestrians [5, 7, 22, 36]. For example, Gheissari *et al*. [7] introduced a spatiotemporal segmentation algorithm to generate normalized color and salient edgel histograms, which are robust to variations in person's appearance clothing. However, the design of sophisticated hand-crafted descriptors is time-consuming and challenging. CNN [15] has liberated the field from the tedious task of manual feature design. Methods that utilize CNNs for the automatic extraction of features from personal images have garnered substantial success [9, 12, 25, 26, 31]. For example, Qian *et al*. [32] developed a multi-scale deep representation learning model to capture discriminative cues at various scales. Considering that fine-grained local clues are useful for distinguishing different pedestrians, Sun *et al*. [38] partitioned the feature map of pedestrians into horizontal stripes, which serves to augment the model's capacity for local region representation. Subsequently, harnessing the capabilities of the Vision Transformer architecture, a variety of Transformer-based approaches for person ReID have come to the forefront. Such as, He *et al*. [10] proposed a pure transformer-based object ReID framework which generates robust features with improved discrimination ability. Zhu *et al*. [60] proposed the auto-aligned transformer to automatically locate both human and non-human parts at the patch level. Due to the limited scale of person ReID datasets, Transformer-based and CNN-based methods frequently encounter the problem of overfitting.

Recently, Li *et al*. [20] proposed CLIP-ReID, which introduces the concept of the large pre-trained visual-language model into the realm of person ReID, surmounting the limitations of traditional CNN and Transformer models hampered by the compact scale of person ReID datasets. Remarkable performance was been achieved with only minimal fine-tuning. Furthermore, Zhai *et al*. [47] introduced MP-ReID, which leveraged the generated multiple person attributes as prompts with CLIP, enhancing the accuracy of retrieval results. Additionally, Li *et al*. [18] directly fine-tuned the image encoder of CLIP by introducing a prototypical contrastive learning loss. However, the single-perspective, single-modality prompts employed by them are too coarse to fully and meticulously describe pedestrians and to effectively harness the multi-modal semantic knowledge inherent in CLIP. It also lacks sufficient robustness, particularly when addressing common challenges in person ReID, such as occlusions and misalignments. In contrast to the aforementioned methods, our MMPL successfully achieves multi-level descriptions of pedestrians and synergizes high-level semantic information to extract robust features from images through the proposed multi-granularity and multi-modal prompts.

2.2. Vision-Language Pre-training

The integration of language supervision with natural images has garnered significant interest in the computer vision community [2, 16, 23, 50]. In contrast to models trained with image supervision alone, vision-language models encode abundant multi-modal representations. These vision-language pre-trained models aim to explore the semantic correspondence between the vision and language modalities through large-scale pre-training. For example, Radford *et al*. [33] proposed CLIP, a pioneering model that synergistically pre-trains image and text encoders on a vast array of text-image pairs sourced from the Internet, which efficiently aligns the representations of images and text using a contrastive loss function. Moreover, Li *et al*. [19] proposed Bootstrapping Language-Image Pre-training, which employs the multi-modal mixture of encoder-decoder to achieve cross-modal information flow for effective multi-task pre-training and flexible transfer learning. In addition, Chen *et al*. [24] proposed Language and Vision Assistant, which connects a vision encoder and large language mod-

els to facilitate general-purpose visual and language understanding. Building on this success, we envision harnessing the extensive multi-modal knowledge embedded in vision-language models to propel the field of person ReID forward.

2.3. Prompt Learning

Complementing the vision-language model is the concept of prompt learning, a paradigm that adapts the large vision-language models to downstream tasks [2, 14, 27, 44, 57]. This innovation allows models to engage in zero-shot or few-shot learning, applying their pre-trained knowledge to novel tasks with minimal additional training.

Traditional prompt learning methods typically involve manually designing a prompt. For example, Zhang *et al*. [51] proposed Contrastive Learning of Medical Visual Representations for generating candidate prompts through text mining and paraphrasing, subsequently selecting the most effective prompts based on the highest training accuracy. Nevertheless, the design of a prompting function is intricate and relies on heuristics. Zhou *et al*. [56] introduced the notion of learnable textual prompt, which have been shown to significantly outperform manually designed prompts in adapting the CLIP model to a variety of tasks. Building on this foundation, Zhou *et al*. [55] proposed Conditional Context Optimization which utilizes dynamic prompts that adapt to each instance and are thus more robust to class shift. The integration of CLIP with prompt learning signifies a notable advancement in domains such as zero-shot recognition [41], object detection [28], and image segmentation [42], showcasing the potential of embedding rich linguistic contexts into visual models to enhance their interpretative prowess.

3. Method

In this section, we first provide an overview of the CLIP-ReID, as detailed in Section 3.1. Subsequently, we elaborate on our proposed MMPL framework. As illustrated in Fig. 2, MMPL encompasses a two-stage training procedure, which consists of Hierarchical Prompt Learning and Collaborative Prompt Learning.

3.1. Preliminaries

CLIP-ReID incorporates a two-stage training strategy designed to adapt the CLIP to the person ReID. We define the pre-trained text and image encoders of CLIP as $\mathcal{T}(\cdot)$ and $\mathcal{I}(\cdot)$, respectively.

In the first training stage, the focus is on optimizing global learnable tokens under the guidance of CLIP-style supervision. The global textual prompt $\boldsymbol{P}_t$ is formulated as *"A photo of* $[X_1][X_2]\cdots[X_N]$ *person"*. Here, $[X]$ is a learnable text token with the same dimension as word embedding, designed to capture discriminative information corresponding to pedestrian identities, and N denotes the total number of learnable tokens. We acquire image embedding $\boldsymbol{V}_{yi}$ and text embedding $\boldsymbol{T}_{yi}$ using an image $\boldsymbol{I}_i$ with an ID label y_i and the corresponding global textual prompt $\boldsymbol{P}_{t_i}$ through a frozen image encoder $\mathcal{I}(\cdot)$ and a frozen text encoder $\mathcal{T}(\cdot)$ as follows:

$$\boldsymbol{V}_{y_i} = \mathcal{I}(\boldsymbol{I}_i); \quad \boldsymbol{T}_{y_i} = \mathcal{T}(\boldsymbol{P}_{t_i}) \tag{1}$$

