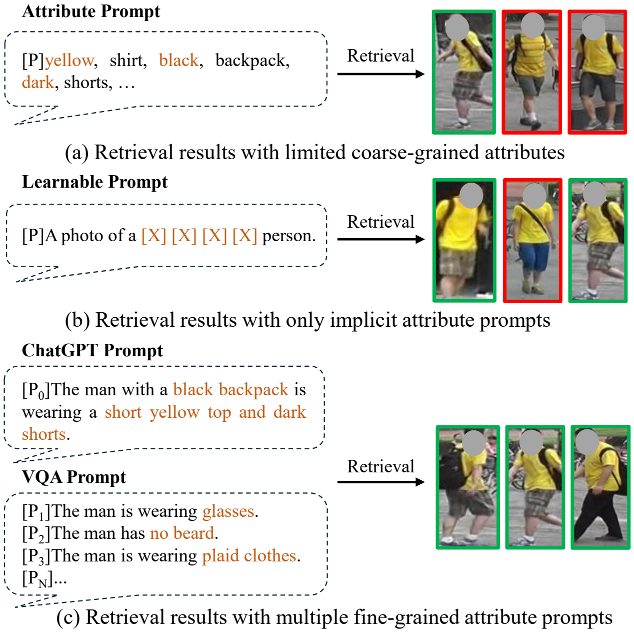
Multi-Prompts Learning with Cross-Modal Alignment for Attribute-Based Person Re-identification

Yajing Zhai[1,2*], Yawen Zeng[1*], Zhiyong Huang[3], Zheng Qin[1†], Xin Jin[2†], Da Cao[1]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[2]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3]National University of Singapore, NUS Research Institute in Chongqing
{yajingzhai9,yawenzeng11}@gmail.com, huangzy@comp.nus.edu.sg,
zqin@hnu.edu.cn, jinxin@eitech.edu.cn, caoda0721@gmail.com

Abstract

The fine-grained attribute descriptions can significantly supplement the valuable semantic information for person image, which is vital to the success of person re-identification (ReID) task. However, current ReID algorithms typically failed to effectively leverage the rich contextual information available, primarily due to their reliance on simplistic and coarse utilization of image attributes. Recent advances in artificial intelligence generated content have made it possible to automatically generate plentiful fine-grained attribute descriptions and make full use of them. Thereby, this paper explores the potential of using the generated multiple person attributes as prompts in ReID tasks with off-the-shelf (large) models for more accurate retrieval results. To this end, we present a new framework called Multi-Prompts ReID (MP-ReID), based on prompt learning and language models, to fully dip fine attributes to assist ReID task. Specifically, MP-ReID first learns to hallucinate diverse, informative, and promptable sentences for describing the query images. This procedure includes (i) explicit prompts of which attributes a person has and furthermore (ii) implicit learnable prompts for adjusting/conditioning the criteria used towards this person identity matching. Explicit prompts are obtained by ensembling generation models, such as ChatGPT and VQA models. Moreover, an alignment module is designed to fuse multi-prompts (i.e., explicit and implicit ones) progressively and mitigate the cross-modal gap. Extensive experiments on the existing attribute-involved ReID datasets, namely, Market1501 and DukeMTMC-reID, demonstrate the effectiveness and rationality of the proposed MP-ReID solution.


Attribute Prompt
[P]yellow, shirt, black, backpack, dark, shorts, …
Retrieval
Learnable Prompt
[P]A photo of a [X] [X] [X] [X] person.
Retrieval
ChatGPT Prompt
[P$_0$]The man with a black backpack is wearing a short yellow top and dark shorts.
VQA Prompt
[P$_1$]The man is wearing glasses.
[P$_2$]The man has no beard.
[P$_3$]The man is wearing plaid clothes.
[P$_N$]...
Retrieval

(a) Retrieval results with limited coarse-grained attributes

(b) Retrieval results with only implicit attribute prompts

(c) Retrieval results with multiple fine-grained attribute prompts

Figure 1: Comparison of various usages of attributes for person ReID, where red boxes represent negative images, green boxes indicate positive results and key attribute words have been marked in red. We can see that using multiple fine human attributes as prompts in ReID brought advancements.

Introduction

Person re-identification (ReID) is a challenging task due to the dramatic visual appearance changes from pose, viewpoints, illumination, occlusion, low resolution, background clutter, etc. (Jin et al. 2020a; Ye et al. 2021; Zhang, Zhang, and Liu 2021). Fine-grained person attributes are robust to these variations and are often exploited as efficient supplements with local descriptions that aid in the learning of more discriminative feature representations (Jia et al. 2022; Wang et al. 2022a). In particular, the common attributes include clothing color, shoes, hairstyle, gender, age, and other specific characteristics. They serve as additional information that complements and aligns images, reducing the impact of the above factors, thereby improving the overall performance of ReID (Yu et al. 2022).

Recently, some researchers (Jia, Chen, and Huang 2021; Niu et al. 2022; Specker, Cormier, and Beyerer 2023; Zheng et al. 2022) begin to investigate the importance of attributes w.r.t ReID task, which demonstrate that attributes are in-

[*] These authors contributed equally. This work was done when Yajing Zhai was an intern at Eastern Institute of Technology, Ningbo, China.

[†] Corresponding authors.

deed an effective piece of information that could enhance the retrieval performance in ReID. However, current attribute-based ReID algorithms fail to leverage the full potential of the abundant contextual information available. That's mainly because they rely on simplistic and naive utilization of coarse-grained attributes, as well as the complexity of accurately capturing and descriptions with the limitation of AI technology in the past. As shown in Figure 1a, certain coarse, separate, and ambiguous attributes, such as "yellow", "shorts", and "shirt", are directly used to retrieve pedestrians, which are less effective compared to clear and complete abundant contextual descriptions as presented in Figure 1c. Thus, it is essential but has not been well investigated to efficiently take full advantage of fine-grained attribute information for improving ReID accuracy.

