Categorical Attention: Fine-grained Language-guided Noise Filtering Network for Occluded Person Re-Identification

Minghui Chen[1,2], **Dayan Wu**[1*], **Chenxu Yang**[1,2], **Qinghang Su**[1,2], **Zheng Lin**[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{chenminghui, wudayan, yangchenxu, suqinghang, linzheng}@iie.ac.cn

Abstract

Person Re-Identification (ReID) aims to match individuals across different camera views, but occlusions in real-world scenarios, such as vehicles or crowds, hinder feature extraction and matching. Current occluded ReID methodologies typically leverage visual augmentation techniques in an attempt to mitigate the disruptive effects of occlusion-induced noise. However, relying solely on visual data fail to effectively filter out occlusion noise. In this paper, we introduce the Fine-grained Language-guided Noise Filtering Network (FLaN-Net) for occluded ReID. FLaN-Net innovatively employs categorical attention mechanism to generate adaptive tokens that capture the following three distinct types of visual information: comprehensive descriptions of individuals, detailed visible attributes, and characteristics of occluding objects. Subsequently, a cross-attention mechanism aligns these prompts with the image, guiding the model to focus on relevant regions. To generate robust and discriminative features for occluded pedestrians, we further introduce a dynamic weighting fusion module that integrates visual, textual, and cross-attention features based on their reliability. Experimental results demonstrate that FLaN-Net outperforms existing methods on occluded ReID benchmarks, offering a robust solution for challenging real-world conditions.

1 Introduction

Person Re-Identification (ReID) aims to identify and match the same target individual across different and non-overlapping camera views [Ye *et al.*, 2021]. However, people and objects often move randomly, and surveillance devices typically cover wide areas in the real-world scenario, which leads to a high likelihood of individuals being partially occluded. This occlusion creates a major challenge for person re-identification, as it introduces significant noise during feature extraction and feature matching. To cope with occluded ReID [Zheng *et al.*, 2015b; Zhuo *et al.*, 2018; Hou *et al.*, 2021], various strategies have been proposed to mitigate the effects of noisy information caused by occlusion. Common approaches, such as auxiliary models [Hou *et al.*, 2021; Wang *et al.*, 2022a; Dou *et al.*, 2023] and attention mechanisms [He *et al.*, 2021; Tan *et al.*, 2022; Jia *et al.*, 2023] help the model distinguish key information from occlusion-induced noise. PFD [Wang *et al.*, 2022a] is a notable method that integrates the auxiliary model and attention mechanism, employing pose-guided feature disentangling to reduce occlusion noise by associating features with human body parts. While auxiliary models provide external information and attention mechanisms adaptively focus on unobstructed regions, their reliance solely on visual data often fails to completely filter out occlusion noise. This leaves occlusions still being misinterpreted as identity-relevant features, yielding less discriminative representations. As shown in Fig. 1, the attention maps reveal the varying performance of different methods under diverse occlusion scenarios. It can

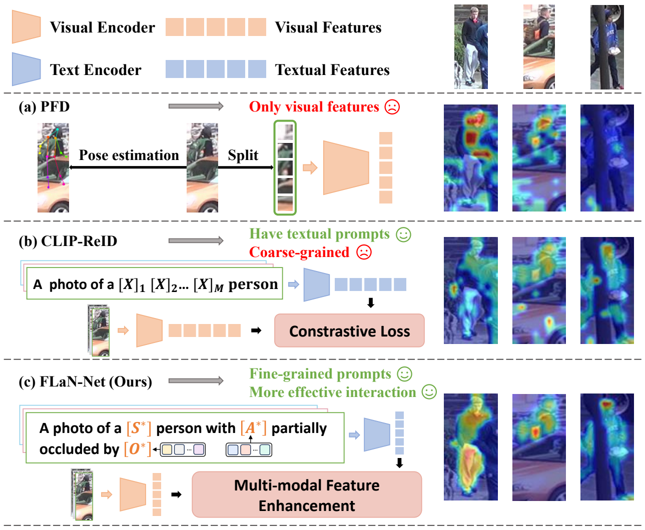

Figure 1: Comparison of baselines and our fine-grained language-guided noise filtering network. (a) PFD, (b) CLIP-ReID, (c) Our proposed FLaN-Net method, which incorporates fine-grained textual descriptions and a more effective multi-modal interaction module, enables the model to achieve more robust noise suppression.

[*] Corresponding Author

be observed from Fig. 1(a) that the attention maps generated by PFD highlight many irrelevant regions.

Inspired by the potential of vision-language models, recent works have explored the integration of textual descriptions with visual features to enhance feature extraction. As a representative work, CLIP-ReID [Li *et al.*, 2023] deploys the vision-language model CLIP [Radford *et al.*, 2021] to enhance the learning of visual features by training a set of tokens for each pedestrian ID. However, existing language-guided ReID methods are not well-suited for the occluded ReID task, as their prompts tend to capture global information from images, failing to provide the necessary contextual information in occluded scenarios. Furthermore, textual information in these methods is not fully utilized, as it is primarily employed for contrastive loss with images rather than explicitly guiding feature extraction. As shown in Fig. 1(b), although CLIP-ReID helps reduce noise in occlusion scenarios, its attention maps still frequently emphasize occluding objects and irrelevant background regions, retaining significant noise in the extracted features.

