A Pedestrian is Worth One Prompt: Towards Language Guidance Person Re-Identification

Zexian Yang[1,2] Dayan Wu[1*] Chenming Wu[3] Zheng Lin[1] Jingzi Gu[1] Weiping Wang[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences [3] Baidu Inc
{yangzexian,wudayan,linzheng,gujingzi,wangweiping}@iie.ac.cn,wuchenming@baidu.com

Abstract

Extensive advancements have been made in person ReID through the mining of semantic information. Nevertheless, existing methods that utilize semantic-parts from a single image modality do not explicitly achieve this goal. Whiteness the impressive capabilities in multimodal understanding of Vision Language Foundation Model CLIP, a recent two-stage CLIP-based method employs automated prompt engineering to obtain specific textual labels for classifying pedestrians. However, we note that the predefined soft prompts may be inadequate in expressing the entire visual context and struggle to generalize to unseen classes. This paper presents an end-to-end ***Prompt****-driven* ***S****emantic* ***G****uidance (****PromptSG****) framework that harnesses the rich semantics inherent in CLIP. Specifically, we guide the model to attend to regions that are semantically faithful to the prompt. To provide personalized language descriptions for specific individuals, we propose learning pseudo tokens that represent specific visual contexts. This design not only facilitates learning fine-grained attribute information but also can inherently leverage language prompts during inference. Without requiring additional labeling efforts, our PromptSG achieves state-of-the-art by over 10% on MSMT17 and nearly 5% on the Market-1501 benchmark. The codes will be available at* `https://github.com/YzXian16/PromptSG`

1. Introduction

Person Re-Identification (ReID) is a crucial research area in computer vision that focuses on identifying individuals across different camera views or time instances [4, 44, 45, 57], which is a sub-task of image-based retrieval. Features of the same individual, as captured by various cameras, are prone to alterations due to changes in lighting, background, and body posture. Consequently, the effectiveness of a sophisticated ReID model fundamentally depends on its capability to learn discriminative features that are impervious to camera-specific variations, thereby enhancing the model's capacity to generalize to previously unseen classes.

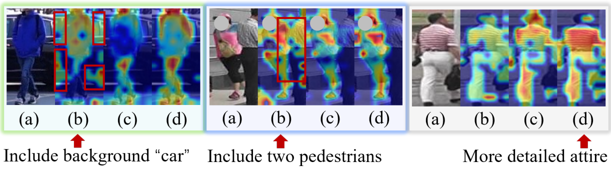

Figure 1. Transformer visualization [2] of attention maps. (a) Original images, (b) CLIP-ReID, (c) Our method w/o inversion, and (d) Our method guided by the composed prompts captures both the exact semantic parts and the external appearance details.

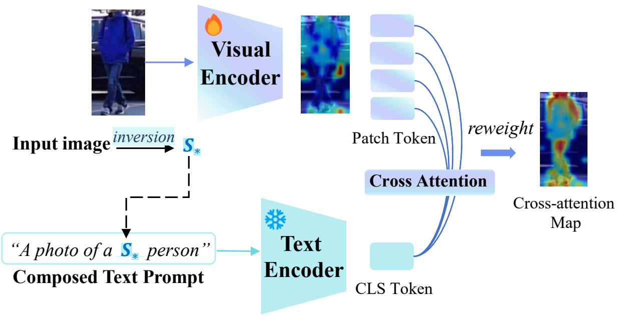

Figure 2. The core idea of our method. Our method inverts input images into pseudo-word tokens S_*, which are then composed into a textual prompt to describe the specific visual context. The attention map of patch tokens is further controlled by the semantics of the textual prompt.

Modern ReID models, constructed upon uni-modal architectures such as the Convolutional Neural Network (CNN) [22] or Vision Transformer (ViT) [14, 19, 35, 40], have made significant advancements within the field. A substantial portion of these solutions focus on the extraction of pertinent regions to rectify misalignment issues. These strategies are dedicated to the extraction of semantic

[*]Corresponding author.

data, such as the human body structure, primarily facilitated through the integration of identity classification [33, 43] and metric learning [24, 55]. However, it is worth noting that these attention regions generally highlight only specific locally discriminative parts without explicit semantic control. When a distinct mask or skeleton direction is necessitated [27, 31], the need for additional, labor-intensive, and time-consuming manual labeling becomes inevitable.

Large-scale Vision Language (VL) models, exemplified by Contrastive Language-Image Pre-Training (CLIP) [26], have recently shown remarkable abilities in reasoning across multi-modal data. CLIP model, when provided with text prompts such as 'A photo of a [CLASS]', displays exceptional zero-shot classification performance at the image level. This leads to a question: *Can we further direct attention to regions of interest through natural language descriptions, such as 'A photo of a person'?* However, due to the resulting visual representation lacking fine-grained information necessary for distinguishing between identities, integrating CLIP straightforwardly into person ReID is non-trivial. Additionally, the query 'A photo of a person' presents a challenge due to the absence of specific descriptors, thereby lacking a personalized prompt for individual identification. The pioneering CLIP-ReID [21] introduces automated prompt engineering on CLIP by incorporating additional ID-wise learnable vectors customized for specific identities. Particularly, CLIP-ReID employs a two-stage training process that first optimizes the learnable vectors with the frozen CLIP model, and then restricts the image encoder with the learned textual descriptions. However, the disentangled usage, i.e., only the visual embedding is utilized during inference, renders the learned soft prompts ineffective for unseen prompts. As a result, the attention regions potentially do not entirely encompass the body part, and may inadvertently include background elements, such as cars and additional pedestrians captured in the scene, as illustrated in the first two examples in Fig. 1(b). In addition, even though CLIP-ReID adheres to training objectives aimed at vision-language alignment, such predefined soft prompts may not be sufficient to characterize the entire visual context of the specified pedestrian.

