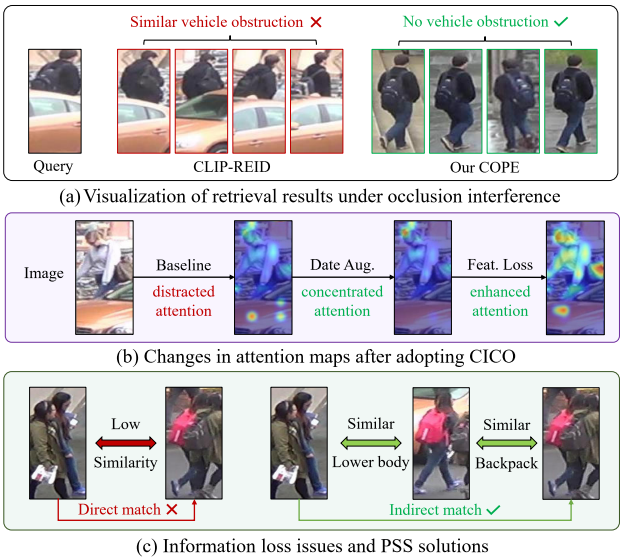
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

COPE: Consistent Occlusion and Prompt Enhancement Network for Occluded Person Re-identification

Siyi Sun[1], Jinliang Lin[1], Juanjuan Weng[2], Zhihui Liu[3], Shaozi Li[1,4], Zhiming Luo[1,4*]
[1] Department of Artificial Intelligence, Xiamen University, Xiamen, China
[2] College of Information Science and Technology, Jinan University, Guangzhou, China
[3] Truesight Technology Co. Ltd, Xiamen, China
[4] Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, Xiamen, China

Abstract

Occlusion presents two critical challenges for person re-identification (Re-ID): feature interference and information loss. While existing efforts have explored occlusion-aware data augmentation and feature reconstruction to mitigate these issues, the former often fails to address erroneous matches caused by similar occlusion patterns and background distractions, whereas the latter typically introduces significant computational overhead. To overcome these limitations, we propose a Consistent Occlusion and Prompt Enhancement (COPE) network. COPE incorporates a Cross-Identity Consistent Occlusion (CICO) module that applies identical occlusions across different identities and encourages feature similarity in the same occluded regions across different identities to reduce occlusion feature interference. A Prompt Background Filling (PBF) module leverages vision-language alignment to generate foreground heatmaps and performs random background filling, enhancing feature robustness under varying backgrounds. Additionally, a lightweight Prompt Similarity Scoring (PSS) module refines retrieval similarity by utilizing prompt-guided reliability scores. Extensive experiments on both occluded and holistic Re-ID benchmarks demonstrate that COPE consistently outperforms existing methods. Notably, it achieves 82.4% Rank-1 accuracy and 76.4% mAP on the challenging Occluded-Duke dataset. The code is available at https://github.com/Cecoming/COPE.



(a) Visualization of retrieval results under occlusion interference

(b) Changes in attention maps after adopting CICO

(c) Information loss issues and PSS solutions

Figure 1. Existing challenges on Occluded-Duke dataset and our solutions. (a) shows incorrect matches with similar occlusions for CLIP-REID and clean correct matches for our COPE under occlusion interference. (b) shows the attention map of the query with different CICO settings. Without CICO, attention is scattered; with data augmentation, attention on the occlusion decreases; with feature loss, attention on the target person increases. (c) shows the difficult matching problem caused by information loss, which we address by using similar intermediate samples.

1. Introduction

In practice, pedestrian images frequently suffer from occlusions caused by surrounding objects, posing two major challenges for accurate person Re-ID: 1) extracted features are easily disturbed by occlusions, and 2) significant portions of discriminative information are lost. With the rapid development of deep learning, the feature extraction of existing models has steadily improved. To enhance feature alignment, some methods [12, 13] attempt to extract global features, while others [9, 24, 34] leverage external cues to extract part-based features. However, due to persistent interference and missing information caused by occlusions,

*Corresponding author: zhiming.luo@xmu.edu.cn

these methods have reached a performance bottleneck.

Recent works have explored two main directions to address these issues: data augmentation and feature reconstruction. Data augmentation-based methods [1, 21] simulate various occlusions to enhance model generalization, while feature reconstruction-based methods [8, 32] restore missing features using external cues or gallery neighbors. However, these approaches overlook several important aspects. (1) Most augmentation methods emphasize diversity but ignore the fact that similar occlusions often induce misleading similarity. As shown in Figure 1(a), when querying occluded samples, retrieval results of transformer-based CLIP-REID [12] showed similar vehicle occlusions, which proves the existence of similar occlusion interference. (2) Most gallery images in occluded person Re-ID remain holistic. For these images, simple occlusion simulation cannot fully resolve background distraction issues in different environments. (3) Reconstruction-based methods typically entail high computational costs due to their reliance on similar neighbor features for feature recovery, resulting in high inference latency and poor scalability in real-time deployment scenarios.

To address these challenges, we propose a novel framework named **Consistent Occlusion and Prompt Enhancement Network (COPE)**. Specifically, to alleviate the interference of recurrent occlusions across different pedestrians, we propose the **Cross-Identity Consistent Occlusion (CICO)** module. At the data level, CICO applies identical occlusion group to different identities to simulate visually misleading scenarios. At the feature level, Transformer-based backbones rely on global self-attention between all patch tokens. Without guidance, occlusion tokens may undesirably interact with foreground tokens, degrading feature quality. To address this, we propose an occlusion consistency loss that encourages feature similarity in the same occluded regions across different identities. This alignment helps the model recognize occlusion-specific tokens and suppress their impact during attention, thereby reducing occlusion interference and shifting attention toward informative foreground regions. As shown in Fig. 1(b), after data augmentation and feature constraints using CICO, the attention to the occluded region is significantly reduced, and the representation of the target region is further enhanced.