To address these limitations, we propose the **F**ine-grained **La**nguage-guided **N**oise Filtering **Net**work (**FLaN-Net**) for Occluded Person Re-Identification. FLaN-Net innovatively employs a categorical attention mechanism to generate adaptive tokens that capture three distinct types of visual information: comprehensive descriptions of individuals, detailed visible attributes, and characteristics of occluding objects. Specifically, our method transforms image information into subject and detail tokens, and then represents each image with a occlusion-aware fine-grained textual prompt in the format:"A photo of a $[S^*]$ person with $[A^*]$ partially occluded by $[O^*]$." Here, S^* denotes the primary subject token, which encapsulates the essential identity of the pedestrian. The attribute token A^* and the context token O^* serve as detail tokens, capturing identity-relevant attributes and occluding objects respectively. FLaN-Net learns detailed information for each pedestrian through the guidance of multiple learnable queries. Moreover, to ensure effective utilization of these prompts, FLaN-Net incorporates a cross-attention mechanism that dynamically aligns textual tokens with image patch tokens. This alignment enables the image encoder to focus on semantically relevant and visible regions of the pedestrian, filtering out noise caused by occlusions. Finally, FLaN-Net combines the features from the image encoder, text encoder, and cross-attention in a dynamic weighting fusion module to generate a robust feature representation for each occluded pedestrian. The dynamic fusion module assigns weights to each feature based on its uncertainty, highlighting the influence of more reliable features. As shown in Fig. 1(c), our method effectively highlights the visible regions of the pedestrian while filtering out noise from occluding objects, demonstrating improved focus and robustness in occluded scenarios. We summarize our contributions as follows:

- We introduce FLaN-Net, a novel method that employs a fine-grained language-guided mechanism to construct adaptive occlusion-aware prompts for occluded ReID. This technique enables meticulously detailed descriptions of the visible aspects of an individual, while effectively filtering out noise caused by occlusions.
- We propose an advanced multi-modal feature enhancement paradigm that combines a cross-attention mechanism with a dynamic weighting fusion module. This integration is designed to produce robust and discriminative representations of occluded pedestrians.
- Experimental results demonstrate that FLaN-Net significantly improves retrieval performance on both occluded and holistic ReID benchmarks, outperforming state-of-the-art methods.

2 Related Work

2.1 Occluded Person Re-identification

Occluded Person Re-Identification (Occluded ReID) presents significant challenges due to the noise introduced by occlusions, which hinder the model's ability to extract and match features accurately. One common approach to mitigate this issue is the incorporation of auxiliary information, such as pose estimation [Hou *et al.*, 2021; Wang *et al.*, 2022a] and human parsing models [Gao *et al.*, 2020a; Dou *et al.*, 2023]. Another widely used approach involves attention mechanisms [He *et al.*, 2021; Tan *et al.*, 2022; Jia *et al.*, 2023; Li *et al.*, 2024], which enhance robustness by adaptively focusing on relevant regions of an image. To further aid the attention learning process, various data augmentation strategies, such as random erasing [Wang *et al.*, 2022b] and artificially generated occlusions [Chen *et al.*, 2021; Xia *et al.*, 2024; Tan *et al.*, 2024], have been incorporated to help the model better handle occlusions across diverse scenarios. However, these methods rely only on visual features, neglecting the potential of text information to help filter out noise. In recent years, language-guided methods [Li *et al.*, 2023; Yang *et al.*, 2024] have been explored to address the challenges of ReID. CLIP-ReID [Li *et al.*, 2023] is the pioneering work that uses the CLIP [Radford *et al.*, 2021] model to integrate textual prompts with visual features.

2.2 Image-to-Word Mapping

In the field of text-to-image generation, [Gal *et al.*, 2022] was the first to use novel pseudo-words in the word embedding space to represent an object or a style. Recently, this technique, known as textual inversion, has been widely applied to zero-shot compositional image retrieval tasks [Saito *et al.*, 2023; Suo *et al.*, 2024]. These methods map a reference image to a pseudo-token in the CLIP embedding space, which is then combined with a descriptive query to facilitate text-to-image retrieval. However, a limitation of these models is that they map the entire image to a single pseudo-token, which can introduce noise and overlook important details relevant to the retrieval task. Recent studies have proposed using learnable queries to capture fine-grained features within an image. For example, in instance segmentation task, learnable queries are used to explicitly represent an object's class, location, and mask [Dong *et al.*, 2021]. In object detection task, learnable queries are used to capture object relationships and global context for parallel predictions [Carion *et al.*, 2020]. Inspired by these works, this paper proposes incorporating learnable queries into occluded ReID tasks to perform fine-grained textual inversion.

