CLIP-driven View-aware Prompt Learning for Unsupervised Vehicle Re-identification

Jiyang Xu[1], Qi Wang[1]*, Xin Xiong[2], Di Gai[1], Ruihua Zhou[3], Dong Wang[3]
[1]School of Mathematics and Computer Sciences, Nanchang University
[2]Department of Information, The First Affiliated Hospital, Jiangxi Medical College, Nanchang University
[3]School of Software, Nanchang University
xujiyang@email.ncu.edu.cn;wangqi@ncu.edu.cn;xiongxinxx@ncu.edu.cn;gaidi@ncu.edu.cn;
zrh@email.ncu.edu.cn;dongwang@email.ncu.edu.cn

Abstract

With the emergence of vision-language pre-trained models, such as CLIP, some textual prompts have been gradually introduced recently into re-identification (Re-ID) tasks to obtain considerably robust multimodal information. However, most textual descriptions based on vehicle Re-ID tasks only contain identity index words without specific words to describe vehicle view information, thereby resulting in difficulty to be widely applied in vehicle Re-ID tasks with view variations. This case inspires us to propose a CLIP-driven view-aware prompt learning framework for unsupervised vehicle Re-ID. We first design a learnable textual prompt template called **view**-aware **co**ntext **op**timization (ViewCoOp) based on dynamic multi-view word embeddings, which can fully obtain the proportion and position encoding of each view in the whole vehicle body region. Subsequently, a cross-modal mutual graph is constructed to explore the connections between inter-modal and intra-modal. Each sample is treated as a graph node, which extracts textual features based on ViewCoOp and the visual features of images. Moreover, leveraging the inter-cluster and intra-cluster correlation in the bimodal clustering results in the determination of connectivity between graph node pairs. Lastly, the proposed cross-modal mutual graph method utilizes supervised information from the bimodal gap to directly fine-tune the image encoder of CLIP for downstream unsupervised vehicle Re-ID tasks. Extensive experiments verify that the proposed method is capable of effectively obtaining cross-modal description ability from multiple views.

Introduction

The objective of vehicle re-identification (Re-ID) task (Zheng et al. 2023; Khorramshahi, Shenoy, and Chellappa 2023; He et al. 2023) is to retrieve specified targets in a cross-camera surveillance system. Previously supervised Re-ID works have achieved impressive performance through training with massive amounts of labeled data. To reduce the expensive cost of manual annotations, numerous studies (Lu et al. 2023; Wang et al. 2023b) have focused on obtaining informative representations through clustering-based algorithms in an unsupervised manner. However, relying solely on visual modality information limits the cross-modal capacity of unsupervised Re-ID tasks, resulting in difficulty for the training models to robustly distinguish the appearance differences among various vehicle models.

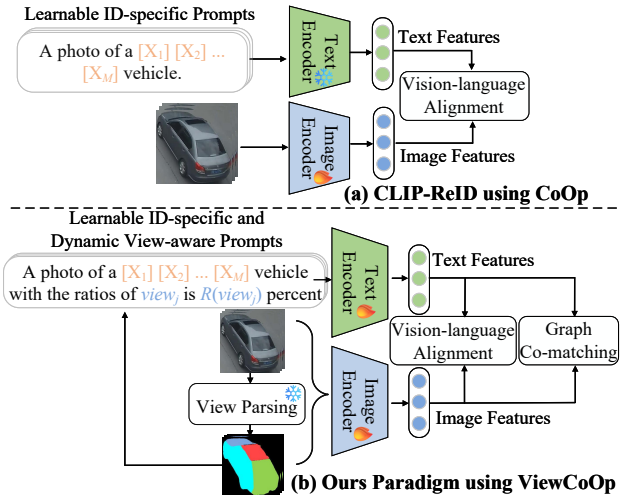

Figure 1: (a) Illustration of the CLIP-ReID (Li, Sun, and Li 2023) using CoOp. (b) In this work, we introduce ViewCoOp to achieve vision-language alignment and cross-modal mutual graph co-matching.

Previous studies (Liu et al. 2023; Zhao et al. 2024; Tan et al. 2024) have used textual descriptions to assist the Re-ID model to learn considerably robust visual features, especially in the text-to-image Re-ID paradigm. This vision-language contrastive learning paradigm requires annotating large-scale textual descriptions and then aligning features with images to obtain the cross-modal performance. Relevant vision-language pre-trained models, such as CLIP (Radford et al. 2021) have achieved impressive performance in downstream zero-shot vision recognition tasks. Given that the text encoder of CLIP can leverage the fixed prompt template, "a photo of class" as textual input to make zero-shot predictions of image categories, many researchers have been inspired to design reasonable prompt templates for fine-tuning CLIP, thereby alleviating the annotation cost of manual textual descriptions. Note that the vehicle Re-ID task ap-

*Corresponding author.

plies the ID index as label, and using "a photo of class" as the textual description is incapable of specifically describing any vehicle image. A feasible solution is to introduce ID-specific learned tokens via context optimization (CoOP) (Zhou et al. 2022b) to form textual descriptions of vehicle identities. Accordingly, Li *et al.* (Li, Sun, and Li 2023) first proposed CLIP-ReID, which utilizes the ID-specific textual expression of CoOP to provide additional supervised information for visual modalities. In particular, each training sample is automatically annotated with a textual description (i.e., "A photo of a $[V_1], [V_2], ..., [V_M]$ vehicle") containing M learnable tokens. Subsequently, M tokens are assigned informative ID properties by aligning the image and text representation, as shown in Figure 1(a). However, ID-specific textual prompts are incapable of accurately describing each view of the vehicle, resulting in extracted textual features that are not sensitive to view variations. This case motivates us to explore a learnable textual prompt template based on view-aware embeddings to provide specific descriptions of vehicle multi-view information.