With the fast-development of large models (Fu et al. 2022; Jin et al. 2022; Li et al. 2023), ReID methods gradually become more practical for real-world scenarios and gain superior performance. Besides, prompt learning (Wu et al. 2022; Zhou et al. 2022b), as a paradigm of strategies that leverages pre-trained models by incorporating additional textual description information, has achieved improved performance in many complex AI tasks (Zeng 2022; Lüddecke and Ecker 2022; Liu et al. 2022). Building upon this inspiration, we investigate the feasibility of utilizing prompts to provide fine-grained attribute information for the ReID task.

Intuitively, there are two strategies for applying attributes as prompts, explicit attribute prompts and implicit attribute prompts, as shown in Figure 1. (i) Explicit attribute prompts refer to an attribute-based sentence prompt generation method, where the production process utilizes some attribute words, among which ChatGPT and visual question answer (VQA) models (Yu et al. 2019; Wang et al. 2022b) with stronger interactivity and feedback mechanism are often used. (ii) While implicit attribute prompts use a learnable textual prompt generation method, where the process does not have specific attribute information, as depicted in Figure 1b. We can see that, the better retrieval result is obtained via the implicit attribute prompt method, but it is still not accurate enough. In contrast, as shown in Figure 1c, the ReID scheme that learning from multiple attribute prompts significantly improves the retrieval performance with more fine-grained information. From this we can infer that, the utilization of fine attribute information could enable the ReID model to learn more auxiliary features and relationships, thereby improving ultimate accuracy.

However, prominent challenges still remain that need to be further addressed. Firstly, the lack of such a required prompt-related ReID dataset in the large-scale practical ReID task has led to few studies have been exploited. The second challenge is that there is a gap between the attribute-based text prompt and the image, making it essential to address the alignment of these two modalities. As a result, despite utilizing rich prompts for improved ReID performance is a promising approach that can lead to efficient and comprehensive results, it remains an under-explored area with the potential for further optimization.

In this paper, we make the first attempt to employ the large-scale multi-prompts information in the attribute-based ReID task and propose a novel **M**ulti-**P**rompts Learning framework, dubbed as **MP-ReID**, to support this challenging task. MP-ReID aims to retrieve one person based on a variety of fine-grained attribute information as a complement for image information to improve the retrieval performance with ChatGPT, VQA and CLIP (Radford et al. 2021). As mentioned above, the multi-prompts include explicit attribute prompts and implicit attribute prompts. 1) **Explicit attribute prompts** — a prompt sentence generation paradigm, which is ensembling generation models based on attribute words. 2) The other is **implicit attribute prompts** — a learnable prompt paradigm without intuitive attributes, which models a prompt's context words with learnable vectors, that could be initialized with either random values or pre-trained word embeddings. Furthermore, image information with the promptable semantic feature is optimized under a cross-modal space to mitigate the cross-modal semantic gap. After that, the learned prompts are regarded as a booster to apply to the person retrieval. By conducting experiments on two well-known datasets, we validate that MP-ReID is superior to various existing methods. The main contributions of this work are summarized as follows:

- To the best of our knowledge, this is the first attempt that introduces the concept of multi-prompts learning strategies to generate diverse, informative, and promptable sentences for ReID improvements.
- We introduce two prompts generation paradigms: explicit attribute prompt and implicit attribute prompt, for applying fine-grained attributes to fully use the comprehensive semantics and enhance the retrieval performance.
- We contribute a Multi-Prompts ReID framework, dubbed MP-ReID to mitigate the cross-modal semantic gap for this attribute-based ReID task. Meanwhile, we collect a prompt-related ReID dataset containing multiple attribute prompts about the same person, and we have released the dataset to facilitate the research community[1].

Related Work

Prompt Learning

In recent years, the use of prompt learning, which concerns providing suggestive information, has become a popular technique for incorporating knowledge in natural language processing problems (Petroni et al. 2019; Song et al. 2022; Jin et al. 2023). It involves adding language-specific instructions to the input text, enabling the pre-trained model to comprehend the downstream task and enhance the performance. ChatGPT[2] and GPT-4[3] offer tremendous opportunities to improve open-source large language models using instruction-tuning (Peng et al. 2023) and transfer to downstream tasks with powerful generalization (Zhang et al. 2023). Moreover, there has been a recent trend towards utilizing prompt learning for improving the quality of visual representations in vision-language models (Ju et al. 2022; Rao et al. 2022).

[1] https://github.com/zyj20/MPReID.