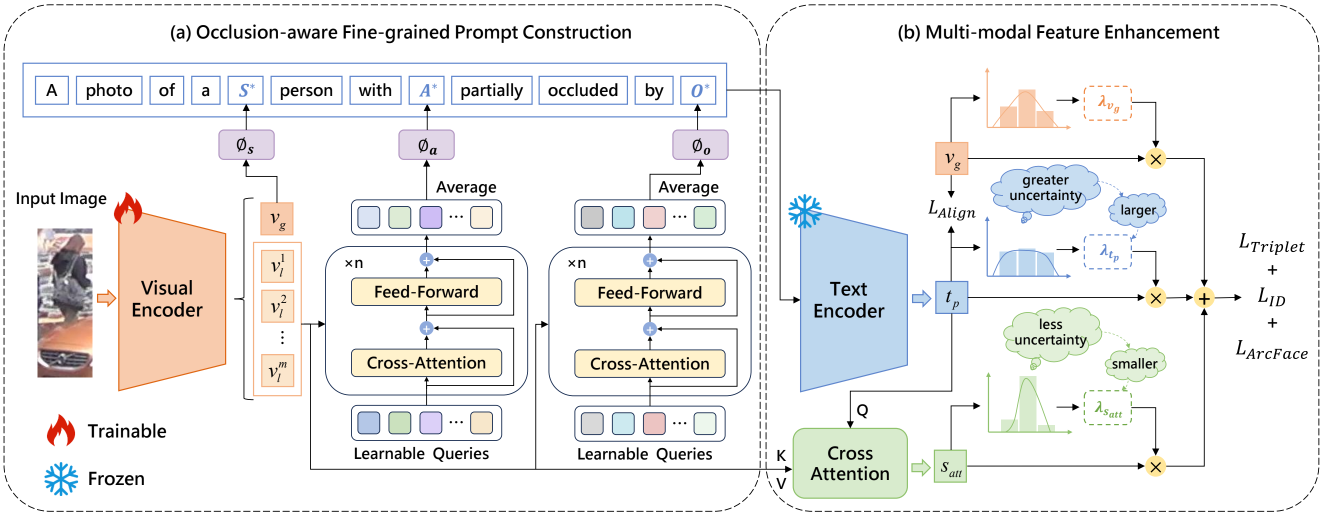

Figure 2: The framework of FLaN-Net: (a) Occlusion-aware Fine-grained Prompt Construction, which generates subject and detail tokens to capture identity-specific features; (b) Multi-modal Feature Enhancement, which aligns textual prompts with visual regions via cross-attention mechanism and combines visual, textual, and cross-attention features through a dynamic weighting fusion module.

3 Methodology

We introduce the Fine-grained Language-guided Noise Filtering Network (FLaN-Net) for Occluded Person Re-Identification, which consists of two key components, as shown in Fig. 2. The first component, **Occlusion-aware Fine-grained Prompt Construction**, generates adaptive tokens that capture three types of visual information: comprehensive descriptions of individuals, detailed visible attributes, and characteristics of occluding objects. These tokens are used to generate fine-grained textual descriptions, which provide a comprehensive representation of the pedestrian. The second component, **Multi-modal Feature Enhancement**, integrates visual and textual information to improve feature extraction. This is achieved through a cross-attention mechanism that aligns textual prompts with specific image regions, and a dynamic weighting fusion module, which adaptively combines the features to ensure that the most relevant and reliable features are given larger weight.

3.1 Occlusion-aware Fine-grained Prompt Construction

To effectively represent occluded images, our method projects the image into pseudo-word tokens, including a subject-focused token S^* for the individual's essential identity and detail-focused tokens A^* and O^* for visible attributes and occluding objects. These tokens are then combined to construct a personalized prompt:"A photo of a $[S^*]$ person with $[A^*]$ partially occluded by $[O^*]$." This occlusion-aware prompt enhances the model's ability to distinguish individuals in occluded ReID tasks.

Subject-focused Token Generation

In order to get the subject-focused token, we leverage the image encoder f_v of the pre-trained CLIP model. Specifically, given an image I, the visual encoder extracts visual feature $v = f_v(I) = \{v_g, v_l^1, \ldots, v_l^n\}$. Here $v_g \in \mathbb{R}^{d\times 1}$ denotes the global visual feature and $\{v_l^i\}_{i=1}^n \in \mathbb{R}^{d\times n}$ represent the local patch features, where d is the feature dimension and n is the number of patches. Then we apply a simple mapping network ϕ_s to transform the global image feature v_g into a subject-focused pseudo-word token. Formally, we define:

$$S^* = \phi_s(v_g), \tag{1}$$

where ϕ_s is a three-layered fully-connected network and S^* serves as a comprehensive description of the main subject.

Detail-focused Token Generation

The detail tokens are divided into two types: attribute tokens A^*, which describe identity-relevant visible attributes, and context tokens O^*, which represent characteristics about the occluding object. Both A^* and O^* are generated using the same network architecture, leveraging learnable queries to extract relevant features from the input image. To illustrate this process, we take the generation of A^* as an example and provide a detailed explanation of its network structure.

Let $V_l = \{v_l^i\}_{i=1}^n \in \mathbb{R}^{d\times n}$ represent the local patch features. Next, these patch features are fed into a fine-grained noise filtering network. This network interacts with a set of m learnable queries $X = \{x_i\}_{i=1}^m \in \mathbb{R}^{d\times m}$ through cross-attention, allowing these queries to capture attribute information from the corresponding semantic regions in the image. For A^*, this process enables these queries to focus on specific visual details, capturing exposed attribute features such as clothing colors, accessory types, and hairstyles. Similarly, for O^*, these queries could extract information about the occluding objects. Specifically, we compute the query Q_d, key K_d, and value V_d matrices as follows:

$$Q_d = XW_Q, \quad K_d = [X, V_l]\,W_K, \quad V_d = [X, V_l]\,W_V, \tag{2}$$

where W_Q, W_K, W_V are different linear transformations and $[X, V_l]$ denotes the concatenation of the learnable queries and local patch features, facilitating their interaction. The cross-attention output H^i for the learnable queries in the i-th attention block is then computed as:

$$H^i = \mathrm{CrossAttn}(Q_d, K_d, V_d) = \mathrm{softmax}\left(\frac{Q_d K_d^\top}{\sqrt{d}}\right) V_d. \tag{3}$$

Subsequently, we feed H^i into a two-layer feed-forward network $\mathrm{FFN}(\cdot)$, producing $\tilde{X}^i$, which represents the updated features of the learnable queries after the i-th attention block. This process is formulated as:

$$\tilde{X}^i = \mathrm{FFN}\left(H^i + \tilde{X}^{i-1}\right) + H^i. \tag{4}$$

Afterwards, we perform average pooling over the refined query embeddings after multiple transformer blocks. The result is then passed through a simple mapping network ϕ_a to derive the final attribute token A^*. It can be formulated as:

$$A^* = \phi_a(\mathrm{AvgPool}(\tilde{X}_{final})), \tag{5}$$

where $\tilde{X}_{final}$ is the final output query embeddings from multiple transformer blocks, $\mathrm{AvgPool}(\cdot)$ denotes average pooing, and ϕ_a denotes a three-layer feed-forward network.