Despite introducing ID-specific textual prompts that can provide informative descriptions, how to apply the powerful image-text alignment ability of CLIP to the vehicle Re-ID task is a critical issue. The vision-language modal alignment can be classified into global and multi-grained alignments. Early studies have utilized the feature projection of image-text pairs onto a latent space to directly achieve alignment via contrastive loss. For multi-grained alignment, some remarkable studies (Jiang and Ye 2023; Zhai et al. 2024; Yan et al. 2023) have designed multiple attribute prompt templates or extracted different types of local features for semantic alignment operations at different levels. Zhai *et al.* (Zhai et al. 2024) attempted to introduce three types of textual descriptions, namely, VQA, GPT, and learnable prompts, to achieve explicit and implicit cross-modal alignments. Although these methods have achieved remarkable performance in various vision-language tasks, they have only considered modal alignment within instance-level image-text pairs. How to mine the positive and negative correlation between different image-text pairs in the cross-modal feature distribution remains a key issue that needs to be solved. We aim to explore a novel vision-language alignment paradigm based on graph-based constraint to facilitate latent relationships between image-text pairs, as illustrated in Figure 1(b).

We present a CLIP-driven view-aware prompt learning framework for unsupervised vehicle Re-ID. Our main contributions are summarized in the following three points.

- We first propose a view-aware learnable textual prompt called ViewCoOp, specifically for unsupervised vehicle Re-ID tasks. The description content of ViewCoOp via the view paring module is enabled to obtain dynamic multi-view clues and express ID-specific textual information that adapts to view variations.
- To facilitate the cross-modal capability of the Re-ID model, a cross-modal mutual graph method is developed, which establishes semantic relationships between image-text pairs to reduce the discrepancy between textual and visual modalities. The training process achieves image-text feature alignment by fine-tuning the text and image encoders of CLIP.
- Extensive experimental results on the VeRi-776 and VehicleID datasets demonstrate that our method outperforms other related ones in terms of view variations and cross-modal performance.

Related Work

Prompt Learning

The development of advanced large-scale pre-training vision-language models, like CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), and exploring how to effectively adapt these models to specific downstream tasks have become particularly important. Reasonable textual description is required during the inference stage to guarantee precise image-text matching. To avoid tedious manual annotations, recent works (Zhou et al. 2022b,a; Huang and Chu 2022; Miyai et al. 2024) have introduced prompt learning into the visual recognition field. CoOp (Zhou et al. 2022b) first attempted to propose a learnable continuous prompt representation instead of manual prompts. Additionally, several researchers (Zhai et al. 2024; Huang and Chu 2022) use prompt learning to create pseudo-labels and self-supervised information for application in unsupervised domain adaptation tasks and pure unsupervised ones. Due to the impressive performance of prompt learning, it has been explosively empolyed into cross-modal Re-ID. Some pioneer works (Li, Sun, and Li 2023; Zhai et al. 2024; Yan et al. 2023) attempt to use learnable ID-specific prompt, multi-attribute prompts, and fine-grained prompts as auxiliary textual description to provide valuable information for visual modalities. As the above cases only consider how to design ID-related descriptions, their prompt does not involve multi-view description of the vehicle, making it unable to be directly applied to vehicle Re-ID.

Multi-view Vehicle Re-ID

Multi-view vehicle Re-ID injects view information during the training process to encourage the model to cope with the challenge of view variations (Li et al. 2023a, 2022b). There are currently two mainstream paradigms, including multi-view representation learning and multi-view label refinement. The methods (Zhang et al. 2022a; Li et al. 2023b, 2022a) in the former are to enhance the model expression of view representations by supervised multi-view annotations, focusing on learning the appearance information of vehicles under different views. The methods (Meng et al. 2020; Dong et al. 2024) in the latter focus on how to adaptively detect view variations, which smooth label weights through the distribution relationship of multi-view features. For instance, Meng *et al.* (Meng et al. 2020) introduced the Perspective Viewpoint Network (PVEN) to align fine-grained feature representations of vehicles across various viewpoints. The above efforts only consider extracting view-related information from visual modalities to overcome the view variation issue. In contrast, this paper seeks to establish a cross-modal view-aware framework for unsupervised vehicle Re-ID.

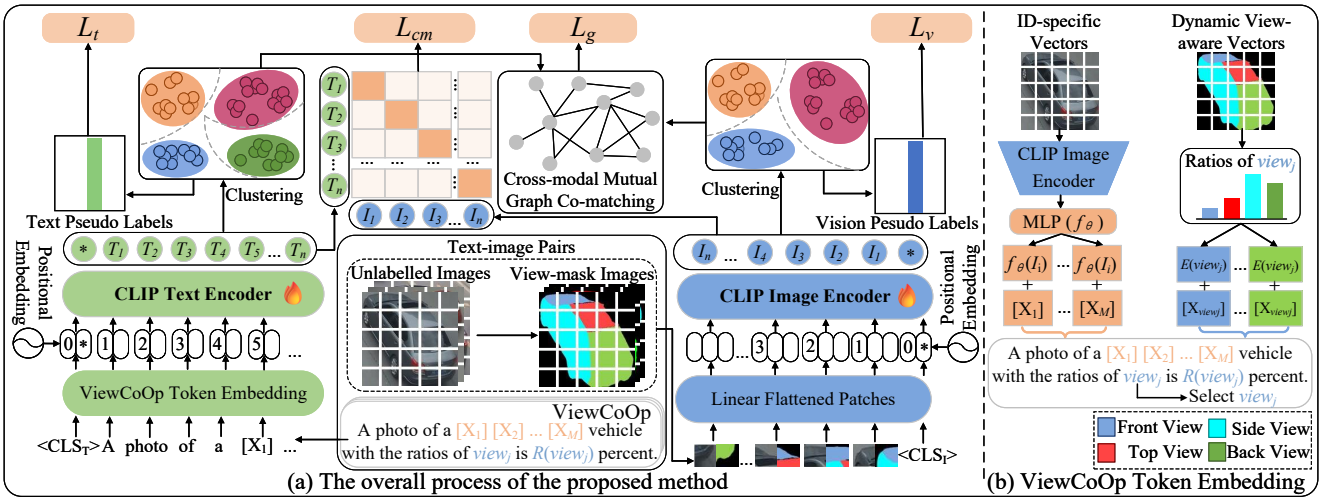

Figure 2: (a) Firstly, it includes a view-aware context optimization(ViewCoOp). Subsequently, a cross-modal mutual graph co-matching is devised to differentiate between positive and negative sample pairs in images and provide reliable supervisory information. (b) The original images and view-mask ones are encoded in parallel as a set of learnable ID-specific and view proportion vectors. Subsequently, these vectors are embedded into the ViewCoOp prompt to obtain the view-aware textual tokens.