In the first training stage, only the text tokens $[X]$ are learned by optimizing the contrastive learning losses $\mathcal{L}_{i2t}$ and $\mathcal{L}_{t2i}$ as detailed below:

$$\mathcal{L}_{t2i}(y_i) = -\frac{1}{|D(y_i)|}\sum_{d\in D(y_i)} \log \frac{\exp(s(\boldsymbol{V}_p, \boldsymbol{T}_{y_i}))}{\sum_{a=1}^{B}\exp(s(\boldsymbol{V}_a, \boldsymbol{T}_{y_i}))} \tag{2}$$

$$\mathcal{L}_{i2t}(y_i) = -\frac{1}{|D(y_i)|}\sum_{d\in D(y_i)} \log \frac{\exp(s(\boldsymbol{T}_{y_i}, \boldsymbol{V}_d))}{\sum_{a=1}^{B}\exp(s(\boldsymbol{T}_{y_i}, \boldsymbol{V}_a))} \tag{3}$$

$$\mathcal{L}_{stage1} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i} \tag{4}$$

where $D(y_i) = \{d \mid y_d = y_i, d \in \{1, 2, \ldots, B\}\}$ is the set of indices corresponding to positives for $\boldsymbol{T}_{y_i}$ within a batch of size B and $s(\cdot,\cdot)$ represents the cosine similarity.

During the second training stage, updates are restricted to the parameters within the image encoder $\mathcal{I}(\cdot)$. CLIP-ReID employs the ID loss $\mathcal{L}_{id}$ and triplet loss $\mathcal{L}_{tri}$ [29] with label smoothing for optimization that are widely used in supervised person ReID. Furthermore, an image-to-text cross-entropy loss $\mathcal{L}_{i2tce}$ is implemented by capitalizing the global fine-tuned textual prompts as follows:

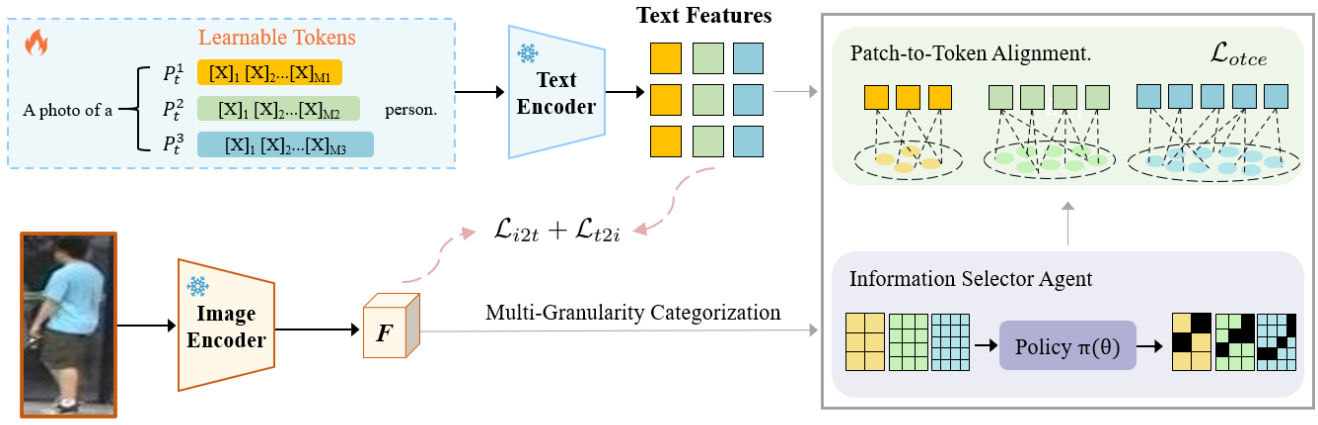
$$\mathcal{L}_{i2tce} = \sum_{l=1}^{L} -q_l \log\left(\frac{\exp(s(\boldsymbol{V}_i \cdot \boldsymbol{T}_{y_l}))}{\sum_{y_a=1}^{L}\exp(s(\boldsymbol{V}_i \cdot \boldsymbol{T}_{y_a}))}\right) \tag{5}$$

$$\mathcal{L}_{stage2} = \mathcal{L}_{id} + \mathcal{L}_{tri} + \mathcal{L}_{i2tce} \tag{6}$$

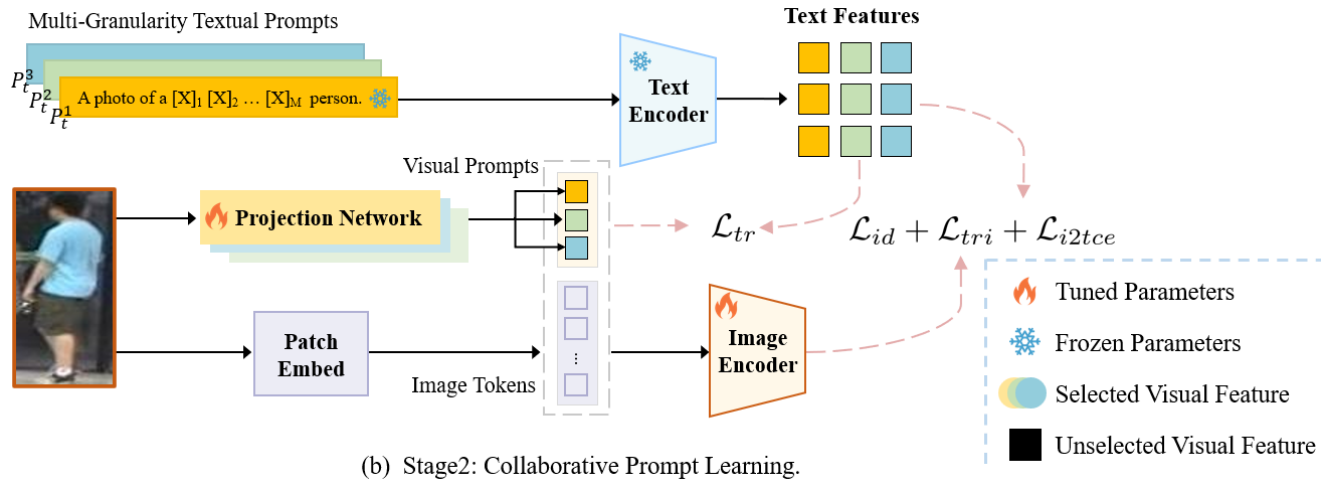
where q_l is the smoothed label, L is the number of identities and i denotes the person image index.