The above network produces the refined attribute token A^*. Similarly, a parallel network with the same structure but different parameters is employed to generate the corresponding context token O^*. In this network, ϕ_o serves as the associated mapping function.

3.2 Multi-modal Feature Enhancement

To effectively capture the intricate relationships between multi-modal information, we integrate a cross-attention mechanism with a dynamic fusion strategy, enhancing the model's capacity to represent nuanced identity features.

Cross-Attention Mechanism

The cross-attention mechanism allows the model to interactively align specific textual cues with corresponding visual regions, focusing on subject and detail tokens in the constructed prompt. Let T represent the constructed prompt "A photo of a $[S^*]$ person with $[A^*]$ partially occluded by $[O^*]$", which is then fed into the frozen CLIP text encoder f_t to obtain the textual representation t_p. Formally, this process can be expressed as follows:

$$t_p = f_t(T). \tag{6}$$

To implement cross-attention, we treat the textual embedding t_p as the query Q_c, while the visual feature v extracted from the image serves as both the key K_c and value V_c. The output of the cross-attention mechanism can be expressed as:

$$s_{att} = \mathrm{CrossAttn}(Q_c, K_c, V_c) = \mathrm{softmax}\left(\frac{Q_c K_c^\top}{\sqrt{d}}\right) V_c. \tag{7}$$

By leveraging the cross-attention mechanism, the textual tokens dynamically guide the image encoder to focus on the visible and discriminative features of the pedestrian while suppressing noise from occluded regions.

Dynamic Weighting Fusion Module

To further enhance feature representation, we introduce a dynamic weighting fusion module that assigns adaptive weights to each feature based on prediction uncertainty, allowing the model to prioritize more reliable features. Specifically, the global visual feature v_g, the textual feature t_p, and the cross-attention feature s_{att} are fed into the fusion module. For each feature f_m ($m = 1, \dots, M$, where $M = 3$, representing the visual, textual, and cross-attention features), the associated weight λ_m is determined according to the uncertainty of feature f_m, which is quantified by the entropy of its prediction distribution. The uncertainty E_m is computed as follows:

$$E_m = -\sum_{k=1}^{C} p_m(k) \log p_m(k), \tag{8}$$

where $p_m(k)$ is the softmax probability assigned to the k-th individual's feature f_m, and C represents the total number of individuals in the dataset. A lower E_m reflects less uncertainty in the prediction, resulting in an increased weight for that feature during the fusion process. The weight λ_m for feature f_m is then determined as follows:

$$\lambda_m = \frac{\exp\left(\mathop{Max}\limits_{m=1,\dots,M}(E_m) - E_m\right)}{\sum_{q=1}^{M} \exp\left(\mathop{Max}\limits_{m=1,\dots,M}(E_m) - E_q\right)}. \tag{9}$$

The final fused representation $\hat{f}$ is formulated as:

$$\hat{f} = \sum_{m=1}^{M} \lambda_m f_m, \tag{10}$$

where f_m represents each prediction feature. This adaptive fusion approach enables balanced contributions from the visual, textual, and cross-attention features.

3.3 Loss Function and Inference

Our framework incorporates four loss functions: Cross-modal Contrastive Loss $\mathcal{L}_{\text{Align}}$ [Radford *et al.*, 2021], Triplet Loss $\mathcal{L}_{\text{Triplet}}$ [Hermans *et al.*, 2017], ID Classification Loss $\mathcal{L}_{\text{ID}}$ [Zheng *et al.*, 2017], and ArcFace Loss $\mathcal{L}_{\text{ArcFace}}$ [Deng *et al.*, 2019].

Cross-modal Contrastive Loss

To align visual and textual representations of each individual, we employ a cross-modal contrastive loss that encourages high similarity between images and their corresponding prompts in the embedding space. It is formulated as:

$$\mathcal{L}_{\text{Align}} = \mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}, \tag{11}$$

$$\mathcal{L}_{\text{i2t}}(i) = -\sum_{p^+ \in P(n)} \log \frac{\exp\left(\mathrm{sim}\left(v_i, t_{p+}\right)/\tau\right)}{\sum_{n=1}^{N} \exp\left(\mathrm{sim}\left(v_i, t_n\right)/\tau\right)}, \tag{12}$$

$$\mathcal{L}_{\text{t2i}}(i) = -\sum_{p^+ \in P(n)} \log \frac{\exp\left(\mathrm{sim}\left(t_i, v_{p+}\right)/\tau\right)}{\sum_{n=1}^{N} \exp\left(\mathrm{sim}\left(t_i, v_n\right)/\tau\right)}, \tag{13}$$

where $P(n)$ denotes the set of positive samples that correspond to the same identity i and τ is a temperature parameter.

Triplet Loss

The triplet loss is employed to increase the distinction between identities by minimizing the distance between positive pairs while maximizing the distance between negative pairs. This loss is formulated as:

$$\mathcal{L}_{\text{Triplet}} = \max\left(d_p - d_n + m, 0\right), \tag{14}$$

where d_p and d_n are the distances of the positive and negative pairs respectively and m is the margin.

ID Classification Loss

The ID classification loss ensures correct identification of each individual and is defined as:

$$\mathcal{L}_{\text{ID}} = -\sum_{k=1}^{N} q_k \log\left(y_k\right), \tag{15}$$

where y_k is the predicted probability of an individual, and q_k is the corresponding ground truth label.