Method

Overall Framework

This section proposes a CLIP-driven view-aware prompt learning framework for unsupervised vehicle Re-ID, as shown in Figure 2(a). Our framework consists of three key modules: ViewCoOp, cross-modal mutual graph method, and CLIP-driven fine-tuning for unsupervised vehicle Re-ID. Given a set of unlabeled images $I = \{I_1, I_2, ..., I_N\}$, we first extract the view-mask of the images through a vanilla view parsing model to form the ViewCoOp textual descriptions $T = \{T_1, T_2, ..., T_N\}$. In particular, the description of ViewCoOp includes a set of ID-specific learnable vectors and four dynamical view proportion vectors. When ViewCoOp is input into the text encoder of CLIP, we also simultaneously input the images and its mask into the image encoder, attempting to optimize the view perception ability of ViewCoOp by integrating visual position embeddings. More importantly, text encoder is not frozen to update ViewCoOp. Subsequently, we consider the potential modal gaps between visual and textual modalities and design a cross-modal mutual graph method based on the clustering results of each modality to further explore the feature distribution relationships between positive and negative image-text pairs. Lastly, the extracted textual and visual features undergo image-text alignment to directly fine-tune CLIP. Thereafter, the optimized image encoder is applied to the feature extractor for unsupervised vehicle Re-ID.

View-aware Context Optimization

Owing to the inefficiency and time-consuming nature of manual annotation, CoOP was a pioneering work that designed a set of learnable tokens for prompt engineering. If we use the template of CoOP, which is "a photo of a $[X_1][X_2]...[X_M]$ vehicle" to generate textual descriptions for each sample in the vehicle Re-ID dataset, then we observe that this is not a suitable choice. The main reason is that vehicles with the same ID have significant appearance discrepancies from different views. The set of learnable vectors in CoOP can only learn ID-specific information and is not sensitive to view information, which cannot effectively deliver the different multi-view features of vehicles.

This paper proposes a view-aware learnable textual prompt called ViewCoOp, which can embed multi-view information specifically for vehicle Re-ID datasets, as shown in Figure 2(b). The ViewCoOp format includes a set of learnable ID-specific vectors and a set of learnable dynamic patch-level view proportion vectors, which are "A photo of a $[X_1][X_2]...[X_M]$ vehicle with the proportions of $view_j$ is $R(view_j)$ percent", where $[X_1][X_2]...[X_M]$ represents a set of learnable ID-specific vectors, $view_j \in \{front\ view;\ back\ view;\ side\ view;\ top\ view\}$, and $R(view_j)$ refers to the patch-level ratio of the $view_j$. Note that to enhance the flexibility of the ViewCoOp format, we randomly shuffle the word set of each view (i.e., "$view_j$ is $R(view_j)$ percent") throughout the whole sentence during each epoch training to achieve data augmentation of the textual modality. For the ratio calculation of each view, we utilize the segmentation model proposed in Meng *et al.* (Meng et al. 2020) to obtain the view-mask image I_{mask} of each sample. In particular, the I_{mask} obtained from each sample can be parsed into four views for segmentation results. The patch-level proportion of each view $R(view_j)$ can be encoded as Eq. (1):

$$R\left(view_j\right) = E\left(\frac{num(P_{view_j})}{\sum_{view_j} num\left(P_{view_j}\right)}\right) + X_{view_j} \quad (1)$$

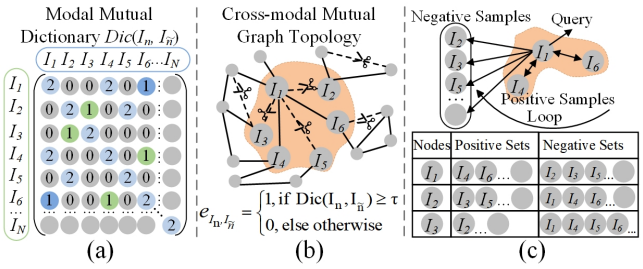

Figure 3: The core components of the cross-modal mutual graph co-matching method include (a) Modal mutual dictionary construction, (b) cross-modal mutual graph topology, and (c) construction of positive and negative sample pairs.

where $P_{view_j} \in I_{mask}$, P_{view_j} and $num(P_{view_j})$ belongs to the patch set and the patch count functions, respectively, for $view_j$, and $E(\cdot)$ denotes embedding the word for patch-level proportion of each view. Simultaneously, X_{view_j} is a learnable view bias term for $view_j$, which drives the patch-level proportion of $view_j$ to become dynamic.

For the ID-specific learnable vectors $[X_1][X_2]...[X_M]$, the idea of CoOp is a static design that becomes fixed once learned. For this purpose, our ViewCoOp introduces a lightweight network MLP (f_θ) to extract image I_n into the visual vector $f_\theta(I_n)$. Thereafter, $f_\theta(I_n)$ with the same dimension and learnable text vectors $[X_1][X_2]...[X_M]$ is combined to obtain the final ID-specific prompt, which is suitable for unsupervised vehicle Re-ID tasks, in which the number of clusters in each epoch dynamically changes. To this end, the m-th learnable textual vetor is represented as $X_m(I_n) = X_m + f_\theta(I_n), m \in \{1, 2, ..., M\}$, following (Li, Sun, and Li 2023), we empirically set M to 4. The encoding process of the textual description of T_n can be described in Eq. (2):

$$E(T_n) = cat(w,\ cat(X_m(I_n)), cat(R(view_j)) \quad (2)$$

where w represents non-learnable word embedding, $cat(X_m(I_n))$ and $cat(R(view_j))$ express concatenations of the ID-specific learnable vectors and of each dynamic view proportion vector, respectively. Subsequently, we feed the ViewCoOp into the text encoder of CLIP to extract the textual features of all samples.