[2] https://openai.com/blog/chatgpt

[3] https://openai.com/product/gpt-4

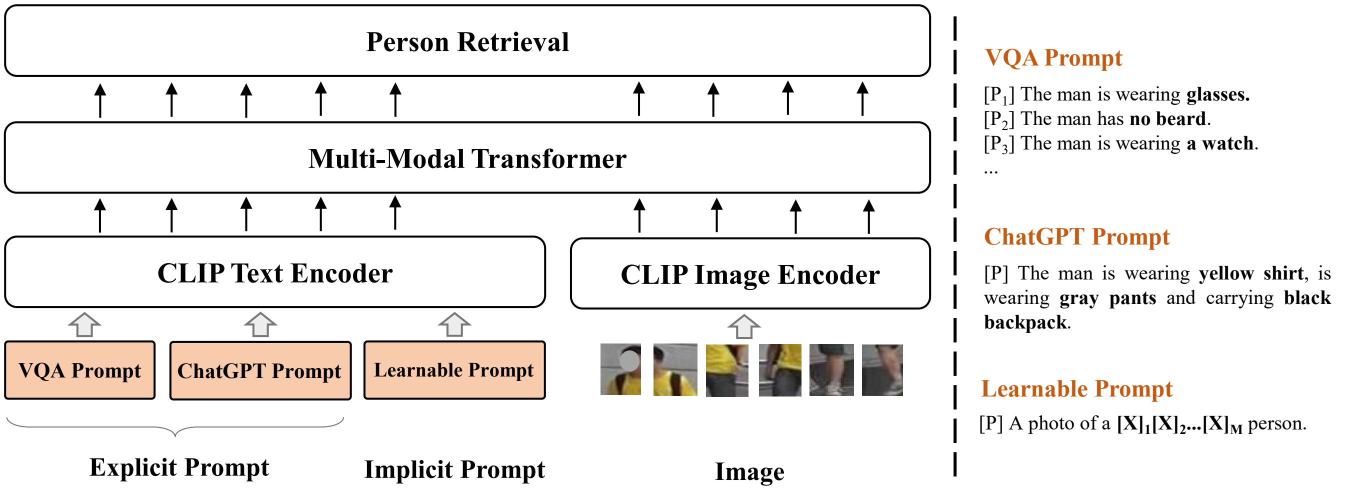

Person Retrieval
Multi-Modal Transformer
CLIP Text Encoder
CLIP Image Encoder
VQA Prompt
ChatGPT Prompt
Learnable Prompt
Explicit Prompt
Implicit Prompt
Image
VQA Prompt
[P_1] The man is wearing **glasses.**
[P_2] The man has **no beard**.
[P_3] The man is wearing **a watch**.
...
ChatGPT Prompt
[P] The man is wearing **yellow shirt**, is wearing **gray pants** and carrying **black backpack**.
Learnable Prompt
[P] A photo of a $[X]_1[X]_2...[X]_M$ person.

Figure 2: The graphical representation of MP-ReID for ReID. Under the prompt learning paradigm, the multi-prompts generated by ChatGPT and VQA are regarded as the textual input to the multi-modal Transformer, which can enhance the retrieval of the matching person images. It is built upon three components: 1) Multi-Prompts Generation Learning; 2) Cross-Modal Alignment; 3) Person Retrieval.

Attributed-based Person ReID

Recently, some deep learning methods are proposed to exploit the discriminative attributes. In particular, Lin et al. (Lin et al. 2019) manually labeled pedestrian attributes for the Market1501 dataset and the DukeMTMC-reID dataset[4]. Besides, the authors proposed a novel attribute-based person recognition framework with an attribute re-weighting module. This aims to learn discriminative embedding and correct prediction. Zhang et al. (Zhang, Niu, and Zhang 2020) leveraged the feature aggregation strategy to make use of attribute information. Jeong et al. (Jeong, Park, and Kwak 2021) presented a new loss for learning cross-modal embeddings in the context of attribute-based person search and regarded attribute dataset as a category of people sharing the same traits. Li et al. (Li, Sun, and Li 2023) fully exploited the cross-modal description ability through a set of learnable text tokens for each person ID and gave them to the text encoder to form ambiguous descriptions with a two-stage strategy, facilitating a better visual representation.

Inspired by the above work, we optimize the descriptive and visual features under the multi-prompts generation paradigm for ReID task, which contains explicit prompts and implicit prompts. In this way, textual prompts and visual features are learned from each other, achieving a win-win effect.

Methodology

This section provides a detailed explanation of our solution, with Figure 2 illustrating the overall framework of our MP-ReID. Generally speaking, our proposed framework comprises three components: multi-prompts generation, cross-model alignment, and person retrieval. 1). The approach of multi-prompts generation learning leverages ChatGPT, VQA, and learnable methods to generate three different prompts, which is given in Figure 3. These prompts are then fused together using a cross-attention mechanism; 2). Cross-modal alignment module aligns prompts-images pairs by feeding them into a multi-modal Transformer to learn the context; and 3). Person retrieval involves creating a feature representation in the prompt-visual space for identifying individuals.

Multi-Prompts Generation Learning

Given an attribute-based ReID dataset, images are defined as $\mathbb{M} = \{m_1, m_2, ..., m_n\}$, the corresponding attributes are denoted as $\mathbb{A} = \{a_1, a_2, ..., a_n\}$, respectively. The MP-ReID first generates prompts $\mathbb{P} = \{p_1, p_2, ..., p_n\}$, which contains P_i^e ensembling explicit prompts and P_i^l learnable implicit prompts. For visual representation and prompt representation, we adopt the image encoder and the text encoder from CLIP as the backbone for feature extractor respectively. They are all implemented as Transformer architecture (He et al. 2021b; Zhu et al. 2023). And the ViT-B/16 network architecture (Dosovitskiy et al. 2021) is utilized for the images, which contains 12 transformer layers. With respect to prompts embedding, we convert each word into a unique numeric ID using byte pair encoding with a $49,512$ vocab size (Sennrich, Haddow, and Birch 2016). To enable parallel computation, we set the context length of each text sequence to 77, including the start [SOS] and end [EOS] tokens. Within a batch of images, we denote the index of each image as $i \in \{1...N\}$. We calculate the similarity between the [CLS] token embedding of the image feature m_i and the corresponding [EOS] token embedding of the text feature p_i. And in this module, we obtain the image feature f_i^m,