To alleviate background inference in holistic pedestrian images, we propose the **Prompt Background Filling (PBF)** module. By leveraging vision-language alignment, PBF generates foreground heatmaps that accurately localize pedestrian regions. The remaining background areas are then randomly filled to simulate diverse environmental contexts. To ensure consistent foreground feature extraction between the original and augmented images, a foreground alignment loss is applied. This design encourages the model to focus on identity-relevant foreground features while reducing its reliance on background cues.

Moreover, to resolve information loss from severe occlusion, we design a **Prompt Similarity Scoring (PSS)** module. As shown in Fig. 1(c), direct matching becomes challenging when query regions are only partially visible. However, if a reliable reference is available, two low-similarity samples can still be associated through shared visible parts (e.g., lower body, backpack). To measure the reliability of an instance, we compute its similarity to the class center, defined as a prompt score. Since feature reliability is closely related to foreground integrity, we use the PBF prompt heatmaps to learn prompt scores. During retrieval, PSS selects high-prompt-score gallery samples to refine query-gallery similarity. Compared to common methods for mitigating information loss, our approach optimizes matching using only neighbor feature similarity without feature reconstruction, greatly reducing inference cost.

The main contributions are summarized as follows:

- We propose the CICO module to simulate realistic occlusion scenarios and enforce feature alignment in occluded regions, effectively mitigating feature interference caused by occlusion.
- We introduce the PBF module, a prompt-guided augmentation strategy, which improves foreground feature extraction consistency across diverse backgrounds.
- We design a lightweight PSS module that refines retrieval results during inference using prompt-guided reliability cues, achieving substantial performance improvements with negligible inference overhead.
- Comprehensive experiments demonstrate that COPE achieves state-of-the-art performance on both occluded and holistic person Re-ID benchmarks.

2. Related Work

2.1. Occluded Person Re-Identification

To mitigate occlusion interference and information loss, recent methods primarily explore data augmentation and feature reconstruction. **Data augmentation** approaches simulate occlusions at the input level via noise injection [1], pasting diverse obstacles cropped from external datasets [21, 27], or transferring salient patches to new backgrounds [25]. However, these methods often emphasize occlusion diversity while overlooking the misleading interference from similar occlusion patterns and the need to augment full-body gallery images. Conversely, **feature reconstruction** methods focus on recovering missing features post-extraction. Approaches like RFCnet [8], FRT [31], and FPC [32] prune invalid regions and leverage external cues or k-nearest gallery neighbors to restore semantic features. This neighbor reliance typically introduces significant computational overhead, limiting real-time scalability.

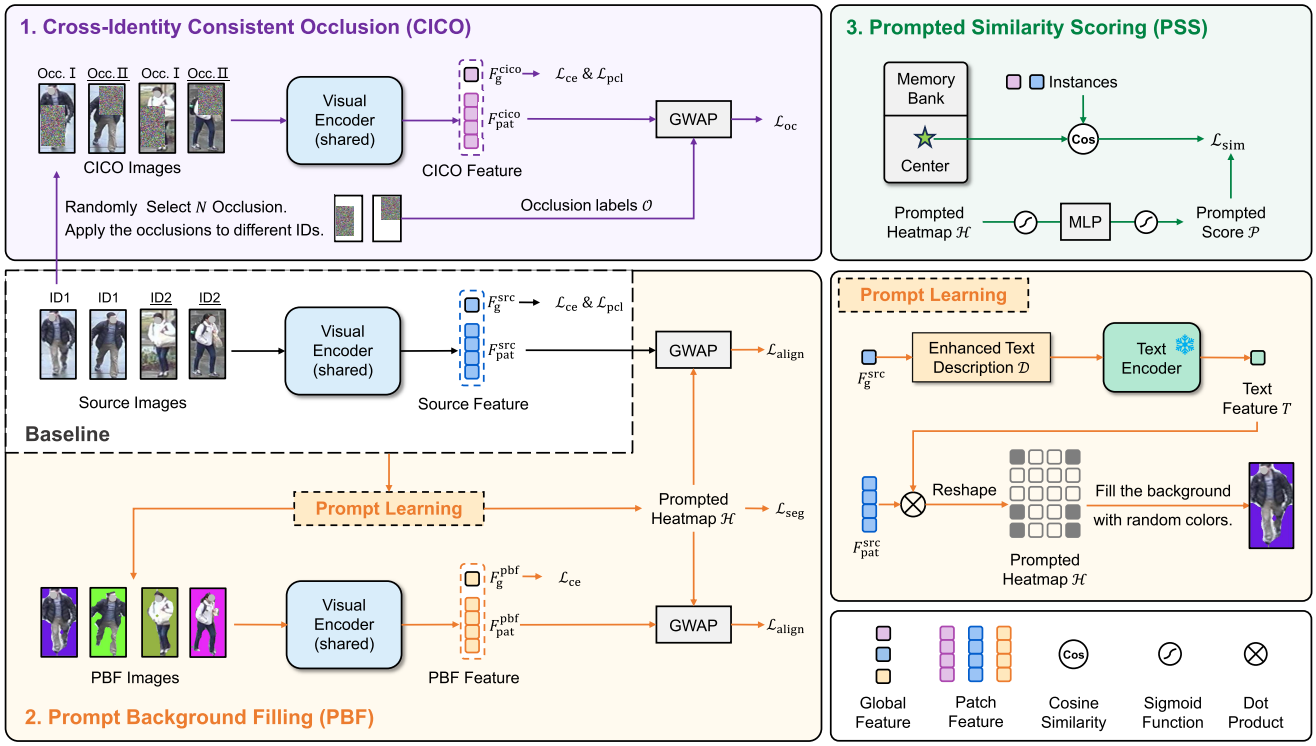


Figure 2. The training of the proposed Consistent Occlusion and Prompt Enhancement (COPE) Network. First, the source image is passed through the Visual Encoder to extract the original image features. Then, data augmentation is applied through the Cross-Identity Consistent Occlusion (CICO) and Prompt Background Filling (PBF) modules, which introduce occlusion and background filling, respectively, to extract corresponding enhanced features. Finally, in the Prompt Similarity Scoring module, the prompt scores are learned based on the similarity between features instances and their corresponding prototypes center.