Cross-modal Mutual Graph Co-matching

The introduction of ViewCoOp has enabled the automatic generation of view-aware textual description. However, how to narrow the difference between visual and textual modalities has become the emphasis of our next focus. For this purpose, a cross-modal mutual graph co-matching method is designed to explore potential correlations between positive and negative image-text pairs, as shown in Figure 3.

Modal mutual dictionary construction. Modal mutual dictionary construction. We first perform separate feature clustering on the visual modality features $f^i = \{f_1^i, f_2^i, ..., f_N^i\}$ and text modality features $f^t = \{f_1^t, f_2^t, ..., f_N^t\}$ extracted from CLIP. Specifically, we feed the vehicle Re-ID dataset to the DBSCAN clustering algorithm to obtain the visual clustering output and the textual clustering output. The visual one-hot pseudo labels $y^i = \{y_1^i, y_2^i, ..., y_N^i\}$ and textual ones $y^t = \{y_1^t, y_2^t, ..., y_N^t\}$ are assigned to each sample. We use the two clustering results as bases to construct a modal mutual dictionary Dic to store the indexes of the bimodal constraints. The size of the dictionary is $N \times N$ and it is dynamically updated in each iteration. Figure 3(a) presents the saving rule of the dictionary Dic containing bimodal constraint indexes. Assuming that the visual and textual clustering indexes of samples I_n and $I_{\tilde{n}}$ are in the same cluster, samples I_n and $I_{\tilde{n}}$ can be considered to have high similarity, and $Dic(I_n, I_{\tilde{n}}) = 2$. If the bimodal clustering of I_n and $I_{\tilde{n}}$ are not in the same cluster, it indicates that they do not have any potential relationship between modalities, then $Dic(I_n, I_{\tilde{n}}) = 0$. Evidently, there exists only one modal clustering index for samples I_n and $I_{\tilde{n}}$ in the same cluster, where $Dic(I_n, I_{\tilde{n}}) = 1$. The representation of $Dic(I_n, I_{\tilde{n}})$ is defined as Eq. (3):

$$Dic(I_n,\ I_{\tilde{n}}) = \begin{cases} 0,\ s.t.\ y_n^i \neq y_{\tilde{n}}^i \& y_n^t \neq y_{\tilde{n}}^t \\ 1,\ s.t.\ y_n^i = y_{\tilde{n}}^i \mid y_n^t = y_{\tilde{n}}^t \\ 2,\ s.t.\ y_n^i = y_{\tilde{n}}^i \& y_n^t = y_{\tilde{n}}^t \end{cases} \quad (3)$$

Cross-modal graph topology. After constructing a modal mutual dictionary Dic, a graph topology is developed to explore the correlation between positive and negative image-text pairs in the entire data distribution. Taking the eight samples in Figure 3(b) as an example, we treat each sample as a graph node and determine whether or not the node pairs $\{v_1, v_2, \cdots, v_8\}$ are connected by searching for the corresponding index values in the dictionary. Moreover, we set the dictionary index threshold τ to be greater than or equal to 1 as a condition for constructing edges for graph nodes, and vice versa. The main reason is that an index value greater than or equal to 1 indicates a potential correlation of at least one modality between two samples. For example, given a graph node pair $(v_1,\ v_4)$ and $(v_1,\ v_6)$, based on the index values of the dictionary in Figure 3(a), which are $Dic(v_1,\ v_4) = 2$ and $Dic(v_1,\ v_6) = 1$, an edge is determined to be constructed between the graph node pairs $(v_1,\ v_4)$ and $(v_1,\ v_6)$. In addition, note that the dictionary index value $Dic(v_1,\ v_2) = 0$, indicating that there is no edge connection between the graph node pairs $(v_1,\ v_2)$. The construction method of the cross-modal mutual graph Graph(V, E) is shown as Eq. (4):

$$Graph(V, E) = \begin{cases} V = \{v_n | v_n = I_n, n = \{1, 2, ..., N\}\} \\ E = \{e_{v_n, v_{\tilde{n}}} | Dic(v_n, v_{\tilde{n}}) \geq \tau\} \end{cases} \quad (4)$$

where V and E represent the node set and edge combination of the graph, respectively. v_n belongs to an element of V. $e_{v_n,\ v_{\tilde{n}}}$ indicates the existence of an edge between v_n and $v_{\tilde{n}}$, and τ is the threshold for connecting edges.

Graph-aware constraint loss. Through the construction of the cross-modal mutual graph, we transform the discrete bimodal features into the connection relationship of edges between nodes. For a certain sample pair, we determine its positivity or negativity based on the connectivity of the

edges in the graph. That is, if there are edges between node pairs, then it forms a positive sample set; otherwise, it forms a negative sample group. Figure 3(c) illustrates the positive and negative sample sets of sample I_1 in Eq. (5):

$$P_{I_1} = \begin{cases} P_{I_1}^{+} = \{I_4,\ I_6,\ ...\}, if\ e_{I_1,I_n} \in Graph \\ P_{I_1}^{-} = \{I_2, I_3, I_5, ...\}, otherwise \end{cases} \tag{5}$$

where $P_{I_1}^{+}$ and $P_{I_1}^{-}$ represent the set of positive and negative sample set, respectively, for I_1.

On the basis of the cross-modal constraint information contained in the graph, we design a graph-aware constraint loss L_g to mine the potential distribution distance between unlabeled samples. This design ensures that the similarity distance between positive sample pairs is minimized and the similarity distance between negative sample pairs is maximized. L_g can be defined as Eq. (6):

$$L_g = \sum_{1}^{N} \frac{\sum_{I_k \in P_{I_j}^{+}} \left(1 - S(f_{I_j}^{i}, f_{I_k}^{i})\right) + \sum_{I_k \in P_{I_j}^{-}} \left(1 + S(f_{I_j}^{i}, f_{I_k}^{i})\right)}{N} \tag{6}$$

where $S(f_{I_n}^{i},\ f_{I_{\tilde{n}}}^{i})$ is the cosine similarity distance of visual features between samples I_n and $I_{\tilde{n}}$, with a range of $[-1, 1]$.