In this paper, we propose **Prompt-**driven **S**emantic **G**uidance (**PromptSG**), that aims to streamline the two-stage pipeline by leveraging the foundational CLIP model effectively and efficiently. As outlined in Fig. 2, our core insight is straightforward: we strive to activate CLIP's cross-modal comprehension using explicit language prompts, and the regions extracted can then be fine-tuned to enhance semantic discriminativeness. Specifically, given a textual prompt, we refine the patch tokens by injecting cross-attention maps, determining which patch attends to the corresponding semantics. Following this rationale, we revisit the fundamental issue that the term 'person' serves as a coarse descriptor, lacking personalized language descriptions for individual identities. Beyond semantic information related to the 'person', appearance information is also crucial for identification purposes [5]. While semantic information aids the model in better body part localization, appearance information further refines the focus on an individual's attire. Hence, we employ the *textual inversion* technique [10], which learns to represent visual context through unique token. We use a lightweight inversion network that maps the image to a pseudo-token. This pseudo-token can then be incorporated into the textual prompt, creating an embedding that closely mirrors the original image. Compared to CLIP-ReID, our solution offers two primary advantages: 1) The textual prompt emphasizes regions in the image via a cross-attention map, capturing the precise semantic part (Fig. 1(c)), and can also be utilized for unseen classes during inference. 2) The model can learn the personal token of the query image in an end-to-end manner, providing more detailed guidance specific to an identity (Fig. 1(d)). **Importantly, our proposed method is free, i.e. there is no need to supply additional information, such as masks, bounding boxes, or precise descriptions.** We summarize the contribution of this paper as follows.

- Leveraging the exceptional multi-modal reasoning capabilities of CLIP, we propose PromptSG, a novel framework for the person ReID task. This approach uniquely utilizes language prompts, providing explicit assistance to the visual encoder in efficiently capturing semantic information.
- To create a more personalized description for the individual, we propose learning to represent the specific, more detailed appearance attributes, by employing the inversion network.
- Without any additional labelling efforts, PromptSG surpasses previous SOTA method [21] by over 10% on the MSMT17 dataset. It also exhibits superior performance on the Market-1501 benchmark, surpassing previous SOTA method [46] by nearly 5%.

2. Related Work

Person Re-identification remains an important yet challenging task due to the subtle inter-class differences. To learn more discriminative representations, a category of CNN-based techniques has primarily concentrated on optimizing the distance metric via metric learning [15, 33, 34, 37, 38]. Recognizing the importance of semantic information, a substantial body of research [3, 23, 31, 43] explores the use of attention mechanisms, which guide the network to extract attention-aware features for body parts. For example, AAnet [36] adopts a unified learning framework that incorporates attribute attention maps through extra attribute labels. Pioneering work TransReID [14] introduces a self-attention-based architecture, Vision Transformer (ViT) [8],

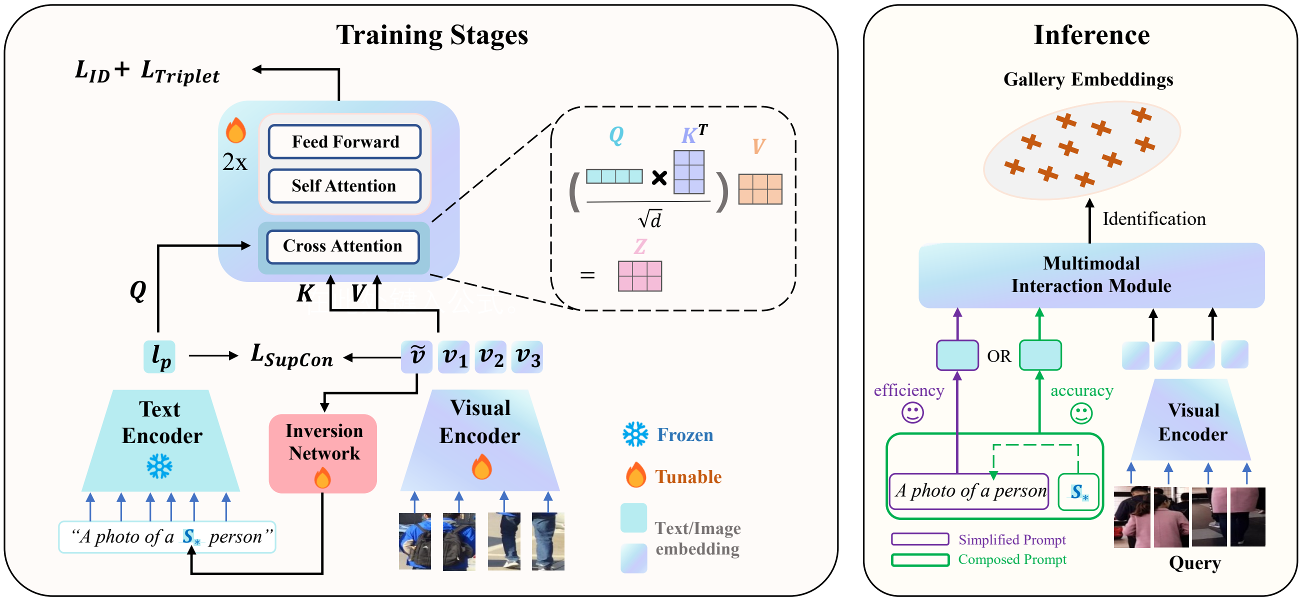

Figure 3. Overview of our framework. PromptSG learns pseudo token S_* from the specific visual embedding, and the visual encoder learns semantic faithful representations with the guidance of language prompts that occur in the Multimodal Interaction Module.