3.2. Hierarchical Prompt Learning

However, the global textual prompt in CLIP-ReID is inherently limited in capturing the subtleties of individual characteristics, which restricts its broader application in person ReID. Accordingly, we propose Hierarchical Prompt Learning to obtain a comprehensive and discriminative text representation of pedestrians, furnishing effective supervision for the image encoder in CLIP. In the first training stage, we integrate learnable tokens to encapsulate multi-granularity fine-grained textual descriptions of pedestrians. These are formulated as hierarchical textual prompts $\{\boldsymbol{P}_t^k\}_{k=1}^{K}$, with K represents the total number of granularity levels. And $\boldsymbol{P}_t^k$ is the textual prompt at the granularity



(a) Stage1: Hierarchical Prompt Learning.



(b) Stage2: Collaborative Prompt Learning.

Figure 2. The overall architecture of the proposed MMPL. To describe pedestrians with precision and detail, in the first training stage, we meticulously designed Hierarchical Prompt Learning to generate detailed multi-granularity textual prompts, facilitated by the Information Selector Agent and Patch-to-Token Alignment. Subsequently, in the second stage, we introduce Collaborative Prompt Learning to foster multi-modal interactive learning, which assists CLIP's image encoder in bridging the modal gap and leveraging multi-modal knowledge to develop discriminative pedestrian representations.

level k, which is designed as "*A photo of* $[X_1^k][X_2^k]\cdots[X_N^k]$ *person*". Here, N is the number of learnable tokens, which varies depending on the granularity level k.

Information Selector Agent. In contrast to global textual prompts, our hierarchical textual prompts provide a detailed and comprehensive description of pedestrians across multiple levels. To effectively update these prompts, the initial requirement is to capture key pedestrian features at various hierarchical levels. We meticulously engineer a deep reinforcement learning module to facilitate the extraction of multi-level salient information at the patch level. Firstly, the feature map $\boldsymbol{F}$ extracted from the image encoder is segmented into M_c non-overlapping, equal-sized patches, denoted as $\{\boldsymbol{f}_j\}_{j=1}^{M_c}$, with varying granularity levels k corresponding to different M_c, enabling the extraction of multi-level salient information via multi-granularity categorization. By segmenting the feature map into a collection of patches at hierarchical levels, we capture intricate details and maximally retain the structural integrity of the human body. Subsequently, we harness a deep reinforcement learning agent to identify patches that encapsulate salient discriminative information, offering accurate and efficacious guidance for the cultivation of our hierarchical textual prompts. We formulate the process of mining key patches as a one-step Markov Decision Process (MDP) and train an information selector agent. Specifically, the agent interacts with the environment, receiving rewards and updating its state to maximize the cumulative expected re-

ward by learning an optimal patch selection strategy for acquiring as many discriminative clues as possible, all within the constraint of a fixed number of patches.

The MDP comprises states, actions, and rewards, which are detailed as follows (for clarity, the granularity level k is omitted in the subsequent discussion):

State. The state $\{\boldsymbol{s}_j\}$ comprises two components: $(\boldsymbol{s}_j^f, s_j^e)$. Here, $\boldsymbol{s}_j^f$ is the feature of single patch f_j at various granularity levels. s_j^e is the similarity score, which assesses the unity between the global feature map $\boldsymbol{F}$ and the individual patch by calculating the Euclidean distance. We utilize s_j^e as the weighting factor for updating the reward.

Agent. We employ a Bidirectional Long Short-Term Memory network (Bi-LSTM) as the patch selection $\mathcal{A}(\cdot)$, with a fully connected layer and sigmoid function on top for predicting the patch selection probability p_j for each patch, which is formulated as follows:

$$p_j = \mathcal{A}(\boldsymbol{s}_j^f) \tag{7}$$

Action. The action corresponds to the selection of each patch. Our information selector agent $\mathcal{A}(\cdot)$ sequentially assigns a binary label, either 1 or 0, to each patch according to a Bernoulli distribution $\mathcal{F}(\cdot)$.

$$a_j = \mathcal{F}(p_j) \tag{8}$$

Here, $a_j \in \{0, 1\}$ denotes the selection status of the patch, with 1 indicating retained and 0 indicating discarded.

Reward. The reward quantifies the utility of the agent's action with respect to the current state. We define the reward as follows:

$$\text{reward} = \frac{p - p_0}{1 - p_0} \tag{9}$$

$$p_0 = \frac{1}{M_c}\sum_{m=1}^{M_c} s_j^e, \quad p = \frac{\sum_{m=1}^{M_c} a_j \cdot s_j^e}{\sum_{m=1}^{M_c} a_j} \tag{10}$$

where p_0 denotes the scores of all patches, and p represents the scores of the patches chosen by the agent. By maximizing the reward, the information selector agent is able to select patches that contain more prominent clues.

We train the information selector agent employing the Policy Gradient [45], aiming to the identify optimal parameter set $\boldsymbol{\theta}$ that defines the policy function, thereby maximizing the expected cumulative reward of $\pi_{\boldsymbol{\theta}}$. The reward function is formally denoted as $R_n = \sum_j r(\boldsymbol{s}_j, a_j)$, which corresponds to the reward computed at the z_{th} episode. The gradient is approximated by conducting Z episodes with the agent on a consistent image set. We aim to minimize E_m to guide the agent in selecting M patches and define the optimization goals as $O(\boldsymbol{\theta})$:

$$\nabla_{\boldsymbol{\theta}} O(\boldsymbol{\theta}) \approx \frac{1}{Z}\sum_{u=1}^{Z}\sum_{j=1}^{M_c} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_j^{(z)}|\boldsymbol{s}_j^{(z)}) R_n \tag{11}$$

$$E_m = \gamma||_1 \frac{1}{M_c}\sum_{j=1}^{M_c} p_j - \frac{M}{M_c}||_1 \tag{12}$$

where $a_j^{(z)}$ is the action taken by the information selector agent and $s_j^{(z)}$ is the state for the patch j in z_{th} episode.