CLIP-Driven Fine-tuning for Unsupervised Vehicle Re-ID

This section updates the original CLIP by comparing the text and image visual features extracted by ViewCoOp. We activate the parameters of CLIP during the training process. The main reason is that the textual description of ViewCoOp is dynamically learnable in each epoch and can align with instance-level visual features. Subsequently, we input textual features into the DBSCAN clustering algorithm to obtain textual one-hot pseudo labels, which are used as supervised information for textual modality training. The text loss function L_t is defined as Eq. (7):

$$L_t = -\sum_{n=1}^{N} y_n^t \cdot log \frac{exp(S(f_n^t,\ f_{\varphi_n}^t))}{\sum_{c=1}^{C} exp(S(f_n^t,\ f_{\varphi_c}^t))} \tag{7}$$

where y_n^t and $f_{\varphi_n}^t$ represent the pseudo labels and centroid features, respectively, of sample I_N in the textual cluster, respectively. C means the number of clusters, and $S(f_n^t,\ f_{\varphi_n}^t)$ is the cosine distance between the textual features f_n^t and its centroid features $f_{\varphi_n}^t$.

Similar to CLIP, we calculate the cross-modal similarity matrix based on the extracted textual and visual features to maintain cross-modal alignment. The cross-modal features corresponding to the samples on the diagonal of the matrix should be as close as possible, while the cross-modal features of samples on non diagonal lines should be as far away as possible. The calculation method for cross-modal loss L_{cm} is shown in Eq. (8):

$$L_{cm} = \frac{\sum_{j=1}^{N} (1 - S(f_{I_j}^{i}, f_{I_j}^{t})) + \sum_{j=1}^{N} \sum_{k=1,\ j \neq k}^{N} (1 + S(f_{I_j}^{i}, f_{I_k}^{t}))}{N^2} \tag{8}$$

Following the universal setting of unsupervised vehicle Re-ID tasks, we adopt the triplet loss L_{tri} and cross-entropy loss L_{ce} as visual modality loss $L_v = L_{ce} + L_{tri}$, and use the one-hot pseudo label y_N^i obtained from visual feature clustering as the supervised information of the visual modality to optimize the image encoder. The calculation of the total loss L_{total} is defined as Eq. (9):

$$L_{total} = L_v + L_t + L_g + L_{cm} \tag{9}$$

The image encoder is updated by $L_v + L_g + L_{cm}$, and the text encoder is updated by $L_t + L_{cm}$.

Experiments

Datasets and Evaluation Protocols

VeRi-776 (Liu et al. 2016b) is the benchmark dataset for vehicle Re-ID task, which includes 776 unique vehicles captured by 20 cameras. The entire dataset is divided into a training set consisting of 37,778 images from 576 vehicles and a testing set consisting of 11,579 images from 200 vehicles. **VehicleID** (Liu et al. 2016a) is a large-scale dataset for vehicle Re-ID, which includes 221,763 images of 26,267 vehicles. To accommodate the varying testing requirements across scales, the test set is subdivided into three subsets (Test800, Test1600, and Test2400) with separate sizes. **Evaluation Protocols:** we use mean accuracy precision (mAP) and rank-rate proposed by (Zheng et al. 2015) to evaluate the performance of the model.

Implementation Details

We employ a pre-trained CLIP-B/16 as our backbone. During the training phase, we train with batches of 64 and 50 epochs, each epoch consisting of 600 iterations. We employ the Adam optimizer to update model weights, setting the initial learning rate to decay by 10 times every 20 epochs. We utilize random erasure, random cropping, and random horizontal flipping, each with a 0.5 probability, as data augmentation techniques to enhance the dataset. All experiments are conducted on two NVIDIA RTX 3090 GPUs.

Ablation Study

Comparison of different modules. To verify the contributions of the proposed ViewCoOp and cross-modal mutual graph co-matching method (abbreviated as CMG), Table 1 reports the experimental performance of the two ablation modules. Significant performance improvements have been achieved on both VeRi-776 and VehicleID using only the ViewCoOp module, attributed to ViewCoOp to provide multi-view textual features to supplement visual features. In addition, compared to the "Baseline", the ablation setting using CMG increased Rank1 and mAP by 5.7% and 4.0% respectively, on VeRi-776. This performance improvement is due to CMG aligning visual-text pairs to explore positive and negative sample potential connections. When the two are integrated, performance is enhanced even further. This fully demonstrates the complementarity of the two ablation modules.

Effectiveness of different textual prompts. We further investigate the impact of using different textual prompts on the proposed framework and tabulate the experimental results in Table 2. It is worth noting that these text descriptions all set four learnable ID-specific vectors. The quantitative performance shows that "ViewCoOp(Static)" improved

Ablation Modules	VeRi-776			VehicleID(Test2400)		
	Rank1	Rank5	mAP	Rank1	Rank5	mAP
Baseline	80.3	84.0	36.8	56.0	69.9	41.5
w/ ViewCoOp	83.6	88.4	39.9	58.6	72.2	45.1
w/ CMG	84.3	89.9	39.2	58.8	71.4	45.7
Ours	**88.3**	**92.2**	**42.4**	**61.3**	**73.3**	**47.4**

Table 1: Ablation study of different modules. "Baseline" means using the image encoder directly for unsupervised vehicle Re-ID tasks without any ablation modules.

Textual Prompts	VeRi-776			VehicleID(Test2400)		
	Rank1	Rank5	mAP	Rank1	Rank5	mAP
CoOp(Zhou et al. 2022b)	84.3	89.9	39.2	58.8	71.4	45.7
CoCoOp(Zhou et al. 2022a)	85.6	91.4	40.3	60.5	72.1	46.0
UnCoOp(Huang and Chu 2022)	86.5	91.7	40.5	60.6	72.4	46.6
ViewCoOp(Static)	86.6	91.3	41.0	59.7	71.9	46.2
ViewCoOp(Dynamic)	**88.3**	**92.2**	**42.4**	**61.3**	**73.3**	**47.4**

Table 2: Comparison of using different textual prompts.