ArcFace Loss

ArcFace loss improves feature discrimination by introducing an angular margin to enhance the separation between different individuals and compact intra-individual representations:

$$\mathcal{L}_{\text{ArcFace}} = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{e^{s\left(\cos\left(\theta_{y_k}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_k}+m\right)\right)}+\sum_{j=1,j\neq y_k}^{n} e^{s\cos\theta_j}}, \tag{16}$$

where θ_j denotes the angle between a feature and the weight vector of the j-th identity. The angular margin m improves discrimination, and the scale factor s stabilizes optimization.

Total Loss

Overall, the loss function used in FLaN-Net is defined as:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{Align}} + \mathcal{L}_{\text{Triplet}} + \mathcal{L}_{\text{ID}} + \mathcal{L}_{\text{ArcFace}}, \tag{17}$$

where λ is a hyper-parameter that balances the contribution of $\mathcal{L}_{\text{Align}}$ to the total loss.

Inference

During inference, following CLIP-ReID [Li *et al.*, 2023], we rely solely on features extracted from the image encoder for person retrieval. The Euclidean distance is computed between the feature of the query image and those in the gallery set to identify the closest matches. The proposed components work together to enhance the model's ability to accurately identify the target subject, ultimately optimizing the feature representation generated by the image encoder.

4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Protocols

We evaluate the proposed FLaN-Net method on two categories of datasets: occluded datasets, including Occluded-Duke [Miao *et al.*, 2019] and Occluded-REID [Zhuo *et al.*, 2018], and holistic datasets, comprising Market-1501 [Zheng *et al.*, 2015a], DukeMTMC-reID [Zheng *et al.*, 2017] and CUHK03-NP [Li *et al.*, 2014]. As the Occluded-REID dataset lacks a dedicated training set, we utilize Market-1501 for training, consistent with other methods to maintain a fair basis for comparison. To assess the effectiveness of our approach, we adopt Cumulative Matching Characteristic (CMC) curves and the mean Average Precision (mAP).

Methods	**Occ-Duke**		**Occ-REID**	
	R-1	**mAP**	**R-1**	**mAP**
PVPM (CVPR 20)	47.0	37.7	66.8	59.5
HOReID (CVPR 20)	55.1	43.8	80.3	70.2
RFCnet (TPAMI 21)	63.9	54.5	-	-
HCGA (TIP 23)	70.2	57.5	-	-
PAT (CVPR 21)	64.5	53.6	81.6	72.1
TransReID (ICCV 21)	66.4	59.2	-	-
DRL-Net (TMM 22)	65.8	53.9	-	-
PFD (AAAI 22)	69.5	61.8	81.5	83.0
DPM (ACM MM 22)	71.4	61.8	85.5	79.7
SAP (AAAI 23)	70.0	62.2	83.0	76.8
OAT (TIP 24)	71.8	62.2	82.6	78.2
OAMN (ICCV 21)	62.6	46.1	-	-
FED (CVPR 22)	68.1	56.4	86.3	79.3
CAAO (TIP 23)	68.5	59.5	87.1	83.4
ADP (AAAI 24)	74.5	63.8	89.2	85.1
DPM-SPT (AAAI 24)	74.7	63.0	87.8	81.1
CLIP-ReID (AAAI 23)	67.1	59.5	-	-
FLaN-Net (Ours)	**75.2**	**65.5**	**92.6**	**89.5**

Table 1: Performance comparison on Occluded-Duke and Occluded-REID datasets. The compared methods are grouped into four categories: auxiliary model-based, transformer-based, data augmentation and language-guided.

Implementation Details

In this work, we employ the ViT-B/16 pretrained on CLIP as the visual encoder, and the pre-trained CLIP text transformer as the text encoder. We use two independent fine-grained noise filtering networks to get A^* and O^*, each consisting of 3 learnable queries and 6 cross-attention blocks. The model is trained using a batch size of 64, consisting of 16 identities, each with 4 images. All input images are resized to 256 × 128 pixels. For optimization, we use the Adam optimizer with a base learning rate of 5e-5 for the randomly initialized modules and 1e-5 for the visual encoder. The model is trained for 60 epochs, with the learning rate decaying by a factor of 0.1 at epochs 20 and 40. The Triplet Loss uses the margin m = 0.3, and the ArcFace Loss is configured with the margin m = 0.5 and the scale factor s = 30. The optimizer for the Arcface Loss function is separately initialized with SGD, using a learning rate of 0.1 and a weight decay of 5e-4. The λ in Eq.17 is set to 0.5 for all datasets. All components are trained on a single NVIDIA RTX3090 GPU.

4.2 Comparison with State-of-the-Art Methods

Experimental Results on Occluded ReID Datasets

To evaluate the effectiveness of our proposed FLaN-Net, we conducted extensive comparisons with various state-of-the-art methods on the occluded ReID datasets, including Occluded-Duke and Occluded-REID, and show the results in Tab. 1. SOTA methods are divided into four mainstreams:

- Auxiliary model-based methods: PVPM [Gao *et al.*, 2020b]; HOReID [Wang *et al.*, 2020]; RFCnet [Hou *et al.*, 2021]; HCGA [Dou *et al.*, 2023].
- Transformer-based methods: PAT [Li *et al.*, 2021]; TransReID [He *et al.*, 2021]; DRL-Net [Jia *et al.*, 2022];