Patch-to-Token Alignment. Previous methods that exclusively depend on contrastive learning loss functions for global alignment between images and textual prompts are deemed to be rudimentary and inefficient. Building upon this foundation, we develop a patch-to-token alignment, enhancing the nuanced expressive power of prompts. After obtaining hierarchical fine-grained patches, we employ the optimal transport strategy to align these patch features with multi-granularity textual prompts, enabling the learnable tokens of textual prompts to capture more distinctive pedestrian information within the selected patches. The selected patches are denoted as $\{\boldsymbol{f}_m\}_{m=1}^{M}$, where M denotes the number of selected patch features. The corresponding granularity learnable textual prompt is denoted as $\{\boldsymbol{p}_n\}_{n=1}^{N}$ where N is the number of the learnable tokens of textual prompts. We define the total distance between them as follows:

$$\langle \tilde{\boldsymbol{T}}, \boldsymbol{C} \rangle = \sum_{m=1}^{M}\sum_{n=1}^{N} \tilde{\boldsymbol{T}}_{m,n} \boldsymbol{C}_{m,n} \tag{13}$$

where $\boldsymbol{C}$ represents the cost matrix, wherein each element signifies the cost associated with the pairing of $\{\boldsymbol{f}_m\}_{m=1}^{M}$ and $\{\boldsymbol{p}_n\}_{n=1}^{N}$. Furthermore $\tilde{\boldsymbol{T}} \in \mathbb{R}^{M\times N}$ is learned to minimize the total distance between these pairs. It should be noted that $\tilde{\boldsymbol{T}}_{m,n}$ measures the transported probability from the m-th visual patch to the n-th learnable token of the textual prompt. To facilitate rapid optimization, we utilize the Sinkhorn distance [4], utilizing an entropic constraint. The distance is thus formulated as an entropy-regularized optimal transport problem, which is articulated as follows:

$$d_{OT,\lambda}(\boldsymbol{u}, \boldsymbol{v}|\boldsymbol{C}) = \underset{\tilde{\boldsymbol{T}}}{\text{minimize}} \langle \tilde{\boldsymbol{T}}, \boldsymbol{C} \rangle - \lambda h(\tilde{\boldsymbol{T}}) \tag{14}$$

$$\text{subject to } \tilde{\boldsymbol{T}}\mathbf{1}_N = \boldsymbol{u}, \tilde{\boldsymbol{T}}^\top \mathbf{1}_M = \boldsymbol{v}, \tilde{\boldsymbol{T}} \in \mathbb{R}_+^{M\times N} \tag{15}$$

where $h(\cdot)$ denotes the entropy, vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are defined as discrete probability vectors that sum to 1, and λ is a hyper-parameter that influences the entropy regularization. Subsequently, rapid optimization can be achieved with minimal iterations, as detailed below:

$$\tilde{\boldsymbol{T}}^* = \text{diag}(\boldsymbol{u}^{\tilde{t}}) \exp(-\boldsymbol{C}/\lambda) \,\text{diag}(\boldsymbol{v}^{\tilde{t}}) \tag{16}$$

$$\boldsymbol{u}^{\tilde{t}} = \frac{\boldsymbol{u}}{\exp(-\boldsymbol{C}/\lambda)\,\boldsymbol{v}^{\tilde{t}-1}}, \quad \boldsymbol{v}^{\tilde{t}} = \frac{\boldsymbol{v}}{\exp(-\boldsymbol{C}/\lambda)\,\boldsymbol{u}^{\tilde{t}-1}} \tag{17}$$

where $\tilde{t}$ denotes the iteration. Upon obtaining the learned transport plan $\tilde{\boldsymbol{T}}^*$, we proceed to define the prediction probability of the image $\boldsymbol{x}$ with ID label y_x (one-hot label vector) as follows:

$$p(y=l|\boldsymbol{x}) = \frac{\exp\left((1-d_{OT,\lambda}(l))/\tau\right)}{\sum_{l=1}^{L}\exp\left((1-d_{OT,\lambda}(l))/\tau\right)} \tag{18}$$

We optimize the multi-granularity textual prompt with cross-entropy loss:

$$\mathcal{L}_{\text{otce}} = -\frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{l=1}^{L} y_x p(y=l|\boldsymbol{x}) \tag{19}$$

Similarly, the multi-granularity textual prompts $\{\boldsymbol{P}_t^k\}_{k=1}^K$ are also optimized through Eq.(2) and Eq.(3):

$$\mathcal{L}_{stage1} = \sum_{k=1}^{K}\alpha_k * (\mathcal{L}_{i2t}^k + \mathcal{L}_{t2i}^k) + \sum_{k=1}^{K}\beta_k * \mathcal{L}_{otce}^k \tag{20}$$

Through hierarchical prompt learning, the learned multi-granularity textual prompts are capable of furnishing comprehensive and detailed descriptions of pedestrians.

3.3. Collaborative Prompt Learning

The single-modal prompt is insufficient for unearthing the rich multi-modal semantic knowledge inherent in CLIP and fails to offer direct and effective guidance to the image encoder. To surmount this limitation, a conceivable solution is the creation of independent visual prompts. However, such a design may lack synergistic integration between the visual and linguistic components, potentially resulting in semantic misalignment between the visual and textual prompts. Furthermore, the direct mapping of text embeddings to visual prompts faces challenges during the inference phase due to the paucity of textual information.

Consequently, we design extra visual prompts that are aligned with the learned multi-granularity textual prompts. This strategy effectively facilitates the transmission of multi-level semantic information, enhancing the model's capability to capture and utilize rich textual contexts. Based on visual and textual prompts, we construct the multi-modal joint interactive learning to assist the image encoder in bridging the modality gap, fully utilizing the rich multi-modal semantic knowledge within CLIP to learn discriminative representations.

In the second training stage, the objective is to optimize the image encoder $\mathcal{I}(\cdot)$ utilizing the high-level semantic information embedded within the learned multi-granularity textual prompt. We initially construct visual prompts $\boldsymbol{P}_v$ from the input images I through a projection network, ensuring that this process does not introduce additional noise or interference.

$$\boldsymbol{P}_v = MLP(GAP(Conv(\boldsymbol{I})) \tag{21}$$

Here, $MLP(\cdot)$ represents the weights of a learnable multi-layer perceptron, whereas $GAP(\cdot)$ and $Conv(\cdot)$ are global average pooling layer and convolution operations with the filter size of 3×3, respectively. We encourage the visual prompts to capture high-level semantic information corresponding to their respective granularities while ensuring diversity among them to prevent the multi-prompt from converging on redundant information. Follow this idea, we propose the transformation loss $\mathcal{L}_{al}$ as:

$$\boldsymbol{T}^k = \mathcal{T}(\boldsymbol{P}_t^k) \tag{22}$$

$$\mathcal{L}_{tr} = \sum_{k=1}^{K}||\boldsymbol{P}_v^k - \boldsymbol{T}^k||_1 - \sum_{i=1}^{K-1}\sum_{j=i+1}^{K}||\boldsymbol{P}_v^i - \boldsymbol{P}_v^j||_1 \tag{23}$$

where $\mathcal{T}(\cdot)$ is the frozen text encoder, while $\boldsymbol{P}_v^k$ and $\boldsymbol{P}_t^k$ denote the visual and textual prompts, respectively, at their respective granularities k. This methodology empowers the visual prompts to inherit multi-tiered, fine-grained information, thereby facilitating the image encoder's incremental acquisition of subtle discriminative cues.