Threshold τ	VeRi-776			VehicleID(Test2400)		
	Rank1	Rank5	mAP	Rank1	Rank5	mAP
0	83.2	88.9	38.6	59.4	70.5	45.5
1	85.5	89.8	41.3	60.8	72.3	46.7
1 or 2	**88.3**	**92.2**	**42.4**	**61.3**	**73.3**	**47.4**
2	86.2	90.7	40.5	**61.3**	72.8	47.2

Table 3: Impact of modifying the threshold τ for connecting graph edges. "$\tau = 1$ **or** 2" means using at least one modal constrained dictionary index to connect edges.

by 1.0% and 0.7% respectively compared to CoCoOp on the VeRi-776. This suggests that embedding the view proportion into a textual prompt is capable of extracting view-invariance features. In addition, "ViewCoOp (Dynamic)" further introduces a learnable view bias term to dynamically update the view proportion for each epoch and improves Rank1 and mAP by 1.7% and 1.4%, respectively, on the VeRi-776. The above results clearly verify that "ViewCoOp (Dynamic)" removes the static view proportion limitation and provides more flexible view-aware textual annotations.

The impact of different connection edge thresholds. Table 3 compares various thresholds to explore the best performance of graph co-matching methods. It can be observed that the highest performance is achieved when the threshold value is $\tau = 1$ *or* 2. We speculate that there are two main reasons for this experimental result: (1) Only considering single modal constraints will limit the cross modal generalization ability between sample pairs. (2) Leveraging bimodal constraints will affect the feature alignment of the subsequent training phase due to the weak expression ability of textual and visual features during early training. Consequently, we integrate the above two constraint settings and select a threshold of $\tau = 1$ *or* 2 to ensure the effective construction of a cross-modal mutual graph.

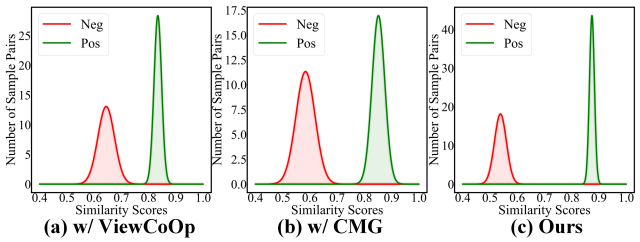

Figure 4: Visualization of distance distribution of sample pairs. Red line indicates a positive sample pair, while green line indicates a negative sample pair.

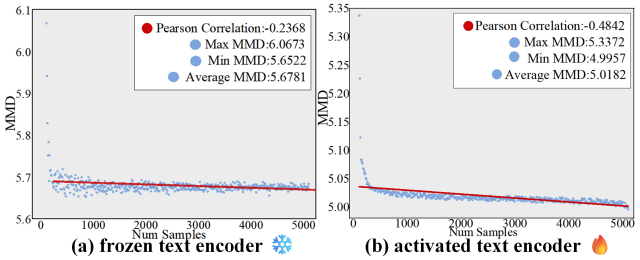

Figure 5: Visualization of text encoder activated to the influence of text-image pair feature differences. Blue scatter points denote the feature differences between the text-image pairs, while red line represents the Pearson Correlation coefficient of the same feature.

Comparison with state-of-the-arts

We compare "Ours" with state-of-the-art unsupervised learning Re-ID methods. Table 4 summarizes their performance in the two main stream vehicle Re-ID datasets, namely, VeRi-776 and VehicleID. Note that all data listed are without re-ranking. Despite being simple, "Ours" achieved competitive results.

Additionally, we explored various view ratio calculation strategies within ViewCoOp. The findings indicate that "Pixel-level" is unable to be adaptively embedded into ViewCoOp to obtain view-aware ability since precise pixel-level view extraction depends on segmentation models, leading to substantial computational errors. In contrast, using appropriately sized patches for calculating view ratios overcomes the limitations associated with segmentation models. When the image is segmented into 16 $\times$ 16 blocks, our method performs optimally, achieving 42.4% map and 88.3% Rank1 in VeRi-776.

Qualitative Visualization Analysis

Visualization of distance distribution of sample pairs. We visualized the cosine distance distribution of sample pairs. Specifically, we randomly selected 5000 samples from the VeRi-776 and constructed positive and negative sample pairs. From Figure 4, CMG outperforms ViewCoOp in distinguishing the distances between positive and negative sample pairs. Simultaneously, the peak values of both positive and negative sample pairs in ViewCoOp are higher than

Methods	References	VeRi-776			VehicleID								
					Test800			Test1600			Test2400		
		Rank1	Rank5	mAP	Rank1	Rank5	mAP	Rank1	Rank5	mAP	Rank1	Rank5	mAP
MMT(Ge et al. 2020a)*	ICLR'20	60.9	69.0	25.4	44.2	55.5	25.0	43.7	53.3	24.4	31.7	42.8	18.1
SPCL(Ge et al. 2020b)*	NIPS'20	65.6	74.0	28.3	70.1	77.6	53.8	66.8	74.5	51.2	59.7	72.0	44.9
ICE(Chen, Lagadec, and Bremond 2021)	ICCV'21	82.1	87.1	37.9	67.1	76.1	51.0	63.5	72.8	48.4	56.1	69.8	41.5
GCMT(Liu and Zhang 2021)*	IJCAI'21	79.0	86.1	34.8	69.8	77.8	53.4	65.8	74.2	50.9	60.3	72.3	45.4
ISE(Zhang et al. 2022b)	CVPR'22	66.0	72.5	27.7	69.6	77.6	54.0	66.4	74.1	51.3	60.3	72.0	45.2
CTFRN(Zheng et al. 2022)*	PR'22	76.7	81.5	37.1	70.1	77.8	54.7	67.2	73.6	51.7	60.5	71.5	45.6
PPLR(Cho et al. 2022)	CVPR'22	85.6	91.1	41.6	70.9	76.6	54.5	66.2	74.3	50.8	60.3	71.8	45.6
Cluster Contrast (Dai et al. 2022)	ACCV'22	86.2	90.5	40.8	67.2	75.5	50.5	63.1	71.3	47.9	56.2	67.8	41.3
AdaMG(Peng, Jiang, and Wang 2023)	TCSVT'23	86.2	90.6	41.0	-	-	-	-	-	-	-	-	-
Lan *et al.* (Lan et al. 2023)	TIP'23	78.5	84.8	35.1	68.6	76.6	52.7	63.6	73.5	48.6	57.7	71.2	42.9
TMGF(Li, Wang, and Gong 2023)	WACV'23	79.4	86.3	33.7	58.9	67.6	47.6	53.9	63.6	43.1	41.3	54.3	34.5
UCF(Wang et al. 2023a)*	TMM'23	85.2	89.2	40.5	69.6	77.5	54.4	65.8	73.8	50.4	59.6	71.6	45.2
STDA(He et al. 2024)	TITS'24	87.4	90.8	42.3	-	-	-	-	-	-	-	-	-
Pixel-level	Ours	86.2	91.3	40.5	70.2	77.3	55.6	65.2	73.7	52.0	59.4	70.8	45.7
Patch-level(8×8)	Ours	86.0	90.1	41.0	70.7	77.5	55.3	66.4	74.7	52.4	59.9	71.5	45.6
Patch-level(16×16)	**Ours**	**88.3**	**92.2**	**42.4**	**72.3**	**79.0**	**56.4**	**67.7**	**74.7**	**53.2**	**61.3**	**73.3**	**47.4**
Patch-level(32×32)	Ours	87.4	91.6	41.4	70.4	78.3	55.5	65.8	74.2	52.3	60.0	71.5	45.8