2.2. Vision-Language Learning

Recent advances in vision-language models like CLIP [18] have demonstrated robust cross-modal representation alignment, which is further enhanced by learnable [40] and image-conditioned prompts [39]. In occluded Re-ID, language cues have been effectively leveraged to guide body region localization [7] and optimize feature representations via prototype contrastive learning [10]. Building upon this foundation of textual alignment and contrastive learning, our approach further refines prompt-guided representations to explicitly address the unique challenges of occlusion and background interference in Re-ID scenarios.

3. Proposed Method

The proposed COPE consists of two stages: training (Fig. 2) and inference (Fig. 3). The training stage mainly involves three key modules: (1) **Cross-Identity Consistent Occlusion (CICO)**, (2) **Prompt Background Filling (PBF)**, and (3) **Prompt Similarity Scoring (PSS)**. During inference, the PSS module utilizes the learned prompt scores to refine retrieval similarity and improve matching performance.

3.1. Baseline

We use the CLIP ViT [18] as the visual encoder. For an input source image, it extracts a class token and patch tokens as the global feature $F_{\mathrm{g}}^{\mathrm{src}} \in R^{D}$ and image feature $F_{\mathrm{pat}}^{\mathrm{src}} \in R^{N \times D}$, respectively. Similarly, augmented images from CICO and PBF produce $F_{\mathrm{g}}^{\mathrm{cico}}$, $F_{\mathrm{pat}}^{\mathrm{cico}}$ and $F_{\mathrm{g}}^{\mathrm{pbf}}$, $F_{\mathrm{pat}}^{\mathrm{pbf}}$. Following PCL-CLIP [10], we adopt prototype-based contrastive learning (PCL) [11] to supervise feature learning. The memory bank is initialized with global features from source images and dynamically updated using both CICO and source global features to balance optimization between occluded and holistic conditions. The prototype-based contrastive learning are as follows:

$$\boldsymbol{F_g}[i] = \{\boldsymbol{F_g}^{\mathrm{src}}[i], \boldsymbol{F_g}^{\mathrm{cico}}[i]\}, \tag{1}$$

$$\mathcal{L}_{\mathrm{pcl}} = -\log \frac{\exp\big(s(\mathcal{K}[i], \mathbf{F}_g[i])/\tau\big)}{\sum_{j=1}^{C} \exp\big(s(\mathcal{K}[j], \mathbf{F}_g[i])/\tau\big)}, \tag{2}$$

where $\boldsymbol{F_g}^{\mathrm{src}}[i]$ and $\boldsymbol{F_g}^{\mathrm{cico}}[i]$ are the i-th class global features of source images and CICO images, $\mathcal{K}[i]$ is the i-th class feature in the memory bank.

3.2. Cross-Identity Consistent Occlusion

The Cross-Identity Consistent Occlusion (CICO) module is designed to suppress the adverse effects of shared occlusion patterns across identities. It operates by applying the same occlusion group to different identities, simulating ambiguous occlusion cases. Specifically, M random Gaussian-shaped occlusions are pre-initialized. In each batch, N occlusion types are sampled and applied to N identities, ensuring each occlusion type is shared across different persons while maintaining diversity for intra-class robustness.

In Transformer-based backbones, all image tokens participate in global self-attention. Without constraint, occlusion tokens may entangle with foreground tokens, introducing harmful interactions. To alleviate this, we introduce an occlusion consistency loss to explicitly guide the model to identify and downweight occlusion tokens. By enforcing occlusion feature similarity across instances sharing the same occlusion type, the model learns a consistent representation for occluded regions, reducing their relevance to identity cues and naturally shifting attention toward visible regions. Concretely, we extract the occlusion feature via Global Weighted Average Pooling (GWAP). Given the patch-level occlusion mask $\mathcal{O}_{\text{pat}} \in \mathbb{R}^{N\times H\times W}$, the occlusion feature of the i-th image under the n-th occlusion is:

$$\begin{aligned} F_n^{\text{cico}}(i) &= \text{GWAP}(\mathcal{O}_n, F_{\text{pat}}^{\text{cico}}) \\ &= \frac{\sum_h^{fH}\sum_w^{fW} \mathcal{O}_n(h,w) F_{\text{pat}}^{\text{cico}}(i,h,w)}{\sum_h^{fH}\sum_w^{fW} \mathcal{O}_n(h,w)} \end{aligned} \tag{3}$$

Then, we compute the mean pairwise MSE distance of occlusion features within each occlusion type:

$$\mathcal{L}_{\text{oc}} = \sum_{n=1}^{N} \frac{1}{|I_n|^2} \sum_{i,j\in I_n} \left\| F_n^{\text{cico}}(i) - F_n^{\text{cico}}(j) \right\|^2, \tag{4}$$

where I_n denotes images augmented with the n-th occlusion. This encourages consistent feature encoding of occlusion tokens, implicitly suppressing their attention weights and enhancing the foreground representation learning.