After obtaining the visual prompts, they are concatenated with the image tokens $\boldsymbol{I}_{tokens}$ after the image through the patch embed, forming the input $[\boldsymbol{I}_{tokens}, \boldsymbol{P}_v^1, \ldots, \boldsymbol{P}_v^K]$ for the image encoder. The explicit conditioning of $\boldsymbol{P}_t$ on $\boldsymbol{P}_v$ facilitates the image encoder in progressively absorbing the nuanced semantic information embedded within the multi-granularity textual prompts, effectively bolstering our model's ability to harness latent multi-modal knowledge. Similarly, we also optimize the image encoder using the identity loss $\mathcal{L}_{id}$ and the triplet loss $\mathcal{L}_{tri}$ [29], as well as the image-to-text cross-entropy loss $\mathcal{L}_{i2tce}$:

$$\mathcal{L}_{stage2} = \mathcal{L}_{id} + \mathcal{L}_{tri} + \mathcal{L}_{i2tce} + \mathcal{L}_{tr} \tag{24}$$

4. Experiments

This section first introduces our experimental datasets, evaluation protocols, and implementation specifics. Then, we compare the proposed MMPL with various state-of-the-art approaches. Lastly, we execute ablation studies to discern the contribution of individual components and present an analysis of some visualization results.

4.1. Datasets and Evaluation Protocols

Our method is assessed on four benchmark datasets for person ReID: MSMT17 [43], Market-1501 [53], DukeMTMC-reID [35], and Occluded-Duke [30]. Table 2 provides a summary of the datasets incorporated in the study. We utilize the cumulative matching characteristics (CMC) at Rank-1 (R1) and the mean average precision (mAP) as performance metrics.

MSMT17 is a large-scale person re-identification dataset from a campus environment, comprising a total of 15 camera views, including 12 outdoor cameras and 3 indoor cameras. The dataset contains 4,101 identities and a total of

Table 1. Performance comparison to the state-of-the-art methods on four datasets.

Methods	References	MSMT17		Market-1501		DukeMTMC		Occluded-Duke	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1
ABD-Net [3]	ICCV (2019)	60.8	82.3	88.3	95.6	78.6	89.0	-	-
HOReID [40]	CVPR (2020)	-	-	84.9	94.2	75.6	86.9	43.8	55.1
SAN [13]	AAAI (2020)	55.7	79.2	88.0	96.1	75.5	87.9	-	-
OfM [48]	AAAI (2021)	54.7	78.4	87.9	94.9	78.6	89.0	-	-
CDNet [17]	CVPR (2021)	54.7	78.9	86.0	95.1	76.8	88.6	-	-
PAT [21]	CVPR (2021)	-	-	88.0	95.4	78.2	88.8	53.6	64.5
AAformer [60]	CVPRW (2021)	63.2	83.6	87.7	95.4	80.0	80.0	58.2	67.0
TransReID [10]	ICCV (2021)	67.4	85.3	88.9	95.2	82.0	90.7	59.2	66.4
CAL [34]	ICCV (2021)	56.2	79.5	87.0	94.5	76.4	87.6	-	-
HAT [49]	ACM MM (2021)	61.2	82.3	89.8	95.8	81.4	90.4	-	-
DCAL [58]	CVPR (2022)	64.0	83.1	87.5	94.7	80.1	89.0	-	-
PASS [59]	ECCV (2022)	71.8	88.2	93.0	96.8	-	-	-	-
HAWK [39]	ACM MM (2022)	68.7	87.9	89.6	96.6	83.1	91.6	58.8	66.2
CLIP-ReID [20]	AAAI (2023)	75.8	89.7	90.5	95.4	83.1	90.8	60.3	67.2
PCL-CLIP [18]	ArXiv (2023)	76.1	89.8	91.4	95.9	-	-	-	-
LFM [8]	IJCNN (2024)	77.1	87.4	86.6	94.8	-	-	-	-
IRM [11]	CVPR (2024)	72.4	86.9	93.5	96.5	-	-	-	-
MMPL		**78.5**	**92.1**	**93.8**	**97.4**	**86.1**	**94.0**	**62.8**	**70.1**

126,441 image samples. The training set consists of 30,248 images, while the test set includes 11,659 query samples and 82,161 gallery samples.

Market-1501 was gathered on campus using five relatively high-resolution cameras and one low-resolution camera. This dataset includes 32,688 images of 1,501 identities. The training set consists of 12,936 images representing 751 identities. The remaining 750 identities are split between the query set, consisting of 3,368 images, and the gallery set, which includes 19,734 images.

DukeMTMC-reID is a dataset collected from eight camera views on campus. It includes 1,404 identities. The dataset is divided into a training set with 16,522 images of 702 identities, a query set with 2,228 images, and a gallery set with 17,661 images of 1,110 identities.

Occluded-Duke is a standard occluded person ReID dataset derived from the DukeMTMC-reID dataset. The training set encompasses 15,618 images of 702 identities, with 9% of the images being occluded. In the test set, there are 2,210 occluded query images and 17,661 gallery images of 1,110 identities, and 10% of them are occluded.

4.2. Implementation Details

We implement the proposed method with PyTorch and training on a single NVIDIA V100 GPU. The pre-trained image and text encoder from CLIP serve as the backbone for our image and text feature extractors, respectively. For the image encoder, we employ the ViT-B/16 with 12 transformer blocks, each with a hidden size of 768 dimensions, pass the output of the encoder through a linear projection layer to reduce the feature dimension from 768 to 512.

Table 2. Statistics of the experimental datasets.