Table 4: Comparison with state-of-the-art unsupervised Re-ID methods. "*" represents unsupervised domain adaptation method. "Pixel-level" is using a pixel-level mask to count view regions for view proportion calculation. "Patch-level (s$\times$ s)" denotes dividing the image into s$\times$s patches to count each view regions for view proportion calculation, where $s = \{8,\ 16,\ 32\}$.

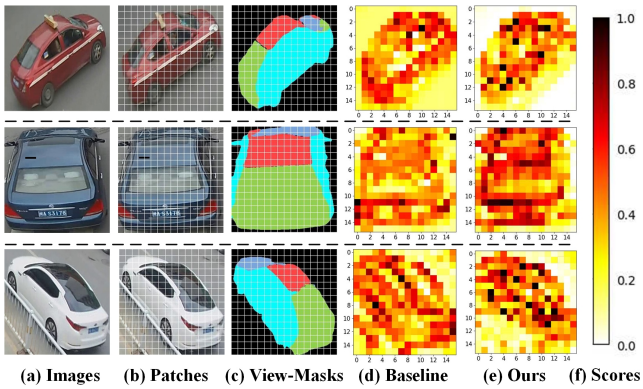

Figure 6: Visualization of multi-view correlation. The color jet bar denotes the correlation scores between the local patch to global features, with darker colors signifying a higher correlation with the global feature.

CMG. This indicates that ViewCoOp is adept at identifying vehicles of the same identity from different views, thereby pulling the distance for positive sample pairs. When the two are combined, i.e., "Ours", it can be seen that the distribution between positive and negative sample pairs is farther from the outside, and the distribution within positive and negative sample pairs is more compact.

The impact of activating the text encoder. We investigate the effect of activating the parameters of the text encoder during CLIP fine-tuning on cross-modal alignment. We randomly select 5000 samples from VeRi-776 and used maximum mean discrepancy (MMD) as the metrics to evaluate text-image pair feature differences. The results in Figure 5 show that when the text encoder is activated, MMD is reduced and the scatter plot distribution is smoother and more concentrated. The key is that the dynamic textual descriptions embedded in ViewCoOp in each epoch need to be aligned with the visual modality.

Visualization of multi-view correlations. To investigate the ability of the proposed method to extract multi-view features, we illustrate in Figure 6 the correlation between each patch and global image. We divide the image into patches of 16×16, and then calculate the cosine correlation between each patch feature and the global feature. We further obtain the segmented patch-level view-mask in Figure 6(c) to more intuitively observe the correlation between the patches contained in each view and the global features. The visualization indicates that, in comparison to the "Baseline", "Ours" produces darker colors in patches within different view regions. This indicates that "Ours" can capture information more relevant to the global features of the vehicle in each view.

Conclusion

This paper proposes a CLIP-driven view-aware prompt learning framework that combines the discriminative ability of textual and visual modalities to overcome the challenge of multiple view variations in unsupervised vehicle Re-ID tasks. Firstly, a ViewCoOp embedded with view region ratios and position encoding is proposed to obtain view-aware textual descriptions. Thereafter, we design a cross-modal mutual graph co-matching strategy to establish the correlation between textual and visual modalities by fine-tuning CLIP to obtain the view-invariant cross-modal capabilities. The experimental results on publicly available datasets demonstrate the superiority of the proposed method. We are convinced that the proposed method is not only limited to overcoming the view variations in vehicle Re-ID but can also be applied to other cross-modal tasks with multiple view objects in future studies.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62461037 and the Jiangxi Provincial Natural Science Foundation under Grant No. 20224BAB212011, 20232BAB202051, 20232BAB212008, and 20242BAB25078.

References

Chen, H.; Lagadec, B.; and Bremond, F. 2021. ICE: Inter-Instance Contrastive Encoding for Unsupervised Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14960–14969.

Cho, Y.; Kim, W. J.; Hong, S.; and Yoon, S.-E. 2022. Part-Based Pseudo Label Refinement for Unsupervised Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7308–7318.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022. Cluster Contrast for Unsupervised Person Re-Identification. In *Proceedings of the Asian Conference on Computer Vision*, 1142–1160.

Dong, N.; Yan, S.; Tang, H.; Tang, J.; and Zhang, L. 2024. Multi-view Information Integration and Propagation for occluded person re-identification. *Information Fusion*, 104: 102201.

Ge, Y.; Chen, D.; Zhao, R.; and Li, H. 2020a. Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. In *International Conference on Learning Representations*.

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; and Li, H. 2020b. Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

He, Q.; Lu, Z.; Wang, Z.; and Hu, H. 2023. Graph-Based Progressive Fusion Network for Multi-Modality Vehicle Re-Identification. *IEEE Transactions on Intelligent Transportation Systems*, 24(11): 12431–12447.