for advancing ReID tasks. DCAL [56] proposes to implicitly extract the local features through a global-local cross-attention mechanism. However, these methods solely apply attention mechanisms to the visual modality, and the lack of explicit language guidance potentially constrains their performance. The work most relevant to ours, CLIP-ReID [21], is the first to utilize vision-language pre-training model CLIP in ReID task. However, CLIP-ReID fails to leverage the linguistic capability of the text encoder in CLIP during inference, since the ID-specific learnable tokens only influence the seen identities.

Large-scale vision-language pre-training model connects the image representation with text embedding in a shared embedding space, has demonstrated effectiveness across a wide range of uni-modal and multimodal downstream tasks. These include classification [6, 48], image captioning [25], and cross-modal retrieval [11, 16, 32, 42, 49]. Foundational VL models, such as CLIP, usually undergo training on extensive image-text pairs with contrastive learning objectives. This foundational pre-training provides the model with strong open-vocabulary classification capabilities. Inherited from prompt learning in NLP [18], CoOp [54] proposes to explore learnable prompt optimization on few-shot classification. Following this soft prompt approach, CLIP-ReID pioneers the adaptation of CLIP for person ReID by classifying images into ID-specific prompts. Differing from CLIP-ReID, which focuses on vision-language alignment, our goal is to exploit rich semantic information from language to explicitly control the weights assigned to each patch or region, and improve the two-stage framework by directly inverting images into the language latent space.

Textual Inversion, originally for personalized text-to-image generation [10], is a learning approach that aims to discover new pseudo-words in the word-embedding space. These pseudo-words are capable of encapsulating both the overall visual content and intricate visual details. Recently, the application of textual inversion has expanded to zero-shot composed image retrieval task [1, 29]. In these studies, a textual inversion network is typically pre-trained using extensive unlabeled image datasets. In this work, we stand out as the first to apply this learning paradigm to person ReID without any additional training data.

3. Preliminary

Contrastive Language-Image Pre-training (CLIP) undergoes pre-training on a large corpus of image-text pairs, aligning visual and linguistic representations within a shared space through the matching of images with their corresponding text descriptions. Specifically, CLIP consists of a visual encoder $\mathcal{V}(\cdot)$ and a text encoder $\mathcal{T}(\cdot)$. The visual encoder $\mathcal{V}(\cdot)$ takes an image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$ as input. The text encoder $\mathcal{T}(\cdot)$ takes a tokenized textual description $\boldsymbol{t} \in \mathbb{R}^{N\times D}$ as input, where N,D are the text's length and token feature dimension respectively. The pre-training objective is based on self-supervised contrastive learning, which minimizes cosine distance for matched image-text pairs. For the downstream tasks such as classification, the description of j-th class is typically obtained through the hand-crafted prompt, *e.g.,* 'A photo of a [CLASS]'. Therefore, the probability of image $\boldsymbol{x}$ being classified as class y can be computed as follows:

$$\mathcal{P}(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\mathcal{V}(\boldsymbol{x}), \mathcal{T}(\boldsymbol{t_y}))/\tau)}{\sum_{j=1}^{K}\exp(\text{sim}(\mathcal{V}(\boldsymbol{x}), \mathcal{T}(\boldsymbol{t_j}))/\tau)}. \tag{1}$$

where τ denotes the temperature, and $\text{sim}(a, b) = \frac{a \cdot b}{\|a\|_2 \|b\|_2}$ is the cosine similarity.

A simple approach to applying CLIP to person ReID involves substituting the linear classifier with image-to-text classification. However, given that labels in ReID tasks are solely index-based, there are no specific words to represent different persons. To tackle this challenge, CLIP-ReID crafts the prompt as 'A photo of a $[X_i]_1[X_i]_2[X_i]_3...[X_i]_M$ person', where $[X_i]_m$, $m \in \{1, ..., M\}$ represents a set of ID-specific learnable tokens for the i-th ID. Nevertheless, CLIP-ReID optimizes ID-specific prompts exclusively bound to training IDs, it overlooks the chance to fully exploit the open-vocabulary capabilities inherent in CLIP.

4. Method

An overview of our framework is depicted in Fig. 3. Starting with the visual embeddings derived from CLIP's visual encoder, our approach employs an inversion network to learn pseudo tokens that encapsulate the visual context. Following this, an interaction between visual and textual modalities is facilitated in the interaction module, leading to the final re-weighted representations. During the inference phase, we are presented with two options for textual inputs: an efficiency-driven simplified prompt and an accuracy-driven composed prompt. Note that the text encoder is frozen in our entire framework.

4.1. Learning the Personalized ID-Specific Prompt

As suggested by prior research, the word-embedding space possesses sufficient expressiveness to encapsulate basic image concepts [7]. However, the inherent limitation lies in the pre-defined prompts in CLIP-ReID, which can only capture limited attributes and may not fully encapsulate the visual context. Contrarily, we propose learning the pseudo token by textual inversion technique that aligns with the context of the query image.