Dataset	ID	Person Image	amera
MSMT17 [43]	4,101	126,441	15
Market-1501 [53]	1,501	32,668	6
DukeMTMC [35]	1,404	36,411	8
Occluded-Duke [30]	1,404	35,489	8

In the first training stage, we train the information selector agent for 50 epoch with Adam optimizer and the learning rate is set to be 1×10^{-4}. The MMPL framework we used for comparison employs prompts of three different granularities, with corresponding prompt lengths (N) of 4, 6, and 8. The numbers of segmentation patches (M_c) are 16, 64, and 128, and the selected numbers of patches (M) are 10, 40, and 80, respectively. The input images are resized to 256×128 and the batch size is 64. The learnable textual prompts are optimized using the Adam optimizer across 120 epochs, with an initial learning rate set at 3.5×10^{-4}, while all other parameters remain frozen. The hyper-parameters α and β are uniformly set at 1 and 0.1. For optimal transport strategy, we set the hyper-parameters in Sinkhorn distances

algorithm [4] as $\lambda = 0.1$. The maximum number of iterations for the inner loop is set at 50, and we will implement early stopping when the average absolute update value is less than 0.05.

In the second training stage, image augmentation is performed through random horizontal flipping, padding cropping, and random erasing [54] to 256×128. We use the Adam optimizer to train the model for 60 epochs, with an initial warm-up phase of 10 epochs during which the learning rate linearly increases from 5×10^{-7} to 5×10^{-6}. Then, the learning rate is reduced by a factor of 0.1 at the 30th and 50th epochs.

4.3. Comparison with State-of-the-Art Methods

Our method is compared with state-of-the-art approaches on three widely recognized person ReID datasets and one occluded person ReID dataset, with the results presented in Table 1.

Observations indicate that the proposed MMPL achieves state-of-the-art performance across all four datasets evaluated. Specifically, within the range of methods compared, there is a discernible trend indicating progressive enhancement in person ReID. Firstly, the vision transformer serves as a more powerful backbone compared to traditional CNN-based methods due to its enhanced capability to capture global dependencies within an image. For example, TransReID [10] demonstrates superior performance that significantly exceeds the capabilities of the previous CNN-based model. Furthermore, the introduction of visual-language pre-training by CLIP-ReID [20] has resulted in leaps forward across various performance metrics. This showcases the tremendous potential inherent in visual-language pre-training for person ReID. In addition, under the same backbone conditions, networks that focus on fine-grained features often achieve better performance than those relying solely on global features (*e.g.*, ABD-Net [3] *vs* CDNet [17]). Our MMPL achieves a 93.8% mAP and 97.4% Rank-1 accuracy on Market-1501 outperforming other methods. Although IRM [11] demonstrates comparable performance on this dataset, a noticeable performance discrepancy emerges on other datasets, demonstrating the superior adaptability and robustness of our MMPL.

Compared to all the aforementioned methods, the strength of our MMPL can be attributed to several aspects: 1) The multi-granularity prompts we fine-tuned successfully capture both global structural information and fine-grained details, aiding the model in learning discriminative pedestrian representation. 2) The multi-modal joint interactive learning framework we proposed aids the image encoder of CLIP in bridging the modal gap. This leverages the extensive, advanced semantic information embedded within CLIP to enhance feature discriminability. 3) We propose the MMPL framework, which leverages Hierarchical Prompt Learning (HPL) and Collaborative Prompt Learning (CPL) to effectively excavate the potential of large-scale vision-language pre-trained models like CLIP for applications in the person ReID. It adeptly harnesses sophisticated semantic information to effectively guide the acquisition of discriminative visual features.

Table 3. Analysis of HPL and CPL on the Market-1501 dataset.

HPL	CPL	R1 (%)	mAP (%)
×	×	95.1	90.3
✓	×	96.6	92.2
×	✓	96.0	91.4
✓	✓	**97.4**	**93.8**

Table 4. Analysis of ISA and PTA for HPL on the Market-1501 dataset.

ISA	PTA	R1 (%)	mAP (%)
×	×	96.0	91.4
✓	×	96.5	92.5
×	✓	96.7	92.8
✓	✓	**97.4**	**93.8**

4.4. Ablation Studies and Analysis

Analysis of each component in MMPL. To verify the impact of each component within MMPL, we present the results of the ablation study in Table 3. We regard the model without HPL and CPL as the baseline, similar to CLIP-ReID. In comparison to the baseline, when only HPL is adopted, we can observe the performance is improved by +1.5% R1 accuracy and +1.9% in mAP. This display of the HPL effectively captures fine-grained information across multiple levels, providing sophisticated semantic guidance to the image encoder in CLIP for learning a comprehensive and precise description of pedestrians. Furthermore, the incorporation of CPL into the baseline leads to additional improvements, with a +0.9% increase in R1 accuracy and +1.1% improvement in mAP. This outcome evidences that CPL assists the model in narrowing the semantic gap between modalities, enabling it to harness abundant multi-modal knowledge for the learning of visual features. Upon integrating HPL and CPL, notable advancements are attained, with a +2.3% increase in R1 accuracy and a +3.5% improvement in mAP. These enhancements confirm that MMPL successfully maximizes CLIP's capabilities for generating discriminative feature representations through the synergistic application of HPL and CPL.

Analysis of each component in HPL. We conduct ablation studies on the Market-1501 dataset to validate the effectiveness of the Information Selector Agent (ISA) and

(a) Baseline (b) MMPL

Figure 3. Showcases some Rank-5 retrieval results obtained by both the baseline and our MMPL on the Market-1501 dataset. The green boxes indicate correct matches and red ones rep- resent incorrect matches.

Patch-to-Token Alignment (PTA) within HPL. In the absence of ISA, we employ attention masks [46] as an alternative. Drawing parallels to [20], we optimize the learnable textual prompts solely using contrastive learning loss in scenarios where PTA is not employed. As evidenced in Table 4, the exclusive application of ISA led to an improvement of +1.1% in mAP and +0.5% in R1 accuracy. Conversely, the independent utilization of PTA yielded respective improvements of +1.4% in mAP and +0.7% in R1 accuracy. These findings substantiate the efficacy of our HPL, which harnesses ISA and PTA to generate multi-granularity prompts. The concurrent deployment of the ISA and PTA, as opposed to not utilizing either, leads to a performance improvement of +2.4% mAP and +1.4% in R1 accuracy. This underscores the synergistic effect between ISA and PTA in enhancing model performance.