$$f_i^m = F_m(m_i) \tag{1}$$

[4] https://vana77.github.io

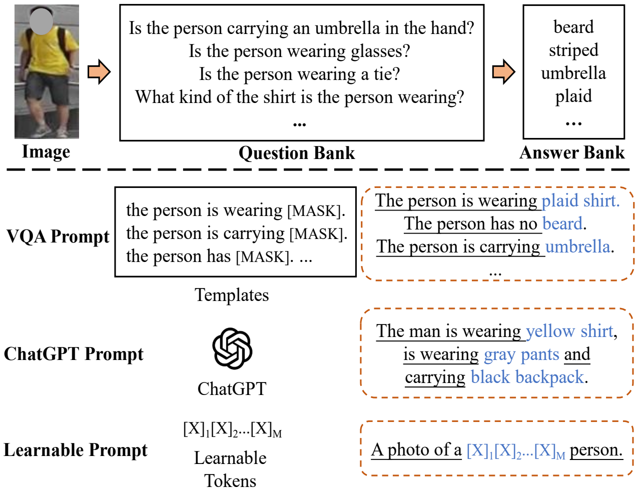

Is the person carrying an umbrella in the hand?
Is the person wearing glasses?
Is the person wearing a tie?
What kind of the shirt is the person wearing?
...
beard
striped
umbrella
plaid
...
Image
Question Bank
Answer Bank
VQA Prompt
the person is wearing [MASK].
the person is carrying [MASK].
the person has [MASK]. ...
Templates
The person is wearing plaid shirt.
The person has no beard.
The person is carrying umbrella.
...
ChatGPT Prompt
ChatGPT
The man is wearing yellow shirt, is wearing gray pants and carrying black backpack.
Learnable Prompt
[X]1[X]2...[X]M
Learnable Tokens
A photo of a [X]1[X]2...[X]M person.

Figure 3: The process of multi-prompts generation learning in the proposed MP-ReID framework. The top is the question bank and answer bank of VQA model; and the bottom is the concrete multi-prompts generation from the VQA prompt, ChatGPT prompt and learnable prompt, respectively.

Accordingly, for textual representation, we obtain the prompt feature f_i^p, which is formally formulated as,

$$f_i^p \leftarrow \begin{cases} f_i^e = F_p(p_i^e) \\ f_i^l = F_p(p_i^l) \end{cases} \tag{2}$$

where $F_m(\cdot)$ and $F_p(\cdot)$ are visual and textual projection function. Besides, f_i^m, f_i^e, f_i^l are extracted image features, explicit prompt features and implicit prompt features, respectively. P_i^e ensembling explicit prompts comprise P_i^c ChatGPT generation prompt and P_i^v VQA generation prompt.

1) Explicit attribute prompts. Specifically, to generate explicit attribute prompts, we adopt a prompt ensembling strategy that utilizes both ChatGPT and VQA models. Firstly, we establish the criteria and guidelines for generating prompt sentences that align with our desired outcome. We have configured it to utilize the specified instructions for sentence creation. This approach necessitates the usage of attribute words to generate prompts.

Subsequently, we transmit these attribute words to ChatGPT, which leverages its large model pre-training prompt learning to automatically generate sentence prompts. Moreover, we design seven related questions and prompt sentence templates to cover as much information as possible about a person, aiming to gain the attributes from VQA are not included in the prompts from ChatGPT. Then the <question, answer> pairs obtained by a VQA model called MCAN (Yu et al. 2019) are converted into seven prompts corresponding to the image. This kind of prompt can be especially applied to situations when attributes cannot be obtained easily. For instance, we can ask these questions as follows: "Is the person wearing a tie?", "Is the person wearing a watch?", and "What kind of shirt is the person wearing?". Next, we randomly assign questions from the question bank to each image and generate several attribute answers. We then pass these answer attributes through predesigned declarative sentence templates and fill in the relevant words to create sub-prompt sentences. Finally, we generate ChatGPT prompts and VQA prompts for 1,501 identities in the Market1501 dataset, as well as for 1,404 identities in the DukeMTMC-reID dataset respectively.

2) Implicit attribute prompts. The implicit prompt strategy in our MP-ReID method uses a learnable prompt approach that does not require intuitive attributes. Specifically, we call it "implicit" because these learnable prompts are training dataset-specific common text descriptions, which is not **directly** corresponding to a sample. Based on CoOp (Zhou et al. 2022b,a) and CLIP-ReID (Li, Sun, and Li 2023), the implicit prompt mainly aims to generate concrete text descriptions through a set of learnable text tokens for fine-grained ReID tasks. That is to say, it provides some attention clues that are somewhat relevant to the tasks. For instance, let the network focus on the human body via "the photo is a [x][x][x][x] person", not a simple/general "the photo is a [x][x][x][x]".

Cross-Modal Alignment

Another technical challenge is how to utilize multi-prompts and alleviate their gaps efficiently. To address it, as shown in Figure 2, we proposed the second component of our MP-ReID — cross-modal alignment, which eases the modality gap between textual prompt features and visual features. Furthermore, similarity learning is used to determine whether feature vectors belong to the same people or not,

$$sim(\mathbb{M}_i, \mathbb{P}_i) = \mathbb{M}_i \cdot \mathbb{P}_i = u_M(m_i) \cdot u_P(p_i) \tag{3}$$

where $u_M(\cdot)$ and $u_P(\cdot)$ are linear layers projecting embedding into a cross-modal embedding space.

1) Aligning for explicit attribute prompts. In this module, we first perform the cross-attention operation (Chen et al. 2022) with the image on both P_i^c ChatGPT generation prompt and P_i^v VQA generation prompt encoded by the CLIP text encoder, respectively. Specifically, the data is sent in a sequential manner to the cross-attention module for processing. In order to integrate prompts f_i^p and images f_i^m more effectively, the textual prompt feature serves as query $(\boldsymbol{Q_i})$. Meanwhile, the image feature and the prompt feature perform the concatenating operation, and are subsequently utilized as key $(\boldsymbol{K_i})$ and value $(\boldsymbol{V_i})$.