Let $f_\theta(\cdot)$ denote an inversion network parameterized by θ, our goal is to invert the global visual embedding $\boldsymbol{v}$ from visual space of CLIP, represented as $\boldsymbol{v} \in V$, into a pseudo token $s_* \in T_*$ by $f_\theta(\boldsymbol{v}) = s_*$, where T_* indicates the token embedding space. Subsequently, this pseudo token can be integrated into natural language sentences. As such, the language prompt for the input image is structured as 'A photo of a s_* person'. It is worth noting that this pseudo-token bears no relationship to an actual word but functions as a representation in the token embedding space. An input language prompt undergoes a tokenization process, resulting in several tokens. The tokenized prompt, denoted as $\boldsymbol{t_p}$, can be fed into the text encoder of CLIP to obtain text embedding $\boldsymbol{l_p} = \mathcal{T}(\boldsymbol{t_p})$. To ensure that the learned pseudo-token effectively tells the context of the image, one can follow to the reconstruction objective of textual inversion by the symmetric contrastive loss, which is formulated as follows:

$$\mathcal{L}_{i2t} = \frac{1}{N}\sum_{n=1}^{N} \log \frac{\exp(\text{sim}(\boldsymbol{v_n}, \boldsymbol{l_p})/\tau)}{\sum_{i=1}^{N}\exp(\text{sim}(\boldsymbol{v_n}, \boldsymbol{l_i})/\tau)}, \tag{2}$$

$$\mathcal{L}_{t2i} = \frac{1}{N}\sum_{n=1}^{N} \log \frac{\exp(\text{sim}(\boldsymbol{l_n}, \boldsymbol{v_p})/\tau)}{\sum_{i=1}^{N}\exp(\text{sim}(\boldsymbol{l_n}, \boldsymbol{v_i})/\tau)}. \tag{3}$$

In this context, $\boldsymbol{v_i}$ or $\boldsymbol{l_i}$ represents the i-th image/text embedding in a batch. $\boldsymbol{l_p}$ is the corresponding prompt embedding for $\boldsymbol{v_n}$ and is constructed in a manner analogous to $\boldsymbol{v_p}$.

The underlying mechanism is grounded in the principle of cycle-consistency, wherein a pseudo token tends to faithfully represent the context of the image only when the text features closely align with corresponding image features. However, the contrastive loss fails to handle cases where images with the same ID are supposed to share the same appearance. Therefore, we aim to encourage the pseudo token to capture visual details exclusive to the same identity. To this end, we exploit the symmetric supervised contrastive loss as follows:

$$\mathcal{L}_{\text{SupCon}} = \mathcal{L}_{i2t}^{\text{Sup}} + \mathcal{L}_{t2i}^{\text{Sup}}. \tag{4}$$

$$\mathcal{L}_{i2t}^{\text{Sup}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{p^+ \in P(i)} \log \frac{\exp(\text{sim}(\boldsymbol{v_n}, \boldsymbol{l_{p^+}})/\tau)}{\sum_{i=1}^{N}\exp(\text{sim}(\boldsymbol{v_n}, \boldsymbol{l_i})/\tau)}, \tag{5}$$

$$\mathcal{L}_{t2i}^{\text{Sup}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{p^+ \in P(i)} \log \frac{\exp(\text{sim}(\boldsymbol{l_n}, \boldsymbol{v_{p^+}})/\tau)}{\sum_{i=1}^{N}\exp(\text{sim}(\boldsymbol{l_n}, \boldsymbol{v_i})/\tau)}. \tag{6}$$

where $P(i)$ represents the positive samples related to v_n, l_n.

4.2. Prompt-driven Semantic Guidance

We refine the language prompt by incorporating the pseudo-token that is linked to the identity, enhancing its ability to convey a more specific visual context for the image. Our commitment extends to meticulously directing the image feature through language. At the core of our approach lies the idea of semantic guidance, wherein we explicitly determine which region of the image aligns with the language prompt. Intuitively, image patches corresponding to the semantic "person" should inherently have substantial influence to facilitate discrimination. As opposed to the interaction between patches in self-attention layers within a single modality. Based on this observation, we explore a patch-to-prompt interaction that occurs in multi-modality.

In particular, we employ a language-guided cross-attention module, which uses the textual embedding as query and the patch-wise embedding of the visual encoder as key and value. More formally, given a pair of image and prompt $(\boldsymbol{x}, \boldsymbol{t_p})$, we first feed the image $\boldsymbol{x}$ into the visual encoder, yielding in a sequence of patch embeddings $\{\tilde{\boldsymbol{v}}, \boldsymbol{v_1}, ..., \boldsymbol{v}_M\}$. Here, $\tilde{\boldsymbol{v}}$ denotes the global visual embedding, while remaining $\boldsymbol{v_i}, i \in [1, M]$ belong to the local patch embeddings. In a similar vein, the prompt is fed into the text encoder to derive the text embedding $\boldsymbol{l_p}$. Subsequently, the text embedding is projected onto a query matrix Q and patch embeddings are projected to a key matrix K and a value matrix V, via three different linear-projection layers. As such, the patch-to-prompt interaction can be achieved by:

$$\mathrm{A}(Q, K, V) = \mathrm{Softmax}(\frac{QK^T}{\sqrt{d}})V. \tag{7}$$

This interaction aggregates the attention map to highlight the regions of high semantic response. Drawing from multi-modal fusion methods [9], we incorporate two transformer blocks following the cross-attention layer to derive final representations. Ultimately, we utilize the standard ReID loss, i.e., the triplet loss and identity classification loss [12], to optimize our framework.

$$\mathcal{L}_{\text{ID}} = \frac{1}{K}\sum_{j=1}^{K} y_j \log p_j, \tag{8}$$

$$\mathcal{L}_{\text{Triplet}} = \max(d_p - d_n + m, 0), \tag{9}$$

where p_j is the prediction probability for the j-th class, and m denotes the margin.