Analysis of the influence of granularity level K. Table 5 illustrate the impact of utilizing the different granularity level K. Apparently, with the increase of K, which corresponds to the adoption of prompts with more granularities, the model's performance progressively improves. For example, MMPL3 demonstrates a +1.3% absolute enhance-

Table 5. Analysis of the influence of the granularity level K on the Market-1501 dataset.

	K	**R1 (%)**	**mAP** (%)
Baseline	×	95.1	90.3
MMPL1	1	96.1	91.8
MMPL2	2	97.0	93.3
MMPL3	3	**97.4**	**93.8**
MMPL4	4	97.1	93.2
MMPL5	5	96.9	92.7

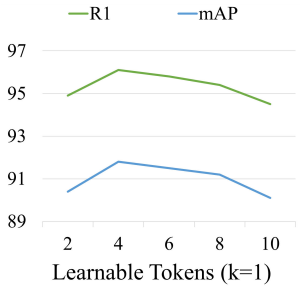


Figure 4. Analysis of learnable tokens N with $k = 1$ on the Market-1501.

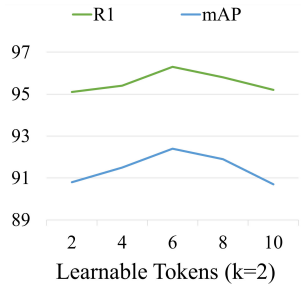


Figure 5. Analysis of learnable tokens N with $k = 2$ on the Market-1501.

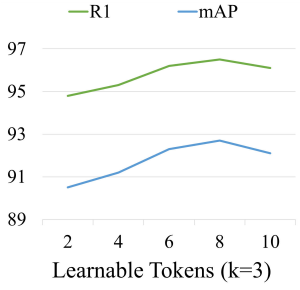


Figure 6. Analysis of learnable tokens N with $k = 3$ on the Market-1501.

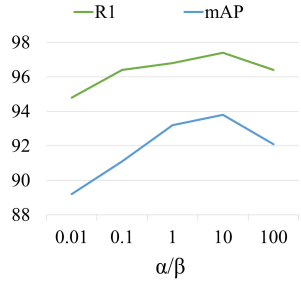


Figure 7. Analysis of hyper-parameter α/β on the Market-1501.

ment in R1 accuracy and a +2.0% improvement in mAP compared to MMPL1. This suggests that our HPL has effectively captured fine-grained information, and the mutual supplementation of prompts at various granularities collectively enhances the representation of pedestrians. However, we find that increasing the granularity level beyond three provides limited performance improvements but results in a significant increase in memory consumption, and also leads to overfitting, as evidenced by the performance decline of MMPL4 compared to MMPL3.

Analysis of the influence of N. We investigate the impact of the number of learnable tokens, denoted by N. As illustrated in Fig. 4, the model performs optimally with $N = 4$ when $k = 1$, suggesting that an insufficient number of tokens fail to adequately describe pedestrians, while an excess leads to overfitting and hinders generalization performance. Comparing with Fig.s 5 and 6, it can be observed that the optimal number of tokens increases as the coarseness of a single granularity level is adjusted. For instance, at granularity level $k = 3$, an optimal prompt length of 8 is identified, compared to 6 at $k = 2$. This finding supports the rationale and effectiveness of employing variable-length textual prompts in our multi-granularity prompts.

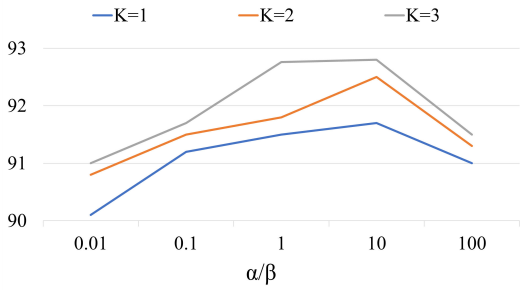


Figure 8. The analysis of hyper-parameters α/β on the Market-1501 dataset, with respect to granularity level K, is conducted using mAP as the metric.

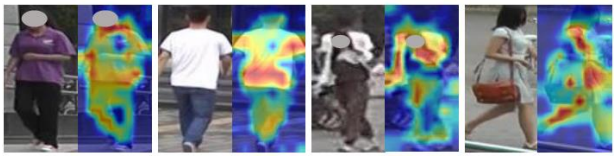
Table 6. Complexity Comparison on Market-1501.

Methods	**Params (M)**	**FLOPs (G)**	**mAP (%)**
TransReID [10]	85.9	17.2	88.9
HAWK [39]	72.3	15.2	89.6
CLIP-ReID [20]	**57.5**	7.92	90.5
MMPL	60.1	**8.27**	**93.8**

Analysis of hyper-parameter α/β. Fig. 8 demonstrates the effect of the hyper-parameter α and β. The values displayed here indicate the relative magnitudes of the loss coefficients for each textual prompt. It is evident that an imbalanced ratio of α/β can detrimentally impact the model's performance with an optimal ratio established at $\alpha/\beta = 10$. Concurrently, we explored the impact of α/β on prompt learning across various granularities, as depicted in Fig. 8. All achieve optimal performance when α/β is set to 10, albeit with differing sensitivities to this parameter.

Analysis of Complexity for MMPL. As illustrated in Tabel 6, we reproduce the SOTA methods on Market-1501. Our method demonstrates significant advantages over TransReID [10] and HAWK [39] in terms of Params (**60.1M** VS 85.9M and 72.3M) and FLOPs (**8.27G** VS 17.2G and 15.2G). Moreover, our method also achieves substantial improvements in mAP (**93.8%** VS 88.9% and 89.6%) compared to these existing methods. Compared with the slightly less parameter-heavy CLIP-ReID, our model undeniably outperforms in terms of performance. These results demonstrate that our method is not only efficient but also effective, exhibiting lower complexity compared to other methods while delivering superior performance.

(a) Baseline



(b) MMPL

Figure 9. The visualization of the class activation maps for the baseline and MMPL on the Market-1501 dataset.

4.5. Visualization

Visualization of Feature Heatmaps. Fig. 9 showcases some visualization experiments using the class activation maps (CAMs) [1]. Our MMPL enables more exhaustive and meticulous observation of pedestrians than the baseline. For example, as shown in the second column, our MMPL attends to diverse discriminative areas, forming a comprehensive and detailed observation of pedestrians. In contrast, the baseline concentrates on certain parts, omitting other intricate details of the human body.