Afterwards, we combine the two gained explicit prompt features, namely the ChatGPT prompt f_i^c and the VQA prompt f_i^v via concatenation (Zhai et al. 2022). Finally, the representation of the explicit prompts is an attentive combination of ChatGPT prompts' and VQA prompts' representations. Moreover, f_i^e is formulated as,

$$f_i^e = MLP(Concat(f_i^c, f_i^v)) \tag{4}$$

Then we construct a Multi-Modal Transformer model that combines prompt and image features to unify them into a cross-modal space that can be aligned (Luo et al. 2019). After each of them receives its respective new features, the obtained features are sequentially fed into the Transformer

Category	Baseline Method	Reference	Market1501 mAP	 R@1	DukeMTMC-reID mAP	 R@1
Image-based	SAN	AAAI 2020	88.00	96.10	75.50	87.90
	PAT	CVPR 2021	88.00	95.40	78.20	88.80
	TransReID	ICCV 2021	88.90	95.20	82.00	90.70
	MSDPA	MM 2022	89.50	95.40	82.80	90.90
	DCAL	CVPR 2022	87.50	94.70	80.10	89.00
Attribute-based	AANet	CVPR 2019	66.89	87.04	55.56	73.92
	AMD	ICCV 2021	88.64	95.94	78.26	89.21
	UCAD	IJCAI 2022	79.50	92.60	-	-
	UPAR	WACV 2023	40.60	55.40	-	-
	CLIP-ReID	AAAI 2023	89.60	95.50	82.50	90.00
Ours	**MP-ReID**	-	**95.50**	**97.70**	**88.90**	**95.70**

Table 1: Performance comparison of various state-of-the-art baselines on both datasets.

model together with the image features, so we can get f_i^s. To further enhance the performance, we use a cross-entropy loss $\mathcal{L}_{\text{cls}}$ for the CLS token f_i^{CLS}, which is responsible for the classification representation of the prompts and images. q_k is the value in the target distribution,

$$f_i^s = [f_i^{CLS}, f_i^e, f_i^m] \tag{5}$$

$$\mathcal{L}_{cls} = \sum_{k=1}^{N} -q_k \log(MLP(f_i^s)) \tag{6}$$

We also design an image-to-prompt contrastive loss $\mathcal{L}_{m2p}$, which is calculated as follows,

$$\mathcal{L}_{m2p}(i) = -\log \frac{\exp\left(sim\left(\mathbb{M}_i, \mathbb{P}_i\right)\right)}{\sum_{a=1}^{N} \exp\left(sim\left(\mathbb{M}_i, \mathbb{P}_a\right)\right)} \tag{7}$$

As for explicit prompt, the text-to-image contrastive loss $\mathcal{L}_{p2m}$ is formulated as,

$$\mathcal{L}_{p2m}(i) = -\log \frac{\exp\left(sim\left(\mathbb{M}_i, \mathbb{P}_i\right)\right)}{\sum_{a=1}^{N} \exp\left(sim\left(\mathbb{M}_a, \mathbb{P}_i\right)\right)} \tag{8}$$

Equation (7) and Equation (8) are the similarities of two embeddings from matched pair.

2) Aligning for implicit attribute prompts. As for implicit prompts, the prompts P_i^l are designed as “A photo of a $[X]_1[X]_2[X]_3...[X]_T$ person”, where each $[X]_t$ (with $t \in 1...T$) is a learnable text token with the same dimension as the word embedding. Here, T represents the number of learnable prompt tokens. Notably, the parameters in X can be trained. We can use the obtained implicit prompt features to calculate image-to-prompt cross-entropy $\mathcal{L}_{m2pce}$,

$$\mathcal{L}_{m2pce}(i) = \sum_{k=1}^{N} -q_k \log \frac{\exp\left(sim\left(\mathbb{M}_i, \mathbb{P}_i\right)\right)}{\sum_{a=1}^{N} \exp\left(sim\left(\mathbb{M}_i, \mathbb{P}_a\right)\right)} \tag{9}$$

Finally, in this module, the losses are summarized as follows,

$$\mathcal{L}_{align} = \mathcal{L}_{cls} + \mathcal{L}_{m2p} + \mathcal{L}_{p2m} + \mathcal{L}_{m2pce} \tag{10}$$

Person Retrieval

Through the above steps, we employ the Euclidean distance to calculate the distance score between query images and gallery images. Therefore, a higher score will be generated for a positive pair of person images than those of negative counterparts. In order to optimize ReID models, two loss functions are introduced: a triplet loss $\mathcal{L}_{tri}$ (Hermans, Beyer, and Leibe 2017) and an ID loss $\mathcal{L}_{id}$ (Zheng et al. 2017). The triplet loss is used to minimize the distance between images of the same person while maximizing the distance between images of different people. The ID loss, on the other hand, is used to concretely optimize for correct identity predictions by smoothing label information. By utilizing both the triplet and ID losses, the model is able to simultaneously reduce intra-class distances and increase inter-class distances, resulting in improved accuracy in re-identifying individuals,

$$\mathcal{L}_{id} = \sum_{k=1}^{N} -q_k \log(y_k) \tag{11}$$

$$\mathcal{L}_{tri} = max(d_p - d_n + \alpha, 0) \tag{12}$$

where y_k represents ID prediction logits of class k. d_p and d_n are feature distances of the positive and negative pair, and α is the margin of triplet loss.

Overall, the objective function of our method MP-ReID is denoted as follows, where λ_{tri} is the balance factor of triplet loss and λ_{id} is the balance factor of ID loss,

$$\mathcal{L}_{reid} = \lambda_{id}\mathcal{L}_{id} + \lambda_{tri}\mathcal{L}_{tri} \tag{13}$$

And ultimately, the loss function used in MP-ReID is as follows,

$$\mathcal{L} = \mathcal{L}_{align} + \mathcal{L}_{reid} \tag{14}$$

Experiments

Experimental Settings

Dataset. In this paper, we evaluate the proposed MP-ReID method on two well-known ReID datasets: Market1501 (Zheng et al. 2015), DukeMTMC-reID (Zheng, Zheng, and Yang 2017), as well as the attribute datasets associated with these two datasets, which were manually annotated (Lin et al. 2019).