4.3. Optimization and Inference

Taining optimization. In summary, the overall objective function for our framework is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{Triplet}} + \mathcal{L}_{\text{ID}} + \lambda\mathcal{L}_{\text{SupCon}}. \tag{10}$$

Similar to CLIP-ReID, the final hidden states of the vision transformer, in conjunction the preceding two layer states, are also employed to calculate $\mathcal{L}_{\text{Triplet}}$.

Improved inference efficiency. Our approach involves the query-specific pseudo token with the textual prompt, which essentially doubles the inference time compared to using only the visual encoder. Fortunately, our empirical findings suggest that providing only 'A photo of a person' as a simplified guideline yields comparable results. In this way, there will be no increase in the inference time caused by the text encoder.

Dataset	#ID	Images	Cams
Market-1501	1,501	32,668	6
MSMT17	4,101	126,441	15
DukeMTMC	1,404	36,411	8
CHUK03-NP	1,467	13,164	2

Table 1. The statistics of dataset in our experiments

5. Experiments

5.1. Experimental Setting

Datasets and Evaluation Protocols. To evaluate and compare various methods, four extensive person re-identification datasets Market-1501 [51], MSMT17 [41], DukeMTMC [52] and CUHK03-NP [22] are exploited. Dataset stats are in Tab. 1. In line with conventions in the ReID community [13], two commonly used metrics, *i.e.*, mean Average Precision (mAP) and Rank-1(R-1) accuracy, are used to evaluate the performance.

Implementation Details. In alignment with prior research, we employ both ResNet-50 and ViT-B/16 Pre-trained from CLIP as our visual encoder and a pre-trained text encoder, *i.e.*, CLIP text Transformer. Our framework additionally features a random-initialized inversion network and a multimodal interaction module. The inversion network is a lightweight model employing a three-layered MLP of 512-dimensional hidden state. A Batch Normalization (BN) layer [17] is placed after the last state of the network. The batch size is configured to 64, encompassing 16 identities with 4 images per identity. All input images are resized to 256 × 128. We use the Adam optimizer with a learning rate of 5e-6 for the visual encoder, whereas the learning rate for random-initialized modules is set to 5e-5. We find λ in Eq. (10) is not sensitive and performs well across a broad range, thus we consistently set $\lambda = 0.5$ for all datasets. The model is trained for 60 epochs, with a learning rate decay factor of 0.1 for every 20 epochs. The entire framework is implemented using PyTorch and runs on a single NVIDIA RTX3090 GPU with 24GB VRAM.

Baseline. Most existing approaches are built upon the strong ReID baseline presented in [24]. Specifically, they employ an ImageNet-21k pre-trained CNN model or ViT as the backbone and incorporate ID loss and triplet loss as crucial components. In contrast, our baseline model deviates by leveraging the pre-trained CLIP model and we fine-tune the visual encoder of CLIP by directly applying the two commonly-used losses.

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

We benchmark PromptSG against the current state-of-the-art, which can generally be divided into three categories: CNN-based, ViT-based, and CLIP-based methods. Tab. 2 summarizes the main results on four widely

Backbone	Method	Reference	Market-1501		MSMT17		DukeMTMC		CUHK03-NP	
			mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
ResNet50	*CNN-based method*									
	OSNeT [53]	ICCV'19	84.9	94.8	52.9	78.7	73.5	88.6	-	-
	ISP [57]	ECCV'20	84.9	94.2	-	-	75.6	86.9	74.1	76.5
	RGA-SC [50]	CVPR'20	88.4	96.1	57.5	80.3	-	-	77.4	81.1
	CDNet [20]	CVPR'21	86.0	95.1	54.7	78.9	76.8	88.6	-	-
	CAL [28]	ICCV'21	87.0	94.5	56.2	79.5	76.4	87.6	-	-
	ALDER* [47]	TIP'21	88.9	95.6	59.1	82.5	78.9	89.9	78.7	81.0
	LTReID* [39]	TMM'22	86.9	94.7	58.6	81.0	80.4	90.5	80.3	82.1
	CLIP-based method									
	Baseline		88.1	94.7	60.7	82.1	79.3	88.6	77.6	79.1
	CLIP-ReID [21]	AAAI'23	89.8	95.7	63.0	84.4	80.7	90.0	78.2	79.4
	PromptSG	Ours	91.8	96.6	68.5	86.0	80.4	90.2	79.8	80.5
ViT-B/16	*ViT-based method*									
	TransReID [14]	ICCV'21	88.9	95.2	67.4	85.3	82.0	90.7	79.6	81.7
	DCAL [56]	CVPR'22	87.5	94.7	64.0	83.1	80.1	89.0	-	-
	AAformer [58]	TNNLS'23	88.0	95.4	65.6	84.4	80.9	90.1	79.0	80.3
	PHA [46]	CVPR'23	90.2	96.1	68.9	86.1	-	-	83.0	84.5
	CLIP-based method									
	Baseline		86.4	93.3	66.1	84.4	80.0	88.8	80.0	80.5
	CLIP-ReID [21]	AAAI'23	89.6	95.5	73.4	88.7	82.5	90.0	81.6	80.9
	PromptSG	Ours	94.6	97.0	87.2	92.6	81.6	91.0	83.1	85.1