3.3. Prompt Background Filling

In addition to simulating occlusion scenarios, we enhance the model's foreground feature extraction through the Prompt Background Filling (PBF) module. Given CLIP's powerful localization capabilities in various occluded pedestrian ReID methods [2, 7], we incorporate a frozen Text Encoder, with text descriptions that guide the model to the foreground region. Specifically, we use $\{\boldsymbol{v}, \text{person}\}$ as the text input, where $\boldsymbol{v} = \{v_1, v_2, v_3, v_4\}$ are learnable tokens. Then, the image feature $F_{\text{g}}^{\text{src}}$ is combined with the learnable prompt to generate enhanced text descriptions $\mathcal{D} = \{\boldsymbol{v'}, \text{person}\}$ following CoCoOp [39], where $\boldsymbol{v'} = \text{MLP}(F_{\text{g}}^{\text{src}}) + \boldsymbol{v}$. The Text Encoder $\mathcal{T}(\cdot)$ then produces text feature $T = \mathcal{T}(\mathcal{D})$, which is compared with the foreground patch features of the image using cosine similarity to yield a prompt heatmap $\mathcal{H} = T \cdot F_{\text{pat}}^{\text{src}}$.

Next, we introduce human parsing labels $\hat{\mathcal{H}}$ [20] and apply the segmentation loss $\mathcal{L}_{seg}$ to supervise the prompted heatmap as follows:

$$\mathcal{L}_{\text{seg}} = \sum_{h}^{H}\sum_{w}^{W}\sum_{j=0}^{1} \mathbb{I}\left[\hat{\mathcal{H}}^{(x,y)} = j\right] \log \text{Up}(\mathcal{H})_j^{(h,w)}, \tag{5}$$

where $\text{Up}(\cdot)$ indicates up-sampling the heatmap from the patch level to the pixel level.

Once the foreground position is identified, we fill the background with random colors, simulating images of pedestrians in different environments. To ensure that the model accurately extracts foreground features under various environmental disturbances, we introduce a foreground alignment loss $\mathcal{L}_{\text{align}}$. For the invariant foreground regions in both the input and augmented images, we use the prompted heatmap as a weighting mask for global weighted average pooling of the foreground regions. We then align the foreground features by MSE loss, denoted as:

$$\mathcal{L}_{\text{align}} = \left\| \text{GWAP}(\mathcal{H}, f_{\text{pat}}^{\text{src}}) - \text{GWAP}(\mathcal{H}, f_{\text{pat}}^{\text{rbf}}) \right\|^2. \tag{6}$$

3.4. Prompt Similarity Scoring

Despite the improvements introduced by CICO and PBF, severe occlusions still lead to retrieval failures due to misaligned or severely limited visible regions. To address this, we propose a **Prompt Similarity Scoring (PSS)** strategy. For human images, the integrity of the foreground area is highly correlated with recognition reliability. More visible parts generally enable more reliable matching. To capture this, we transform the visual foreground probability from the prompted heatmap produced by PBF into a prompt score $\mathcal{P} \in R^1$ by a multi-layer perceptron (MLP):

$$\mathcal{P} = \sigma(\text{MLP}(\sigma(\mathcal{H}))). \tag{7}$$

where σ denotes the sigmoid activation.

Meanwhile, the reliability of features is reflected by their similarity to corresponding class centers. During training, we compute the cosine similarity between each instance feature and its corresponding class center using a memory bank. Then, the prompt score is supervised by minimizing its discrepancy with this similarity score via a MSE loss:

$$\mathcal{S}[i] = \text{Cos}(F_g[i], \mathcal{K}[i]), \tag{8}$$

$$\mathcal{L}_{\text{sim}} = \left\| \mathcal{P} - \mathcal{S} \right\|^2, \tag{9}$$

where $\mathcal{S}[i]$ is the cosine similarity between the i-th class feature center and the corresponding image feature.

3.5. Inference

In inference, following standard Re-ID settings [16], the Euclidean distance between the query feature F_Q and each gallery feature F_G is first computed. The distances are then converted into gallery similarities:

$$\mathcal{S}_G = \frac{1}{1 + D(F_Q, F_G)}. \tag{10}$$

To incorporate occlusion-awareness, we further weigh each gallery similarity with its corresponding prompt score $\mathcal{P}$ as illustrated in Fig. 3, resulting in the prompt similarity:

$$\mathcal{S}_\mathcal{P} = \mathcal{S}_G \cdot \mathcal{P}. \tag{11}$$

We select the top-K_1 samples from $\mathcal{S}_G$ as candidate gallery samples, with features denoted as F_1, and the top-K_2 samples from $\mathcal{S}_\mathcal{P}$ as intermediate reference samples, with features denoted as F_2. Similarly, we compute intermediate similarities between candidate and intermediate samples. Multiplying the intermediate similarity by the previous top-K_2 prompt similarities $\mathcal{S}_\mathcal{P}^{K_2}$ yields the similarity increment:

$$\mathcal{S}_{\text{inter}} = \frac{1}{1 + D(F_1, F_2)}. \tag{12}$$

$$\Delta = \frac{1}{K_2}(\mathcal{S}_{\text{inter}} \times \mathcal{S}_\mathcal{P}^{K_2}) \tag{13}$$

where division by K_2 is to reduce the influence of the number of intermediate samples. The final similarity score for each candidate is refined as:

$$\mathcal{S}^i = \mathcal{S}_G^i + \Delta^i, \quad \text{for } i = 1, \ldots, K_1. \tag{14}$$

We retain only the updated similarity vector $\mathcal{S} \in \mathbb{R}^{K_1}$ for subsequent ranking. These values are converted back to distances via $D' = \frac{1}{\mathcal{S}} - 1$, to remain compatible with standard retrieval procedures.

3.6. Objective Function

Following the standard ReID settings [16], we leverage a cross-entropy loss $\mathcal{L}_{\text{ce}}$ to supervise the global features: F_g^{src}, F_g^{cico}, and F_g^{rbf}. The overall training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{pcl}} + \mathcal{L}_{\text{oc}} + \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{sim}}. \tag{15}$$

4. Experiment

4.1. Datasets and Experimental Setting

Datasets. We evaluate our method on two holistic and four occluded ReID datasets. The holistic datasets include: **Market-1501** [35] and **MSMT17** [29]. The occluded datasets include: **Occluded-Duke** [17], **P-Duke-REID** [19], **Occluded-ReID** [41], and **Partial-REID** [36].