Strategies						Market1501				DukeMTMC-reID			
LP(Baseline)	**AW**	**GC**	**VP**	**CP**	**CP & VP**	**mAP**	**R@1**	**R@5**	**R@10**	**mAP**	**R@1**	**R@5**	**R@10**
✓						89.60	95.50	-	-	82.50	90.00	-	-
✓	✓					89.40	95.60	96.90	97.30	83.90	92.30	96.50	97.20
✓		✓				86.30	94.00	97.60	98.60	78.10	88.60	93.60	95.10
✓			✓			87.60	94.20	97.50	98.70	78.20	84.70	93.20	94.80
✓				✓		90.20	95.90	98.80	99.30	87.20	94.50	97.60	98.30
✓					✓	**95.50**	**97.70**	**99.20**	**99.50**	**88.90**	**95.70**	**98.00**	**98.70**

Table 2: Ablation study of prompt strategies for MP-ReID on both datasets. (Thereinto, "LP" is learnable prompts, "AW" is coarse and separate attribute words, "GC" is generation caption, "VP" is VQA prompts, "CP" is ChatGPT prompts.)

Evaluation Protocols. To evaluate the performance of our approach, we employed Rank@k and mean Average Precision (mAP) as the evaluation metrics for all experiments on the two datasets (Wang et al. 2021; Farooq et al. 2022). Higher values indicate better performance.

Implementation Details. We apply our method on a server equipped with the NVIDIA GeForce RTX 2080 Ti GPU. We use the Transformer-based models and the learning rate is 5×10^{-7} with a linearly growing. And the warming up is set to 10 to make the model converge faster. In our implementation, we set $S = 16$ and $K = 4$ to enable our model to learn from multiple identities and samples per identity. For feature extraction, prompt features and image features are represented as 512-dimensional vectors. Furthermore, we set the ID loss balance factor λ_{id} to 0.25 as a regularization strategy, λ_{tri} and the weight of $\mathcal{L}_{align}$ is set to 1. Regarding the triplet loss, we set the margin parameter ξ to 0.3 to create an adequate boundary between the positive and negative samples. Moreover, we directly use the off-the-shelf ChatGPT 3.5 for the explicit prompts generation.

Overall Performance Comparison

To demonstrate the effectiveness of our proposed method, we compared it with several state-of-the-art approaches. And we employ R@1, R@5, R@10 for convenience of representation.

Table 1 presents the experimental results, and we have the following observations: 1) Our MP-ReID approach achieves better performance on both datasets, significantly outperforming state-of-the-art baselines. It is mainly because the MP-ReID model employs the multi-prompts paradigm to significantly enhance the identification performance. This suggests the presence of highly informative cues in the image and prompt that were neglected in traditional person ReID schemes. 2) Despite recent advancements in attribute-based algorithms for person ReID, several popular methods such as AANet (Chen et al. 2021), AMD (Chen et al. 2021), UCAD (Yan et al. 2022), UPAR (Specker, Cormier, and Beyerer 2023) and CLIP-ReID (Li, Sun, and Li 2023) have demonstrated poor search results due to technological limitations that prevent full utilization of attribute information. On the other hand, image-based methods such as SAN (Jin et al. 2020b), PAT (Li et al. 2021), TransReID (He et al. 2021a), MSDPA (Cheng et al. 2022) and DCAL (Zhu et al. 2022), while effective in some regards, do not take into account attribute information, leading to a potential loss of valuable information for person ReID. The utilization of multi-prompts in MP-ReID significantly improves the retrieval performance of person ReID. In particular, the performance of R@1 on the Market1501 dataset and the DukeMTMC-reID dataset improves significantly by at least 5.9% and 6.1%, respectively.

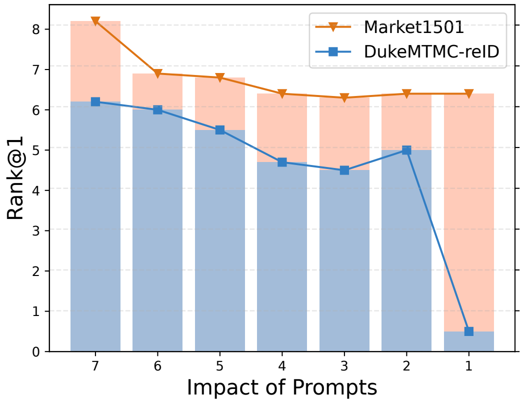


Figure 4: Ablation study on the effect of reducing various sub-prompts for MP-ReID.

Ablation Studies

The overall comparative analysis shows that our proposed MP-ReID solution exhibits superior performance. To further validate the importance of multi-prompts in ReID, we took CLIP-ReID with implicit prompt as a baseline and performed some ablation studies. Firstly, MP-ReID is compared with its several variants: 1) MP-ReID with the coarse and separate attributes prompt. 2) MP-ReID with generation caption prompts from an image captioning model. 3) MP-ReID with/without any ensembling explicit prompts, i.e., ChatGPT generation prompts, as well as VQA generation prompts.