Table 2. Comparison with the state-of-the-art models on Market-1501, MSMT17, DukeMTMC, and CUHK03-NP (labeled) datasets. The superscript star* indicates that the image is resized to a resolution exceeding 256x128. All results are reported without re-ranking. Color Red and blue: the best and second-best results.

used person ReID datasets. We observe that our proposed PromptSG attains the best results and sets a new state-of-the-art performance. Remarkably, PromptSG achieves over 10% improvement on MSMT17 and nearly 5% on Market-1501, surpassing previous state-of-the-art results.

Compared with ViT-based method. Pioneering work TransReID [14] sets a strong baseline for the ViT-based method by leveraging the potentials of the transformer. Building upon this groundwork, PHA [46] further enhances the preservation of key high-frequency elements in images. In contrast to existing ViT-based methods that only capture the patch-wise uni-modal information, our PromptSG method demonstrates that the interaction of different modalities can improve the performance of individual modalities.

Compared with CLIP-based method. Compared with the competing CLIP-based method CLIP-ReID, our PromptSG outperforms it by **5.0%/1.5%** and **13.8%/3.9%** mAP/Rank-1 on Market-1501 and MSMT17 datasets when taking ViT-B/16 as visual backbone. A key distinction between CLIP-ReID and our approach resides in the composition of the query-specific pseudo-token. Our results further underscore that incorporating textual information during the inference process can also enhance performance.

Compared with CNN-based method. To ensure a fair comparison, we also implement PromptSG with a ResNet-50 backbone. Apart from LTReID [39] that utilize higher resolution images, our method consistently surpasses other methods by a significant margin, especially on the most challenging person ReID dataset, MSMT17. This highlights the robustness and superiority of our approach across various architectures.

5.3. Ablation Study

In the following, we conduct an ablation study on the essential elements of PromptSG on Market-1501 and MSMT17 datasets, and all the experiments are conducted on the ViT-B/16 backbone.

Contributions from Different Components. To assess the contribution of various components, we conduct ablation experiments by removing one component at a time. Recall that $\mathcal{L}_{i2t}^{\mathrm{Sup}}$ and $\mathcal{L}_{t2i}^{\mathrm{Sup}}$ are the supervised contrastive losses in Eq. (2), Eq. (3) respectively, and MIM denotes the multimodal interaction module. Comparing rows b) and c) with a), we see a similar conclusion where the removal of text-to-image or image-to-text contrastive loss leads to a decent improvement on both datasets. Further comparing rows a) and d), we observe that the removal of semantic information leads to a larger decrease than solely removing ID-specific appearance information. Notably, as seen in row a), our full model, PromptSG, utilizes both semantic and appearance

	Components			Market-1501		MSMT17	
	$\mathcal{L}_{i2t}^{\text{Sup}}$	$\mathcal{L}_{t2i}^{\text{Sup}}$	MIM	mAP	R-1	mAP	R-1
a)	✓	✓	✓	94.6	97.0	87.2	92.6
b)		✓	✓	92.8	96.7	85.2	91.9
c)	✓		✓	93.0	96.7	84.5	90.2
d)	✓	✓		89.4	95.3	71.4	87.3

Table 3. Ablation study on the effectiveness of each component of PromptSG on Market-1501 and MSMT17.

Method	Market-1501		MSMT17	
	mAP	R-1	mAP	R-1
Training w/ composed	94.6	97.0	87.2	92.6
Training w/o composed	92.0	96.3	85.3	91.6

Table 4. Ablation of training with or without composing the pseudo token on Market-1501 and MSMT17.

Traning Model	Inference Model Visual Encoder	Text Encoder	FPS ↑	mAP
w/o attention module	✓	-	1x	89.4
+ 1 cross-layer	✓	-	0.95x	91.1
+ 1 cross & 1 self-layer	✓	-	0.91x	93.0
+ 1 cross & 2 self-layer	✓	✓	0.48x	94.6
+ 1 cross & 2 self-layer	✓	-	0.88x	94.1

Table 5. The impact of various interaction modules and efficiency comparison with different inference models on Market-1501. Cross and self means cross-attention and self-attention, respectively. FPS denotes the quantity of images processed by the model in one second.

Method	#Params	#Params %CLIP	Training Times ↓	
(a) Market-1501				
CLIP-ReID	89M	0.71	4689s	1x
PromptSG	94M	0.75	2417s	0.51x
(b) MSMT17				
CLIP-ReID	90M	0.73	12904s	1x
PromptSG	94M	0.75	6108s	0.47x

Table 6. Comparison of training times and the number of parameters on Market-1501 and MSMT17. #Params denotes the number of learnable parameters in the whole framework. All models are evaluated on a single 3090Ti GPU.

language supervision during training, achieving a substantial improvement of over a point. The overall conclusion supports that language guidance, through both semantic and appearance cues, plays a crucial role in improving the performance of our model.