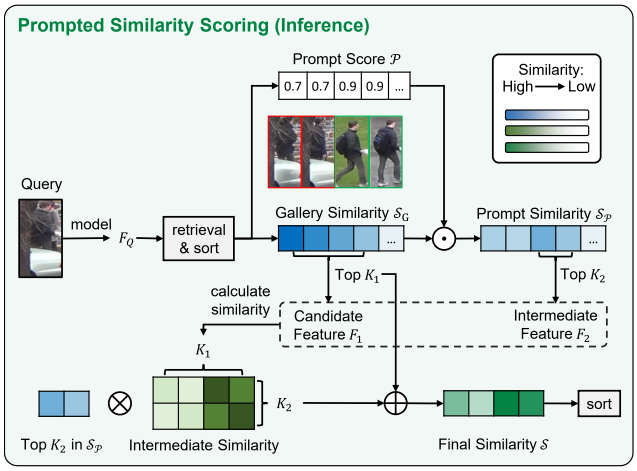


Figure 3. During inference, the Prompt Similarity Scoring module uses prompt scores to further refine final similarity.

Following prior works [4, 23, 24], we train on Market-1501 and test on Occluded-ReID and Partial-REID.

Evaluation Protocol. To ensure a fair comparison with other methods, we adopt the widely used Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) as evaluation metrics, following the evaluation settings provided by existing occluded Re-ID methods.

Implementation Details. We follow the setup in CLIP-ReID [12], using a ViT-based CLIP backbone. The original and reduced features are concatenated as retrieval features. Input images are resized to 256×128, with patch size 16×16 and sampling strides of 16 and 12. During training, images undergo random flipping and cropping. The batch size is 64, with 4 images per identity. We use the SGD with an initial learning rate of 3.5e-4, which is reduced by 0.1 at 30%, 50%, and 70% of the training cycle. For cross-domain training, Occluded-REID and Partial-REID training continues for 60 epochs, while other datasets continue for 100 epochs. The PCL loss temperature τ is set to 0.05 following PCL-CLIP [10]. The CLIP text encoder and the image encoder embedding layer are frozen during both training and inference. In CICO, M and N are set to 20 and 2, respectively. The PBF module is pre-trained for 60 epochs with the segmentation loss. In PSS, K_1 and K_2 are set to 200 and 5. All experiments are conducted on one NVIDIA V100 GPU.

4.2. Comparison with the State of the Arts

We categorize existing occluded person ReID methods into four groups based on their backbones. All comparison methods use input images with the same resolution of 256×128. As shown in Tab. 1, Transformer-based approaches consistently outperform ResNet-50-based methods. CLIP-ViT-based models achieve comparable performance to ViT-based models on occluded datasets, while

Table 1. Comparison with state-of-the-art methods on occluded and holistic person re-identification benchmarks. **Bold** and underline indicate optimal and sub-optimal results, respectively. $*$ indicates that the backbone network uses an overlapping sliding window setting with a stride size of 12 or less. † indicates results reproduced by us.

Type			Occluded								Holistic			
Datasets			Occ-DUKE		Occ-REID		P-Duke		P-REID		Market1501		MSMT17	
Methods	Source	Backbone	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP
PCB[22]	ECCV 18	ResNet-50	42.6	33.7	41.3	38.9	-	-	66.3	63.8	93.8	81.6	40.4	68.2
PVPM[4]	CVPR 20		47.0	37.7	70.4	61.2	51.5	29.2	78.3	72.3	94.0	82.5	-	-
HOReID[26]	CVPR 20		55.1	43.8	80.3	70.2	-	-	85.3	80.3	94.2	84.9	-	-
OAMN[1]	ICCV 21		62.6	46.1	-	-	-	-	86.0	77.4	93.2	79.8	-	-
BPBreID[20]	WACV 23		66.7	54.1	76.9	68.6	91.0	77.8	-	-	95.1	87.0	-	-
PAT[13]	CVPR 21	ViT	64.5	53.6	81.6	72.1	-	-	88.0	-	95.4	88.0	-	-
TransReID*[6]	CVPR 21		66.4	59.2	81.3	72.2	-	-	83.0	77.5	95.2	88.9	85.3	67.4
FED[28]	CVPR 22		68.1	56.4	86.3	79.3	-	-	84.6	82.3	95.0	86.9	-	-
FODN*[15]	MM 23		72.6	61.9	87.1	80.8	-	-	-	-	95.6	89.0	-	-
ADP*[30]	AAAI 24		74.5	63.8	89.2	85.1	-	-	-	-	95.9	89.6	-	-
FPC*[32]	AAAI 24		76.7	72.8	86.3	84.6	-	-	86.3	86.5	95.1	91.4	-	-
RGANet[7]	TIFS 23	CLIP ViT	71.6	62.4	86.4	80.0	-	-	87.2	-	95.5	89.8	88.1	72.3
CLIP-ReID*[12]	AAAI 23		67.2	60.3	-	-	-	-	-	-	95.4	90.5	89.7	75.8
PCL-CLIP[10]†	arXiv 23		70.4	63.3	86.5	85.0	91.9	84.8	85.0	84.4	95.7	91.4	88.5	73.9
ProFD*[2]	MM 24		70.6	63.1	92.3	90.3	92.8	84.7	-	-	95.6	90.8	-	-
SGFNet*[14]	TIFS 25		76.5	67.2	**93.0**	90.3	94.1	86.3	-	-	96.2	91.1	**90.0**	74.4
KPR[21]	ECCV 24	Swin	79.8	67.1	85.4	79.1	-	-	86.0	-	95.9	89.6	-	-
COPE	Ours	CLIP ViT	82.1	75.4	88.7	88.2	93.3	88.4	**88.0**	**88.7**	96.2	93.0	89.0	76.4
COPE*			**82.4**	**76.4**	90.9	**90.7**	**94.3**	**89.5**	87.0	87.4	**96.3**	**93.3**	89.5	**77.2**

showing superior performance on holistic datasets. Notably, the recent Swin-Tranformer-based KPR method [21] achieves strong performance.