1) Ablation study of prompt strategies for MP-ReID. Table 2 displays the performance of different component combinations of MP-ReID. Our conclusions are threefold: a) MP-ReID using coarse and separate attribute words and generation caption shows wicked retrieval results than the prompts generated by the large model ChatGPT. b) both the explicit prompt and the implicit prompt in the table show

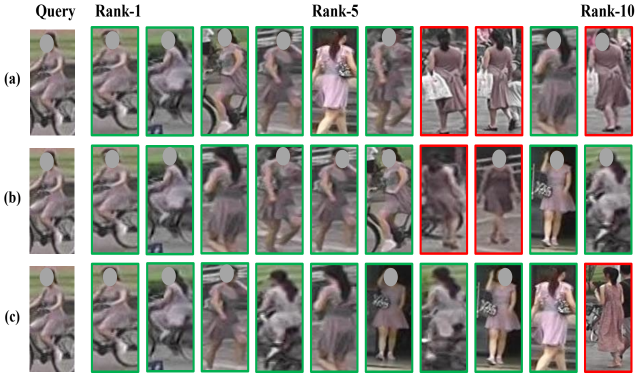


Figure 5: Visualization of three examples that illustrate the retrieval results about a) baseline (implicit learnable prompt); b) + coarse and separate attribute words; and c) our MP-ReID. Thereinto, the green box denotes the same ID as the query image, and the red box indicates a different ID from the query image.

relatively better performance. c) MP-ReID outperforms MP-ReID w/ VQA generation prompt by 7.9% and 10.7% in mAP on Market1501 and DukeMTMC-reID datasets. Furthermore, the scheme of MP-ReID w/ ChatGPT generation prompt proved inferior to MP-ReID by 5.3% and 1.7% in mAP on Market1501 and DukeMTMC-reID datasets. Furthermore, research has shown that using multiple fine prompts is more effective.

2) Ablation study on multiple prompts. As Figure 4 revealed, to further gain deeper insight into the effectiveness of multi-prompts learning in MP-ReID, we compared the effect of different numbers of multi-prompts in R@1 by showcasing on Market1501 and DukeMTMC-reID datasets. Thereinto, the graph is presented by subtracting the base value of 89% from the obtained R@1. This presentation method is utilized to enhance the clarity of the graph. Significantly, we have the following observations: a) we use multiple sub-prompts, including 7 VQA prompts, 1 ChatGPT prompt and 1 learnable prompt. We gradually eliminate 1 - 4 VQA prompts and 1 ChatGPT prompt when only one learnable prompt remains in our experiments. The results have obviously shown that more prompts are more effective than few prompts for ReID, because few prompts cannot be learned much information. b) These findings certify the effectiveness of combining the ChatGPT generation prompts and VQA generation prompts these two explicit prompts and the implicit prompts components in our MP-ReID approach. In addition, the ablated results reveal the necessity of multiple prompts proposed in our framework, jointly resulting in its superior performance.

Visualization

This paper aims to combine multiple prompts features to enhance visual features. As Figure 5 reveals, to better comprehend our MP-ReID network, we examined and evaluated the person retrieval outcomes through visualization and analysis. Figure 5a displays the effects of the baseline CLIP-ReID with implicit learnable prompts. The method adopts implicit prompts integrating coarse and separate attributes see Figure 5b. And Figure 5c is our MP-ReID with multiple prompts enhanced for ReID. The green box means the same ID as the query image, and the red box reveals a different ID from the query image. We can observe that MP-ReID achieves the optimal performance on Rank-10, which is mainly because of the newly introduced multi-prompts learning. Our proposed prompts learning strategies facilitate the discovery of fine-grained discriminative clues by leveraging more relevant characteristic prompts among samples.

Conclusions

This paper introduces a new concept of multi-prompts and a novel framework for attribute-based ReID, named MP-ReID. Specifically, we take the first attempt to explore multiple prompts generation learning strategies with ChatGPT and VQA models, which effectively learn discriminative representations via generated multi-prompts information. For the concrete prompts generation, we classify it into explicit prompts and implicit prompts. Among them, for generating explicit prompts, large model ChatGPT and VQA are used based on a prompt ensembling paradigm, and the implicit prompts are learnable prompts. The model is then refined using well-designed losses that consider textual prompts and visual image constraints to alleviate the modality gap. Our MP-ReID has achieved state-of-the-art performance on two well-known ReID datasets.

Limitations

In this paper, we propose the use of multiple prompts to enhance the person re-identification task, which has been experimentally validated as effective. However, explicit and implicit integration aspects warrant further exploration. For instance, in terms of quantity, additional prompt methods beyond the current three can be considered. Furthermore, the integration strategy can be further refined. Our current integration strategy is relatively straightforward, but we believe that employing more diverse and tighter integration methods will yield even better results. At the model level, we are particularly intrigued by multi-modal large models, but due to dataset and resource constraints, we have not yet conducted extensive experimentation. We anticipate that larger models and corpora will reveal more intriguing findings.

Acknowledgments

The authors are highly grateful to the anonymous referees for their careful reading and insightful comments. The work is partially supported by the National Natural Science Foundation of China (No.61802121, No.62302246, No.U20A20174, No.U22A2030), the National Key R&D Projects (No.2022YFB3103500), the Natural Science Foundation of Hunan Province, China (No.2022JJ30159 and No.2023JJ20013), Technology Projects of Hunan Province (No.2015TP1004), Science and Technology Key Projects of Changsha City (No.kh2103003) and the Natural Science Foundation of Zhejiang Province, China (No.LQ23F010008) and China Scholarship Council.

References

Chen, S.; Zeng, Y.; Cao, D.; and Lu, S. 2022. Video-guided machine translation via dual-level back-translation. *Knowledge-Based Systems*, 245: 108598.

Chen, X.; Liu, X.; Liu, W.; Zhang, X.-P.; Zhang, Y.; and Mei, T. 2021. Explainable person re-identification with attribute-guided metric distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11813–11822.

Cheng, X.; Jia, M.; Wang, Q.; and Zhang, J. 2022. More is better: Multi-source dynamic parsing attention for occluded person re-identification. In *Proceedings of the ACM International Conference on Multimedia*, 6840–6849.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An image is worth 16x16 words: transformers for image recognition at scale. *ICLR*.