Ablation Study on the Personalized Prompt. To better understand whether the learned pseudo tokens s_* can provide more granular guidance for learning visual embeddings, we train a strong baseline model, where the textual prompt dose not composed with the s_* during training and testing, but instead relies on the simplified prompt “A photo of a person” for semantic guidance. Note that we will not use the symmetric supervised contrastive loss in this case. Results in Tab. 4 imply that composing the s_* has a significant impact on the overall performance. When s_* is removed from the training process, the performance decreases by **1.9%** to **2.6%** in terms of mAP. Although we focus on the uni-modal re-identification task, the above formulation could potentially be applied to multimodal test sets, such as text-to-image person retrieval by composing the image feature with the text to achieve better alignment.

Ablation Study on the Interaction Module. We analyze the impact of different designs of the interaction module on performance and inference speed, as well as the impact of not using a composed prompt during inference. Notably, personalized prompts are consistently included during the training. As shown in Tab. 5, without an attention module (w/o attention module), the model achieves a baseline performance, with inference speed being dependent solely on the visual encoder. Introducing a single cross-attention layer (+1 cross-layer) shows a notable performance improvement, indicating the positive effect of incorporating a cross-layer design. Notably, performances can be stably improved with more self-attention layers, but at the cost of lower inference efficiency. Furthermore, our analysis illuminates the impact of employing a composed prompt during the inference phase, revealing that when we follow the same procedure as the training stage—composing text with query images—the Frames Per Second (FPS) is only 0.48 times that of the baseline. This is expected as we need to pass through two encoders for each query. However, we empirically discovered that using a fixed prompt “A photo of a person” for all queries may not lead to significant performance degradation, and it does not compromise efficiency. Therefore, one could opt for this version to achieve a more favorable balance between accuracy and efficiency.

Comparison of training efficiency. In order to showcase the efficiency of our proposed approach, we carry out a comparative analysis between our one-stage PromptSG and the two-stage CLIP-ReID method, focusing on the number of learnable parameters and training speed. The details of this comparison are provided in Tab. 6. In terms of training parameters, on top of CLIP, CLIP-ReID incorporates an additional of parameters mainly through the ID-wise learnable prompt, our approach primarily extends through a fixed-size mapping network and an interaction module. Despite CLIP-ReID having 2%-4% fewer parameters than ours on two datasets, it may experience continuous growth in parame-

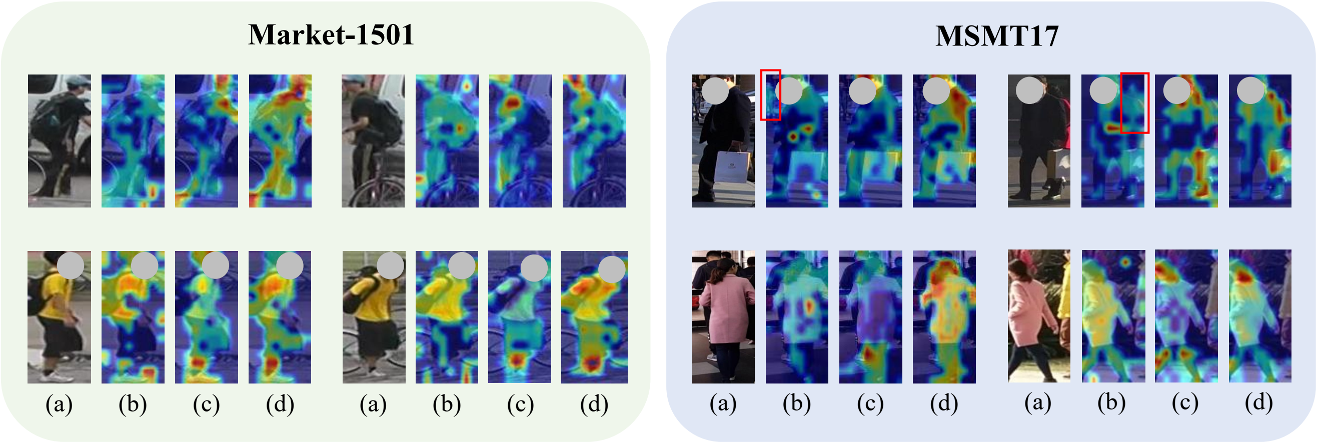

Figure 4. Transformer visualization of attention maps. (a) Original images, (b) CLIP-ReID, (c) PromptSG without composed training, (d) PromptSG. We see our method is effective in simultaneously focusing on the semantic clues and exploring more discriminative parts.

ters in scenarios with a higher number of classes or evolving dynamics. In contrast, PromptSG demonstrates stronger robustness in the number of parameters and achieves significant faster training speed. It can achieve a speedup of approximately 2 times faster during training compared to the two-stage method.

6. Qualitative Analysis

To have an intuitive understanding and validate the effectiveness of our method, we conduct a qualitative analysis where visualization of attention maps is presented in Fig. 4. Specifically, we exhibit examples from Market-1501 and MSMT17 datasets, each with two training images and two gallery images. **We carefully selected some challenging examples, including those with complex backgrounds or images depicting multiple individuals.** To gain a better insight into the regions of interest attended by the model in zero-shot scenarios, we do not use the common protocol GramCAM [30], as it needs the class-prediction scores and might be considered less suitable for Transformer-type backbones. Following [21], we use the Transformer-interpretability method in [2].