Occluded ReID Datasets. Our COPE achieves SOTA results on most occluded datasets. Specifically, it surpasses previous leading methods by +2.6%/+3.6% on Occluded-Duke and +0.2%/+3.2% on P-Duke-REID in Rank-1/mAP. The COPE's effectiveness stems from two key advantages: (1) It incorporates occlusion-aware and background-aware augmentations at the data level, and employs loss functions that encourage robust and discriminative learning in feature level, even under similar occlusions and random backgrounds. (2) It adopts a prompt-guided re-ranking strategy, which refines retrieval results by leveraging high prompt scores, thereby mitigating occlusion effects more effectively than methods FED [28] and FPC [32] that also utilize k-nearest neighbors information. On cross-domain benchmarks, COPE ranks second in Rank-1 on Occluded-REID while achieving the best performance on all other metrics, confirming its strong generalization ability. This is attributed to the use of external cue-based enhancement, enabling the model to focus on foreground features even under domain shifts.

Holistic ReID Datasets. COPE achieves SOTA performance on holistic datasets. On Market-1501, it surpasses previous best methods by +0.1%/+1.9% in Rank-1/mAP, indicating that its robustness extends beyond occluded scenarios. On the more challenging MSMT17, COPE outperforms other methods in mAP, further validating its effectiveness.

4.3. Ablation Study and Analysis

Analysis of proposed components. The results on the Occluded-Duke are shown in Tab. 2. Compared to the CLIP-based baseline in Index-1, CICO improves Rank-1/mAP by +4.6%/+7.6% in Index-2, indicating that CICO enhances the model's robustness to similar occlusion patterns. Next, from **Index-3**, PBF further improves performance by enabling the model to better localize and extract foreground features under complex backgrounds, leading to stronger discriminative representation and improved retrieval performance. Finally, comparing Index-3 and **Index-4**, PSS module contributes an additional gain of +5.3%/+6.6% in Rank-1/mAP. By leveraging foreground localization cues from PBF, PSS refines retrieval results through post-processing, helping the model make better use of global image information to identify more accurate matches.

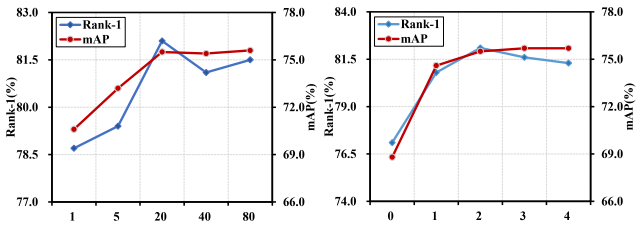
Analysis of proposed loss functions. As shown in Tab. 3, **Index-2** removes the occlusion consistency loss and leads to a 2.9% Rank-1 and 2.8% mAP drop, showing that

Table 2. Ablation study of the proposed components in COPE on Occluded-Duke dataset.

Index	CICO	PBF	PSS	Rank-1	*m*AP
1				70.2	60.3
2	✓			74.8(+4.6)	67.9(+7.6)
3	✓	✓		76.8(+2.0)	68.9(+1.0)
4	✓	✓	✓	**82.1**(+5.3)	**75.4**(+6.5)

Table 3. Ablation study of the proposed loss functions in COPE on Occluded-Duke dataset.

Index	Setting	Rank-1	*m*AP
1	COPE(Full)	82.1	75.4
2	1 - $\mathcal{L}_{\text{oc}}$	79.3 (-2.9)	72.7 (-2.7)
3	1 - $\mathcal{L}_{\text{align}}$	80.9 (-1.2)	75.0 (-0.4)



(a) Impact of predefined number of occlusions M (b) Impact of number of occlusions per batch N

Figure 4. Sensitivity analysis of hyperparameters M and N in CICO module on Occluded-Duke dataset.

Table 4. Sensitivity analysis of hyperparameters K_1 and K_2 in PSS on Occluded-Duke dataset. Inf-Time represents the inference time for all images.

Index	Setting	Rank-1	*m*AP	Inf-Time
1	$K_1 = 50, K_2 = 5$	82.0	73.7	51.09s
2	$K_1 = 500, K_2 = 5$	82.1	75.6	51.61s
3	$K_1 = 200, K_2 = 5$	82.1	75.4	51.42s
4	$K_1 = 200, K_2 = 1$	76.5	68.9	50.65s
5	$K_1 = 200, K_2 = 20$	80.7	73.6	51.57s

CICO alone cannot ensure feature-level consistency. The occlusion consistency loss explicitly pulls together features from the same occluded regions, helping the model focus on visible regions under misleading occlusion patterns. In **Index-3**, removing the foreground alignment loss leads to a 1.2% drop in Rank-1, showing that preserving foreground consistency is also essential under background variation.

Hyperparameters of occlusion diversity and frequency in CICO. As shown in Fig. 4, increasing the number of predefined occlusion types M and the number of occlusion instances per batch N initially improves performance, as more diverse and frequent occlusions enhance the model's robustness. However, performance saturates or slightly drops when $M > 20$ or N becomes too large. This suggests that while moderate occlusion augmentation helps the model learn to handle occlusions, excessive occlusions may lead to overfitting to specific patterns or hinder the learning of complete pedestrian features. Empirically, using two occlusion types and covering two instances per ID (occupying 50% of a batch) achieves the best trade-off.