Farooq, A.; Awais, M.; Kittler, J.; and Khalid, S. S. 2022. AXM-Net: Implicit cross-modal feature alignment for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4477–4485.

Fu, D.; Chen, D.; Yang, H.; Bao, J.; Yuan, L.; Zhang, L.; Li, H.; Wen, F.; and Chen, D. 2022. Large-scale pre-training for person re-identification with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2476–2486.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021a. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15013–15022.

He, T.; Jin, X.; Shen, X.; Huang, J.; Chen, Z.; and Hua, X.-S. 2021b. Dense interaction learning for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1490–1501.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Jeong, B.; Park, J.; and Kwak, S. 2021. Asmr: Learning attribute-based person search with adaptive semantic margin regularizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12016–12025.

Jia, J.; Chen, X.; and Huang, K. 2021. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 962–971.

Jia, J.; Gao, N.; He, F.; Chen, X.; and Huang, K. 2022. Learning disentangled attribute representations for robust pedestrian attribute recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1, 1069–1077.

Jin, X.; He, T.; Shen, X.; Liu, T.; Wang, X.; Huang, J.; Chen, Z.; and Hua, X.-S. 2022. Meta clustering learning for large-scale unsupervised person re-identification. In *Proceedings of the ACM International Conference on Multimedia*, 2163–2172.

Jin, X.; Lan, C.; Zeng, W.; and Chen, Z. 2020a. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11165–11172.

Jin, X.; Lan, C.; Zeng, W.; and Chen, Z. 2023. Domain prompt tuning via meta relabeling for unsupervised adversarial adaptation. *IEEE Transactions on Multimedia*.

Jin, X.; Lan, C.; Zeng, W.; Wei, G.; and Chen, Z. 2020b. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 07, 11173–11180.

Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; and Xie, W. 2022. Prompting visual-language models for efficient video understanding. In *Proceedings of the European Conference on Computer Vision*, 105–124. Springer.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*.

Li, S.; Sun, L.; and Li, Q. 2023. CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; and Wu, F. 2021. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2898–2907.

Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; and Yang, Y. 2019. Improving person re-identification by attribute and identity learning. *Pattern Recognition*.

Liu, Y.; Wei, W.; Peng, D.; and Zhu, F. 2022. Declaration-based prompt tuning for visual question answering. In *Proceedings of the Thirty-first International Joint Conference on Artificial Intelligence*.

Lüddecke, T.; and Ecker, A. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7076–7086.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. 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*.

Niu, K.; Huang, L.; Huang, Y.; Wang, P.; Wang, L.; and Zhang, Y. 2022. Cross-Modal Co-Occurrence Attributes Alignments for Person Search by Language. In *Proceedings of the ACM International Conference on Multimedia*, 4426–4434.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A.; and Riedel, S. 2019. Language models as knowledge bases? Proceedings of the Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the Conference on International Conference on Machine Learning*, 8748–8763.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18082–18091.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, 1715–1725.

Song, X.; Jing, L.; Lin, D.; Zhao, Z.; Chen, H.; and Nie, L. 2022. V2P: Vision-to-prompt based multi-modal product summary generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 992–1001.

Specker, A.; Cormier, M.; and Beyerer, J. 2023. UPAR: Unified pedestrian attribute recognition and person retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 981–990.

Wang, X.; Zheng, S.; Yang, R.; Zheng, A.; Chen, Z.; Tang, J.; and Luo, B. 2022a. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121: 108220.

Wang, Z.; Wang, Z.; Zheng, Y.; Wu, Y.; Zeng, W.; and Satoh, S. 2021. Beyond intra-modality: a survey of heterogeneous person re-identification. In *Proceedings of the International Conference on International Joint Conferences on Artificial Intelligence*, 4973–4980.

Wang, Z.; Zhang, Z.; Lee, C.; Zhang, H.; Sun, R.; Ruoxi, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Wu, H.; Ma, B.; Liu, W.; Chen, T.; and Nie, D. 2022. Fast and constrained absent keyphrase generation by prompt-based learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10, 11495–11503.

Yan, Y.; Yu, H.; Li, S.; Lu, Z.; He, J.; Zhang, H.; and Wang, R. 2022. Weakening the influence of clothing: universal clothing attribute disentanglement for person re-identification. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, 1523–1529.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 2872–2893.

Yu, Z.; Pei, J.; Zhu, M.; Zhang, J.; and Li, J. 2022. Multi-attribute adaptive aggregation transformer for vehicle re-identification. *Information Processing & Management*, 59(2): 102868.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6281–6290.

Zeng, Y. 2022. Point prompt tuning for temporally language grounding. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003–2007.

Zhai, Y.; Zeng, Y.; Cao, D.; and Lu, S. 2022. Trireid: Towards multi-modal person re-identification via descriptive fusion model. In *Proceedings of the International Conference on Multimedia Retrieval*, 63–71.

Zhang, C.; Zhang, C.; Li, C.; Qiao, Y.; Zheng, S.; Dam, S. K.; Zhang, M.; Kim, J. U.; Kim, S. T.; Choi, J.; et al. 2023. One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era. *arXiv preprint arXiv:2304.06488*.

Zhang, J.; Niu, L.; and Zhang, L. 2020. Person re-identification with reinforced attribute attention selection. *IEEE Transactions on Image Processing*, 30: 603–616.

Zhang, Z.; Zhang, H.; and Liu, S. 2021. Person re-identification using heterogeneous local graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12136–12145.

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

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

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1367–1376.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3754–3762.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

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

Zhu, H.; Ke, W.; Li, D.; Liu, J.; Tian, L.; and Shan, Y. 2022. 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*, 4692–4702.

Zhu, J.; Jin, J.; Yang, Z.; Wu, X.; and Wang, X. 2023. Learning CLIP guided visual-text fusion transformer for video-based pedestrian attribute recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2625–2628.