We compare our (d) PromptSG with (b) CLIP-ReID and (c) PromptSG without image-composed training. It can be seen that our method exhibits significant effectiveness, as it adeptly captures semantic information while also concentrating on more detailed appearance details. For example, in the **first row** of the Market-1501 dataset, the attention map of CLIP-ReID is susceptible to interference from background elements like “car”. On the other hand, PromptSG w/o composed training tends to emphasize semantic information related to the ‘person,’ focusing on the location of the head, arms, and legs. In contrast, our method goes beyond this by also exploring appearance features, such as identifying individuals wearing hats or carrying backpacks. Finally, in the **first row** examples of MSMT17, where additional pedestrians appear in the image, our method excels in effectively filtering out unnecessary pedestrians, while CLIP-ReID fails.

7. Conclusion

In this paper, we propose PromptSG, a simple yet effective framework that exploits the foundational model CLIP for the person ReID task. We show that language guidance is an effective way to adapt pre-trained multimodal models for the uni-modal retrieval tasks. Through leveraging the aligned multi-modal latent space provided by CLIP, the textual prompt ”A photo of a person” can naturally address the challenge of the visual encoder in its struggle to capture semantic information. To probe more fine-grained appearance features, we incorporate an inversion network to learn pseudo tokens that describe the image context.

Discussion and Limitation. Despite the considerable potential of language prompt learning in ReID tasks, prompt learning in the vision branch remains a largely untapped area. Fine-tuning the visual encoder for strong supervised performance may lead to poor zero-shot generalization. We hope our work can inspire future research on fully unleashing the potential of large foundation models in challenging ReID tasks.

Acknowledgments This work was supported by the National Key R&D Program of China under Grant 2022YFB3103500, the National Natural Science Foundation of China under Grants 62106258, 62006242 and 62202459, and the China Postdoctoral Science Foundation under Grant 2022M713348 and 2022TQ0363, and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001) and BAST (NO.BYESS2023304).

References

[1] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2023. 3

[2] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021. 1, 8

[3] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2019. 2

[4] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2019. 1

[5] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu 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 (CVPR)*, 2023. 2

[6] Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang. Area: adaptive reweighting via effective area for long-tailed classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[7] Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 4

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[11] Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. Dual alignment unsupervised domain adaptation for video-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023. 3

[12] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 5

[13] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 2023. 5

[14] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2021. 1, 2, 6

[15] Pingting Hong, Dayan Wu, Bo Li, and Weipinng Wang. Camera-specific informative data augmentation module for unbalanced person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022. 2

[16] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning (ICML)*, 2015. 5

[18] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics (ACL)*, 2020. 3

[19] Dengjie Li, Siyu Chen, Yujie Zhong, Fan Liang, and Lin Ma. Dip: Learning discriminative implicit parts for person re-identification. *arXiv preprint arXiv:2212.13906*, 2022. 1

[20] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021. 6

[21] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 2, 3, 6, 8

[22] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2014. 1, 5

[23] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 2

[24] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei 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 (CVPRW)*, 2019. 2, 5

[25] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, 2021. 2

[27] Haocong Rao and Chunyan Miao. Transg: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[28] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6

[29] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017. 8

[31] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 2

[32] Qinghang Su, Dayan Wu, Chenming Wu, Bo Li, and Weiping Wang. From data to optimization: Data-free deep incremental hashing with data disambiguation and adaptive proxies. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2024. 3

[33] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin 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)*, 2018. 2

[34] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020. 2

[35] Lei Tan, Pingyang Dai, Rongrong Ji, and Yongjian Wu. Dynamic prototype mask for occluded person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022. 1

[36] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019. 2

[37] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia (ACM MM)*, 2018. 2

[38] Lin Wang, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. Attack is the best defense: Towards preemptive-protection person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022. 2

[39] Pingyu Wang, Zhicheng Zhao, Fei Su, and Honying Meng. Ltreid: Factorizable feature generation with independent components for long-tailed person re-identification. *IEEE Transactions on Multimedia (TMM)*, 2022. 6

[40] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 1

[41] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 5

[42] Dayan Wu, Qi Dai, Jing Liu, Bo Li, and Weiping Wang. Deep incremental hashing network for efficient image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019. 3

[43] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019. 2

[44] Zexian Yang, Dayan Wu, Wanqian Zhang, Bo Li, and Weipinng Wang. Handling label uncertainty for camera incremental person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 2023. 1

[45] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2021. 1

[46] Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6

[47] Quan Zhang, Jianhuang Lai, Zhanxiang Feng, and Xiaohua Xie. Seeing like a human: Asynchronous learning with dynamic progressive refinement for person re-identification. *IEEE Transactions on Image Processing (TIP)*, 2021. 6

[48] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[49] Wanqian Zhang, Dayan Wu, Yu Zhou, Bo Li, Weiping Wang, and Dan Meng. Binary neural network hashing for image retrieval. In *Proceedings of the 44th international ACM SIGIR*

conference on research and development in information retrieval (SIGIR)l, 2021. 3

[50] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person reidentification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[51] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2015. 5

[52] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017. 5

[53] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person reidentification. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2019. 6

[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 3

[55] Xiao Zhou, Yujie Zhong, Zhen Cheng, Fan Liang, and Lin Ma. Adaptive sparse pairwise loss for object reidentification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[56] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi 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 (CVPR)*, 2022. 3, 6

[57] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, 2020. 1, 6

[58] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2023. 6