Methods	**Market-1501**		**DukeMTMC**	
	R-1	**mAP**	**R-1**	**mAP**
SAN (AAAI 20)	96.1	88.0	87.9	75.5
TransReID (ICCV 21)	95.2	88.9	90.7	82.0
HAT (ACM MM 21)	95.8	89.8	90.4	81.4
DCAL (CVPR 22)	94.7	87.5	89.0	80.1
AAformer (TNNLS 23)	95.4	88.0	90.1	80.9
PHA (CVPR 23)	**96.1**	90.2	-	-
CLIP-ReID (AAAI 23)	95.5	89.6	90.0	82.5
RFCnet (TPAMI 21)	95.2	89.2	90.7	80.7
PFD (AAAI 22)	95.5	89.7	91.2	83.2
FED (CVPR 22)	95.0	86.3	89.4	78.0
DPM (ACM MM 22)	95.5	89.7	91.0	82.6
CAAO (TIP 23)	95.3	88.0	89.8	80.9
HCGA (TIP 23)	95.2	88.4	-	-
SAP (AAAI 23)	96.0	**90.5**	-	-
ADP (AAAI 24)	95.6	89.5	91.2	83.1
DPM-SPT (AAAI 24)	95.5	89.4	91.1	82.4
OAT (TIP 24)	95.7	89.9	91.2	82.3
FLaN-Net (Ours)	95.7	**90.5**	**92.1**	**83.6**

Table 2: Performance comparison on Market-1501 and DukeMTMC-reID. The compared methods are grouped into two categories: holistic methods and occluded methods.

Methods	**Labeled**		**Detected**	
	R-1	**mAP**	**R-1**	**mAP**
RGA-SC (CVPR 20)	81.1	77.4	79.6	74.5
HAT (ACM MM 21)	82.6	80.0	79.1	75.5
NetVLAD-M (TIFS 22)	80.4	76.7	79.7	74.8
MPN (TPAMI 22)	85.0	81.1	83.4	79.1
AAformer (TNNLS 23)	80.3	79.0	78.1	77.2
PHA (CVPR 23)	84.5	83.0	83.2	80.3
OAT (TIP 24)	83.9	81.5	80.6	78.0
FLaN-Net (Ours)	**88.1**	**86.7**	**87.1**	**84.8**

Table 3: Performance comparison on CUHK03-NP.

Index	**Prompts**	**R-1**	**mAP**
1	”A photo of a person”	72.6	63.7
2	”A photo of a $[S^*]$ person”	73.4	64.4
3	”A photo of a $[S^*]$ person with $[A^*]$”	74.6	65.2
4	”A photo of a $[S^*]$ person with $[A^*]$ partially occluded by $[O^*]$”	75.2	65.5

Table 4: Ablation study of different prompts on Occluded-Duke.

PFD [Wang *et al.*, 2022a]; DPM [Tan *et al.*, 2022]; SAP [Jia *et al.*, 2023]; OAT [Li *et al.*, 2024].

- Data augmentation methods: OAMN [Chen *et al.*, 2021]; FED [Wang *et al.*, 2022b]; CAAO [Zhao *et al.*, 2023]; ADP [Xia *et al.*, 2024]; DPM-SPT [Tan *et al.*, 2024].
- Language-guided methods: CLIP-ReID [Li *et al.*, 2023].

The experimental results demonstrate that FLaN-Net achieves outstanding performance on both occluded ReID datasets. For the Occluded-Duke dataset, FLaN-Net achieves a Rank-1 accuracy of 75.2% and an mAP of 65.5%, outperforming the second best method, DPM-SPT, by +0.5% and +2.5%, respectively. On the Occluded-REID dataset, FLaN-Net achieves the best performance with at least +3.4% Rank-1 accuracy and +4.4% mAP compared to other methods. These results indicate that FLaN-Net effectively addresses the noise caused by occlusions.

Experimental Results on Holistic ReID Datasets

We also experiment our proposed method on holistic person ReID datasets, including Market-1501, DukeMTMC-reID and CUHK03-NP. Tab. 2 shows the results on Market-1501 and DukeMTMC-reID datasets. We compare FLaN-Net with two categories of methods:

- Holistic ReID methods: SAN [Jin *et al.*, 2020]; TransReID; HAT [Zhang *et al.*, 2021]; DCAL [Zhu *et al.*, 2022]; AAformer [Zhu *et al.*, 2023]; PHA [Zhang *et al.*, 2023]; CLIP-ReID.
- Occluded ReID methods: RFCnet; PFD; FED; DPM; CAAO; HCGA; SAP; ADP; DPM-SPT; OAT.

We observe that our FLaN-Net achieves competitive results on the Market1501 dataset and achieves SOTA performance on the DukeMTMC-reID dataset. Compared with language-guided method CLIP-ReID, our method surpasses it by +0.2%/+0.9% Rank-1 accruacy/mAP on Market-1501 and +2.1%/+1.1% Rank-1 accruacy/mAP on DukeMTMC.

Additionally, we evaluate FLaN-Net's performance on the CUHK03-NP dataset under both manually labeled and auto-detected bounding box settings. Several methods are compared, including RGA-SC [Zhang *et al.*, 2020], HAT, NetVLAD-M [Zhang *et al.*, 2022], MPN [Ding *et al.*, 2022], AAformer, PHA, OAT. As shown in Tab. 3, FLaN-Net still surpasses all other methods with significant margins. Specifically, it achieves at least +3.1%/+3.7% improvements in Rank-1 accuracy/mAP on the labeled setting and at least +3.7%/+4.5% on the detected setting. Though FLaN-Net is not explicitly designed for holistic ReID tasks, it still guarantees a comparable performance with most of holistic methods, underscoring its robustness and generalization capabilities.