Table 5. Comparison with other occlusion augmentation methods in CICO module on Occluded-Duke dataset.

Methods	ViT	ViT+SPT	ViT+ADM	ViT+CICO
Rank-1	63.5	**68.6**	66.2	67.4
*m*AP	53.1	57.4	57.7	**57.8**

Table 6. Ablation study on the origin of the prompt heatmap in PBF module on Occluded-Duke dataset.

Index	Setting	Rank-1	*m*AP
1	Human parsing label	80.2	74.6
2	"A photo of a person."	81.3	74.8
3	"[learnable tokens] person."	82.1	75.4

Hyperparameters of top-K number in PSS. As shown in Tab. 4, varying K_1 values in **Index 1** to **Index 3**, Rank-1 accuracy remains stable, but mAP increases with K_1 and reach stabilization ($k_1 > 200$), indicating that most relevant matches are among the top 200 results. Scoring lower-ranked samples has minimal impact on performance. Comparing **Index 3** to **Index 5**, different K_2 values significantly impact performance. A small K_2 relies too much on few high prompt score samples, leading to poor outcomes, while a larger K_2 introduces occluded samples with lower prompt scores, degrading retrieval quality. We find that $K_2 = 5$ offers the best balance. Moreover, the inference time remains largely unchanged across different parameter settings, indicating the stability of our method.

Comparison with occlusion augmentation methods. We compare CICO with the SOTA augmentation methods SPT [25] and ADM [30] on the ViT [3] backbone, as shown in Tab. 5. SPT merges salient and non-salient regions across images to synthesize new samples, while ADM employs attack-based masks to simulate realistic occlusion disturbances. Both methods improve performance by diversifying occluded appearances. In contrast, our approach focuses on mitigating similar occlusions to enhance matching reliability. CICO achieves comparable performance, demonstrating that robust occlusion handling can be attained without relying on data diversity expansion.

Discussion on the origin of prompt heatmaps. Accurate prompt heatmaps are crucial for enhancing data quality in PBF and ensuring reliable prompt scores in PSS. As shown in Tab. 6, we compare two generation strategies:

Table 7. Comparison with re-ranking and FPC on Occluded-Duke dataset. † indicates results reproduced by us.

Index	Setting	Rank-1	*m*AP	Inf-Time
1	COPE(w/o PSS)	76.8	68.9	42.30s
2	1 + PSS(COPE)	82.1	75.4	51.42s
3	1 + re-rank	81.1	81.6	173.75s
4	COPE+re-rank	**84.0**	**82.4**	180.35s
5	FPC	76.7	72.8	95.71s†
6	FPC+re-rank	78.6	78.3	221.38s†

directly using human parsing labels and vision–language guided variants. In **Index-1**, automatically generated parsing masks from BPBReID [20] and KPR [21] are directly used as heatmaps, making the model highly dependent on annotation quality and label availability during inference. In contrast, **Index-2** employs a fixed textual prompt to align image–text features, enabling automatic foreground localization and improving Rank-1 accuracy by 1.1%. Building on this, **Index-3** adopts a CoCoOp-based adaptive prompt, where text embeddings are conditioned on visual features to more precisely focus on pedestrian foregrounds, yielding further gains of +0.7%/+0.8% in Rank-1/mAP.

Comparison with post-processing and feature reconstruction methods. We compare PSS against the widely used re-ranking [37] in Tab. 7. **Index-1** shows results without PSS. **Index-2** and **Index-3** incorporate PSS and re-ranking, respectively. Our PSS method is 121 seconds faster than re-ranking while achieving a Rank-1 improvement of over 1.0%. Furthermore, **Index-4** indicates that PSS and re-ranking are compatible and can further enhance retrieval accuracy. Lastly, we introduce the SOTA feature reconstruction method FPC [32] for comparison. Comparing Index-2 and **Index-5**, our method outperforms FPC by +5.4%/+8.8% in Rank-1/mAP while only taking 54% of the time. Comparing Index-4 and **Index-6**, the COPE+re-ranking outperforms FPC+re-ranking by a large margin, further demonstrating the superiority of our method.

Visualization. Fig. 5(a) shows the results under partial occlusion. The baseline suffers from irrelevant occlusions (e.g., vehicles), leading to feature confusion. With CICO and PBF, interference is reduced and matches align better in the visible regions. Adding PSS further improves ranking by leveraging prompt scores to highlight correct matches. In Fig. 5(b), under heavy occlusion, the baseline fails to retrieve relevant images. CICO and PBF retrieve partial matches, but similar backgrounds still cause confusion. Notably, prompt scores for heavily occluded queries (e.g., 0.73 and 0.82) are lower than those for fully visible ones (e.g., 0.91 and 0.92), indicating diminished feature reliability. Our PSS effectively compensates for this by refining the retrieval process based on high-confidence samples.



Figure 5. Visualization of retrieval rankings for two types of queries under different modules. Green/red borders indicate correct/incorrect matches. $\mathcal{P}$ is the prompt score.



Figure 6. Visualization of attention maps on the test set from the Occluded-Duke dataset in the baseline and our COPE.

In Fig. 6, we further visualize the attention maps of occluded test images. The baseline shows scattered and inaccurate attention, reflecting hallucination [33] in CLIP-ViT, where focus drifts to occluded or irrelevant regions. In contrast, our COPE suppresses attention on occlusions and enhances it on visible parts, leading to better localization and utilization of body parts for the final retrieval.