Visualization of Retrieved Results. As depicted in Fig. 3, we selectes eight pedestrians images from the Market-1501 dataset. Significantly, MMPL is able to effectively improve the ranking results by retrieving more correctly matched images than the baseline. For instance, the baseline incorrectly matched three images in the third row, whereas MMPL achieved correct matches for all.

5. Conclusion

In this work, we present an exhaustive review of the current research in person ReID and propose the Multi-Granularity and Multi-Modal Prompt Learning framework for person ReID. This framework incorporates Hierarchical Prompt Learning (HPL) and Collaborative Prompt Learning (CPL) to fully harness the capabilities of pre-trained vision-language models, such as CLIP, for learning robust and discriminative representations. Within the HPL, we develop the Information Selector Agent (ISA) to refine and synchronize key visual information with textual prompts across various granularities at the patch-to-token level, resulting in high-quality, multi-granularity textual prompts. In the CPL, we establish multi-modal interactive learning both visual and textual prompts, aiding CLIP's image encoder in bridging the modality gap and leveraging CLIP's rich multi-modal semantic information to achieve more nuanced person representations. Our method has demonstrated state-of-the-art performance across four widely recognized datasets for person ReID.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62476260, 62225207, 62436008 and 62106245, the Fundamental Research Funds for the Central Universities under Grant WK2100000057.

References

[1] H. Chefer, S. Gur, and L. Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021. 12

[2] G. Chen, W. Yao, X. Song, X. Li, Y. Rao, and K. Zhang. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 3, 4

[3] 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*, pages 8351–8361, 2019. 8, 9

[4] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 6, 9

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 3

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[7] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1528–1535. IEEE, 2006. 3

[8] Y. Gong, C. Zhang, Y. Hou, L. Chen, and M. Jiang. Beyond dropout: Robust convolutional neural networks based on local feature masking. *arXiv preprint arXiv:2407.13646*, 2024. 8

[9] Q. Han, L. Li, W. Min, Q. Wang, Q. Zeng, S. Cui, and J. Chen. Joint training with local soft attention and dual cross-neighbor label smoothing for unsupervised person re-identification. *Computational Visual Media*, 10(3):543–558, 2024. 3

[10] 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*, pages 15013–15022, 2021. 2, 3, 8, 9, 11

[11] W. He, Y. Deng, S. Tang, Q. Chen, Q. Xie, Y. Wang, L. Bai, F. Zhu, R. Zhao, W. Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17521–17531, 2024. 8, 9

[12] Z. Huang, J. Liu, L. Li, K. Zheng, and Z.-J. Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1034–1042, 2022. 3

[13] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11173–11180, 2020. 8

[14] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 4

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 3

[16] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[17] H. Li, G. Wu, and W.-S. Zheng. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6729–6738, 2021. 8, 9

[18] J. Li and X. Gong. Prototypical contrastive learning-based clip fine-tuning for object re-identification. *arXiv preprint arXiv:2310.17218*, 2023. 3, 8

[19] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[20] 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*, volume 37, pages 1405–1413, 2023. 2, 3, 8, 9, 10, 11

[21] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2898–2907, 2021. 1, 8

[22] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3610–3617, 2013. 3

[23] B. Lin, Z. Tang, Y. Ye, J. Cui, B. Zhu, P. Jin, J. Zhang, M. Ning, and L. Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 3

[24] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[25] J. Liu, Z.-J. 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*, pages 7202–7211, 2019. 3

[26] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei. Multi-scale triplet cnn for person re-identification. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 192–196, 2016. 3

[27] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 4

[28] Y. Long, J. Han, R. Huang, H. Xu, Y. Zhu, C. Xu, and X. Liang. Fine-grained visual–text prompt-driven self-training for open-vocabulary object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 4

[29] 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*, pages 0–0, 2019. 1, 4, 7

[30] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 542–551, 2019. 1, 7, 8

[31] H. Park and B. Ham. Relation network for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11839–11847, 2020. 3

[32] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5399–5408, 2017. 3

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[34] 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*, pages 1025–1034, 2021. 8

[35] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 7, 8

[36] L. Shengcai, H. Yang, Z. Xiangyu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA*, pages 7–12, 2015. 3

[37] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1179–1188, 2018. 1

[38] 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 (ECCV)*, pages 480–496, 2018. 3

[39] X. Tian, J. Liu, Z. Zhang, C. Wang, Y. Qu, Y. Xie, and L. Ma. Hierarchical walking transformer for object re-identification.

In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4224–4232, 2022. 8, 11

[40] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6449–6458, 2020. 8

[41] H. Wang, M. Yang, K. Wei, and C. Deng. Hierarchical prompt learning for compositional zero-shot recognition. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1470–1478, 2023. 4

[42] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 4

[43] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 7, 8

[44] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023. 4

[45] J. Wu, G. Li, S. Liu, and L. Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12386–12393, 2020. 6

[46] Y. Yang, W. Huang, Y. Wei, H. Peng, X. Jiang, H. Jiang, F. Wei, Y. Wang, H. Hu, L. Qiu, et al. Attentive mask clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2771–2781, 2023. 10

[47] 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*, volume 38, pages 6979–6987, 2024. 3

[48] E. Zhang, X. Jiang, H. Cheng, A. Wu, F. Yu, K. Li, X. Guo, F. Zheng, W. Zheng, and X. Sun. One for more: Selecting generalizable samples for generalizable reid model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3324–3332, 2021. 8

[49] G. Zhang, P. Zhang, J. Qi, and H. Lu. Hat: Hierarchical aggregation transformers for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 516–525, 2021. 8

[50] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 3

[51] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 4

[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*, pages 2153–2162, 2023. 1

[53] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 7, 8

[54] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 9

[55] 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*, pages 16816–16825, 2022. 4

[56] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 4

[57] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. 4

[58] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4692–4702, 2022. 8

[59] K. Zhu, H. Guo, T. Yan, Y. Zhu, J. Wang, and M. Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *European conference on computer vision*, pages 198–214. Springer, 2022. 8

[60] K. Zhu, H. Guo, S. Zhang, Y. Wang, J. Liu, J. Wang, and M. Tang. Aaformer: Auto-aligned transformer for person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 3, 8