4.3 Ablation Study

Ablation Study on Prompt Variations

We design a series of prompts to evaluate their impact on model performance, as shown in Tab. 4. From index 1 to 2, the addition of subject-specific S^* improves Rank-1 accuracy from 72.6% to 73.4% and mAP from 63.7% to 64.4%, demonstrating the benefit of incorporating identity-specific information. Expanding the prompts with attribute details A^* (index 3) further enhanced Rank-1 accuracy to 74.6% and mAP to 65.2%, underscoring the importance of fine-grained identity attributes. Finally, introducing occlusion details O^* (index 4) yields the highest performance, with Rank-1 reaching 75.2% and mAP 65.5%. These results validate the effectiveness of occlusion-aware fine-grained prompts in guiding feature extraction under occluded scenarios.

Ablation Study on Model Components

In Tab. 5, we evaluate the contribution of occlusion-aware fine-grained prompt ($\mathcal{F}$), cross-attention mechanism ($\mathcal{C}$), and dynamic weighting fusion module ($\mathcal{D}$) on Occluded-Duke. From index 1 to 2, the performance improves by +2.9% in rank-1 accuracy and +1.8% in mAP, demonstrating the importance of utilizing fine-grained descriptions to capture

Index	$\mathcal{F}$	$\mathcal{C}$	$\mathcal{D}$	R-1	R-5	R-10	mAP
1	-	-	-	70.0	83.7	88.0	61.6
2	✓	-	-	72.9	85.0	88.8	63.4
3	-	✓	-	72.1	84.3	88.4	62.7
4	✓	✓	-	74.3	85.7	89.4	64.9
5	✓	✓	✓	75.2	86.3	89.8	65.5

Table 5: Ablation study of occlusion-aware fine-grained prompt ($\mathcal{F}$), cross-attention mechanism ($\mathcal{C}$), and dynamic weighting fusion module ($\mathcal{D}$) on Occluded-Duke.

	Occluded-Duke		DukeMTMC-reID	
	R-1	mAP	R-1	mAP
Average Weighting	74.6	64.9	91.8	82.8
Summation	74.3	65.3	91.7	83.1
Concatenation	73.2	64.2	91.9	83.0
Dynamic Fusion	**75.2**	**65.5**	**92.1**	**83.6**

Table 6: Comparison of different fusion methods.

identity-relevant features. From index 1 to 3, cross-attention also shows its effectiveness. For index 4, combining $\mathcal{F}$ and $\mathcal{C}$ further boosts performance by +4.3% in rank-1 accuracy and +3.3% in mAP, highlighting the necessity of utilizing fine-grained prompts through cross-attention. Notably, Using fine-grained prompts alone (index 2) or applying cross-attention with simple prompts (index 3) proves insufficient for optimal performance. Finally, from index 4 to 5, the addition of $\mathcal{D}$ results in an additional improvement of +0.9% in rank-1 accuracy and +0.6% in mAP, demonstrating the effectiveness of dynamically fusing features.

4.4 Model Analysis

Explore the Optimal Number of Learnable Queries

We investigate the impact of varying the number of learnable queries on model performance, as illustrated in Fig. 3. The results indicate that using 3 learnable queries yields the best performance, with both mAP and Rank-1 reaching their highest values. The performance initially drops as the number of learnable queries increases from 3 to 5. However, the performance begins to improve again as the number rises from 5 to 7 queries. Despite this, the performance gain remains modest compared to the increase in computational cost. Consequently, selecting 3 learnable queries strikes the optimal balance between performance and computational efficiency.

Effectiveness of Dynamic Weighting Fusion Module

To assess the effectiveness of our proposed dynamic weighting fusion module, we compare it against three alternative fusion methods: average weighting, summation, and concatenation. In the average weighting approach, each of the three features is assigned an equal weight of 1/3, disregarding their individual importance. The summation method combines the three features through element-wise addition, while the concatenation approach merges features along a specified axis. As shown in Tab. 6, our dynamic fusion method outperforms all these methods by adaptively assigning weights to each feature based on the entropy of its prediction distribution.

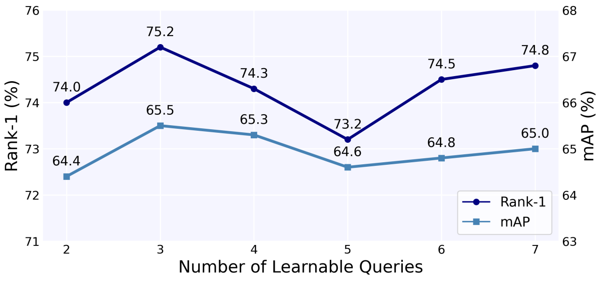

Figure 3: Impact of learnable query numbers on Occluded-Duke performance.

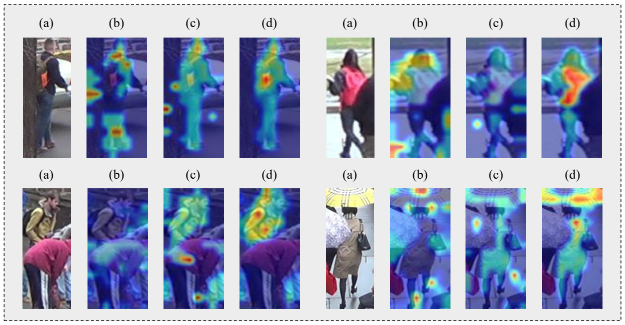

Figure 4: Visualization of attention maps. (a) Input images, (b) PFD, (c) CLIP-ReID, (d) FLaN-Net.