He, Q.; Wang, Z.; Zheng, Z.; and Hu, H. 2024. Spatial and Temporal Dual-Attention for Unsupervised Person Re-Identification. *IEEE Transactions on Intelligent Transportation Systems*, 25(2): 1953–1965.

Huang, T.; and Chu, F., Jack andWei. 2022. Unsupervised Prompt Learning for Vision-Language Models. arXiv:2204.03649.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 4904–4916.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Khorramshahi, P.; Shenoy, V.; and Chellappa, R. 2023. Robust and Scalable Vehicle Re-Identification via Self-Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 5295–5304.

Lan, L.; Teng, X.; Zhang, J.; Zhang, X.; and Tao, D. 2023. Learning to Purification for Unsupervised Person Re-Identification. *IEEE Transactions on Image Processing*, 32: 3338–3353.

Li, H.; Wang, Y.; Wei, Y.; Wang, L.; and Li, G. 2023a. Discriminative-region attention and orthogonal-view generation model for vehicle re-identification. *Applied Intelligence*, 53(1): 186–203.

Li, J.; Wang, M.; and Gong, X. 2023. Transformer Based Multi-Grained Features for Unsupervised Person Re-Identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 1–9.

Li, M.; Liu, J.; Zheng, C.; Huang, X.; and Zhang, Z. 2023b. Exploiting Multi-View Part-Wise Correlation via an Efficient Transformer for Vehicle Re-Identification. *IEEE Transactions on Multimedia*, 25: 919–929.

Li, S.; Sun, L.; and Li, Q. 2023. CLIP-ReID: Exploiting Vision-Language Model for Image Re-identification without Concrete Text Labels. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1): 1405–1413.

Li, Y.; Liu, K.; Jin, Y.; Wang, T.; and Lin, W. 2022a. VARID: Viewpoint-Aware Re-IDentification of Vehicle Based on Triplet Loss. *IEEE Transactions on Intelligent Transportation Systems*, 23(2): 1381–1390.

Li, Z.; Tang, C.; Liu, X.; Zheng, X.; Zhang, W.; and Zhu, E. 2022b. Consensus Graph Learning for Multi-View Clustering. *IEEE Transactions on Multimedia*, 24: 2461–2472.

Liu, H.; Tian, Y.; Wang, Y.; Pang, L.; and Huang, T. 2016a. Deep Relative Distance Learning: Tell the Difference between Similar Vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2167–2175.

Liu, Q.; He, X.; Teng, Q.; Qing, L.; and Chen, H. 2023. BDNet: A BERT-based dual-path network for text-to-image cross-modal person re-identification. *Pattern Recognition*, 141: 109636.

Liu, X.; Liu, W.; Mei, T.; and Ma, H. 2016b. A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance. In *European Conference on Computer Vision*, 869–884.

Liu, X.; and Zhang, S. 2021. Graph Consistency Based Mean-Teaching for Unsupervised Domain Adaptive Person Re-Identification. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 874–880.

Lu, Z.; Lin, R.; He, Q.; and Hu, H. 2023. Mask-Aware Pseudo Label Denoising for Unsupervised Vehicle Re-Identification. *IEEE Transactions on Intelligent Transportation Systems*, 24(4): 4333–4347.

Meng, D.; Li, L.; Liu, X.; Li, Y.; Yang, S.; Zha, Z.-J.; Gao, X.; Wang, S.; and Huang, Q. 2020. Parsing-Based View-Aware Embedding Network for Vehicle Re-Identification. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7101–7110.

Miyai, A.; Yu, Q.; Irie, G.; and Aizawa, K. 2024. LoCoOp: few-shot out-of-distribution detection via prompt learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Peng, J.; Jiang, G.; and Wang, H. 2023. Adaptive Memorization With Group Labels for Unsupervised Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10): 5802–5813.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 8748–8763.

Tan, W.; Ding, C.; Jiang, J.; Wang, F.; Zhan, Y.; and Tao, D. 2024. Harnessing the Power of MLLMs for Transferable Text-to-Image Person ReID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17127–17137.

Wang, P.; Ding, C.; Tan, W.; Gong, M.; Jia, K.; and Tao, D. 2023a. Uncertainty-Aware Clustering for Unsupervised Domain Adaptive Object Re-Identification. *IEEE Transactions on Multimedia*, 25: 2624–2635.

Wang, Q.; Zhong, Y.; Min, W.; Zhao, H.; Gai, D.; and Han, Q. 2023b. Dual similarity pre-training and domain difference encouragement learning for vehicle re-identification in the wild. *Pattern Recognition*, 139: 109513.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2023. CLIP-Driven Fine-Grained Text-Image Person Re-Identification. *IEEE Transactions on Image Processing*, 32: 6032–6046.

Zhai, Y.; Zeng, Y.; Huang, Z.; Qin, Z.; Jin, X.; and Cao, D. 2024. Multi-Prompts Learning with Cross-Modal Alignment for Attribute-Based Person Re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7): 6979–6987.

Zhang, C.; Yang, C.; Wu, D.; Dong, H.; and Deng, B. 2022a. Cross-view vehicle re-identification based on graph matching. *Applied Intelligence*, 52(13): 14799–14810.

Zhang, X.; Li, D.; Wang, Z.; Wang, J.; Ding, E.; Shi, J. Q.; Zhang, Z.; and Wang, J. 2022b. Implicit Sample Extension for Unsupervised Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7369–7378.

Zhao, Z.; Liu, B.; Lu, Y.; Chu, Q.; and Yu, N. 2024. Unifying Multi-Modal Uncertainty Modeling and Semantic Alignment for Text-to-Image Person Re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7): 7534–7542.

Zheng, A.; Zhang, C.; Li, C.; Tang, J.; and Tan, C. 2023. Multi-Query Vehicle Re-Identification: Viewpoint-Conditioned Network, Unified Dataset and New Metric. *IEEE Transactions on Image Processing*, 32: 5948–5960.

Zheng, D.; Xiao, J.; Chen, K.; Huang, X.; Chen, L.; and Zhao, Y. 2022. Soft pseudo-Label shrinkage for unsupervised domain adaptive person re-identification. *Pattern Recognition*, 127: 108615.

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.

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*, 16795–16804.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*.