5. Conclusion

This paper proposes the COPE for occluded person re-identification. To handle strong interference from similar occlusions, the CICO module applies identical occlusions across identities and aligns their features, enhancing robustness under similar occlusion patterns. The PBF module simulates background disturbance and aligns foreground features to improve foreground localization and extraction. Additionally, a lightweight PSS module refines retrieval results by leveraging prompt-guided completeness scores. Extensive experiments show that COPE achieves SOTA performance on both occluded and holistic benchmarks while maintaining high efficiency.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 62276221, 62376232); the Fundamental Research Funds for the Central Universities (No. 21625344); the Postdoctoral Fellowship Program of CPSF (No. GZB20250401); the China Postdoctoral Science Foundation (No. 2025M781471); and the Guangdong Basic and Applied Basic Research Foundation (No. 2026A1515010083).

References

[1] Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuang Liu, Qixiang Ye, Mingliang Xu, Qi'an Chen, and Rongrong Ji. Occlude them all: Occlusion-aware attention network for occluded person re-id. In *ICCV*, pages 11833–11842, 2021. 2, 6

[2] Can Cui, Siteng Huang, Wenxuan Song, Pengxiang Ding, Min Zhang, and Donglin Wang. Profd: Prompt-guided feature disentangling for occluded person re-identification. In *ACM MM*, pages 1583–1592, 2024. 4, 6, 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 7, 1

[4] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *CVPR*, pages 11744–11752, 2020. 5, 6, 4

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[6] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *CVPR*, pages 15013–15022, 2021. 6

[7] Shuting He, Weihua Chen, Kai Wang, Hao Luo, Fan Wang, Wei Jiang, and Henghui Ding. Region generation and assessment network for occluded person re-identification. *IEEE TIFS*, 2023. 3, 4, 6

[8] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Feature completion for occluded person re-identification. *IEEE TPAMI*, pages 4894–4912, 2021. 2

[9] Mengxi Jia, Yifan Sun, Yunpeng Zhai, Xinhua Cheng, Yi Yang, and Ying Li. Semi-attention partition for occluded person re-identification. In *AAAI*, pages 998–1006, 2023. 1

[10] Jiachen Li and Xiaojin Gong. Prototypical contrastive learning-based clip fine-tuning for object re-identification. *arXiv preprint arXiv:2310.17218*, 2023. 3, 5, 6

[11] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *ICLR*, 2020. 3

[12] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *AAAI*, pages 1405–1413, 2023. 1, 2, 5, 6, 4

[13] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *CVPR*, pages 2898–2907, 2021. 1, 6

[14] Guorong Lin, Shunzhi Yang, Wei-Shi Zheng, Zuoyong Li, and Zhenhua Huang. A semantically guided and focused network for occluded person re-identification. *IEEE TIFS*, 20:9716–9731, 2025. 6

[15] Wenfeng Liu, Xudong Wang, Lei Tan, Yan Zhang, Pingyang Dai, Yongjian We, and Rongrong Ji. Learning occlusion disentanglement with fine-grained localization for occluded person re-identification. In *ACM MM*, pages 6462–6471, 2023. 6

[16] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, pages 1487–1495, 2019. 5

[17] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, pages 542–551, 2019. 5, 3

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3, 1

[19] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35, 2016. 5, 3

[20] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person re-identification. In *WACV*, pages 1613–1623, 2023. 4, 6, 8

[21] Vladimir Somers, Alexandre Alahi, and Christophe De Vleeschouwer. Keypoint promptable re-identification. In *ECCV*, pages 216–233, 2024. 2, 6, 8

[22] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline). In *ECCV*, pages 501–518, 2018. 6

[23] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 5, 4

[24] Lei Tan, Pingyang Dai, Rongrong Ji, and Yongjian Wu. Dynamic prototype mask for occluded person re-identification. In *ACM MM*, pages 531–540, 2022. 1, 5, 2, 3, 4

[25] Lei Tan, Jiaer Xia, Pingyang Dai, Mingbo Zhao, Yongjian Wu, and Liujuan Cao. Occluded person re-identification via saliency-guided patch transfer. In *AAAI*, pages 5070–5078, 2024. 2, 7

[26] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *CVPR*, pages 6449–6458, 2020. 6

[27] Tao Wang, Mengyuan Liu, Hong Liu, Wenhao Li, Miaoju Ban, Tianyu Guo, and Yidi Li. Feature completion trans-

former for occluded person re-identification. *IEEE TMM*, 2024. 2

[28] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. Feature erasing and diffusion network for occluded person re-identification. In *CVPR*, pages 4754–4763, 2022. 6

[29] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 5, 3

[30] Jiaer Xia, Lei Tan, Pingyang Dai, Mingbo Zhao, Yongjian Wu, and Liujuan Cao. Attention disturbance and dual-path constraint network for occluded person re-identification. In *AAAI*, pages 6198–6206, 2024. 6, 7

[31] Boqiang Xu, Lingxiao He, Jian Liang, and Zhenan Sun. Learning feature recovery transformer for occluded person re-identification. *IEEE TIP*, pages 4651–4662, 2022. 2

[32] YuTeng Ye, Hang Zhou, Jiale Cai, Chenxing Gao, Youjia Zhang, Junle Wang, Qiang Hu, Junqing Yu, and Wei Yang. Dynamic feature pruning and consolidation for occluded person re-identification. In *AAAI*, pages 6684–6692, 2024. 2, 6, 8, 3

[33] Kejia Zhang, Keda Tao, Jiasheng Tang, and Huan Wang. Poison as cure: Visual noise for mitigating object hallucinations in lvms. *arXiv preprint arXiv:2501.19164*, 2025. 8

[34] Cairong Zhao, Zefan Qu, Xinyang Jiang, Yuanpeng Tu, and Xiang Bai. Content-adaptive auto-occlusion network for occluded person re-identification. *IEEE TIP*, pages 4223–4236, 2023. 1

[35] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 5, 3

[36] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *iccv*, pages 4678–4686, 2015. 5, 4

[37] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 1318–1327, 2017. 8

[38] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020. 1

[39] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 3, 4

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, pages 2337–2348, 2022. 3

[41] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *ICME*, pages 1–6, 2018. 5, 3