5 Visualization

To evaluate the model's ability to handle occluded images, we visualize attention maps generated by different methods, as shown in Fig. 4. The PFD is heavily distracted by occlusions, leading to less effective attention. CLIP-ReID demonstrates improved attention but still captures irrelevant areas in some cases. In contrast, our proposed FLaN-Net focuses precisely on the visible and identity-relevant regions of the pedestrian, avoiding interference from occlusions. This visualization highlights the robustness of FLaN-Net in handling diverse occlusion scenarios and its ability to focus on the most discriminative features for person re-identification.

6 Conclusions

In this paper, we propose FLaN-Net, an innovative framework specifically designed to address the challenges of occluded person re-identification. By employing a categorical attention mechanism, FLaN-Net generates fine-grained prompts capturing individual descriptions, visible attributes, and occluding object characteristics. The integration of cross-attention mechanisms and a dynamic weighting fusion module enables the model to focus on core identity while mitigating the impact of occlusions. Experimental results across multiple datasets demonstrate that FLaN-Net achieves state-of-the-art performance, underscoring its robustness and effectiveness in challenging real-world scenarios. This work paves the way for the development of more advanced occlusion-aware ReID solutions and highlights the potential of integrating vision and language for robust identity recognition.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62472420.

References

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[Chen *et al.*, 2021] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11833–11842, 2021.

[Deng *et al.*, 2019] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[Ding *et al.*, 2022] Changxing Ding, Kan Wang, Pengfei Wang, and Dacheng Tao. Multi-task learning with coarse priors for robust part-aware person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1474–1488, 2022.

[Dong *et al.*, 2021] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. *Advances in Neural Information Processing Systems*, 34:21898–21909, 2021.

[Dou *et al.*, 2023] Shuguang Dou, Cairong Zhao, Xinyang Jiang, Shanshan Zhang, Wei-Shi Zheng, and Wangmeng Zuo. Human co-parsing guided alignment for occluded person re-identification. *IEEE Transactions on Image Processing*, 32:458–470, 2023.

[Gal *et al.*, 2022] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[Gao *et al.*, 2020a] Lishuai Gao, Hua Zhang, Zan Gao, Weili Guan, Zhiyong Cheng, and Meng Wang. Texture semantically aligned with visibility-aware for partial person re-identification. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3771–3779, 2020.

[Gao *et al.*, 2020b] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11744–11752, 2020.

[He *et al.*, 2021] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[Hou *et al.*, 2021] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Feature completion for occluded person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4894–4912, 2021.

[Jia *et al.*, 2022] Mengxi Jia, Xinhua Cheng, Shijian Lu, and Jian Zhang. Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Transactions on Multimedia*, 25:1294–1305, 2022.

[Jia *et al.*, 2023] Mengxi Jia, Yifan Sun, Yunpeng Zhai, Xinhua Cheng, Yi Yang, and Ying Li. Semi-attention partition for occluded person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 998–1006, 2023.

[Jin *et al.*, 2020] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11173–11180, 2020.

[Li *et al.*, 2014] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[Li *et al.*, 2021] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2898–2907, 2021.

[Li *et al.*, 2023] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1405–1413, 2023.

[Li *et al.*, 2024] Yanping Li, Yizhang Liu, Hongyun Zhang, Cairong Zhao, Zhihua Wei, and Duoqian Miao. Occlusion-aware transformer with second-order attention for person re-identification. *IEEE Transactions on Image Processing*, 33:3200–3211, 2024.

[Miao *et al.*, 2019] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 542–551, 2019.

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

[Saito *et al.*, 2023] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023.

[Suo *et al.*, 2024] Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26951–26962, 2024.

[Tan *et al.*, 2022] Lei Tan, Pingyang Dai, Rongrong Ji, and Yongjian Wu. Dynamic prototype mask for occluded person re-identification. In *Proceedings of the 30th ACM international conference on multimedia*, pages 531–540, 2022.

[Tan *et al.*, 2024] Lei Tan, Jiaer Xia, Wenfeng Liu, Pingyang Dai, Yongjian Wu, and Liujuan Cao. Occluded person re-identification via saliency-guided patch transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5070–5078, 2024.

[Wang *et al.*, 2020] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6449–6458, 2020.

[Wang *et al.*, 2022a] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2540–2549, 2022.

[Wang *et al.*, 2022b] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. Feature erasing and diffusion network for occluded person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4754–4763, 2022.

[Xia *et al.*, 2024] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6198–6206, 2024.

[Yang *et al.*, 2024] Zexian Yang, Dayan Wu, Chenming Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. A pedestrian is worth one prompt: Towards language guidance person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17343–17353, 2024.

[Ye *et al.*, 2021] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[Zhang *et al.*, 2020] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3186–3195, 2020.

[Zhang *et al.*, 2021] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. Hat: Hierarchical aggregation transformers for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 516–525, New York, NY, USA, 2021. Association for Computing Machinery.

[Zhang *et al.*, 2022] Mingyang Zhang, Yang Xiao, Fu Xiong, Shuai Li, Zhiguo Cao, Zhiwen Fang, and Joey Tianyi Zhou. Person re-identification with hierarchical discriminative spatial aggregation. *IEEE Transactions on Information Forensics and Security*, 17:516–530, 2022.

[Zhang *et al.*, 2023] Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14133–14142, 2023.

[Zhao *et al.*, 2023] Cairong Zhao, Zefan Qu, Xinyang Jiang, Yuanpeng Tu, and Xiang Bai. Content-adaptive auto-occlusion network for occluded person re-identification. *IEEE Transactions on Image Processing*, 2023.

[Zheng *et al.*, 2015a] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[Zheng *et al.*, 2015b] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 4678–4686, 2015.

[Zheng *et al.*, 2017] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017.

[Zhu *et al.*, 2022] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4692–4702, 2022.

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

[Zhuo *et al.*, 2018] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *2018 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2018.