PART REPRESENTATION LEARNING WITH TEACHER-STUDENT DECODER FOR OCCLUDED PERSON RE-IDENTIFICATION

Shang Gao[1,3]　　*Chenyang Yu*[1]　　*Pingping Zhang*[2,*]　　*Huchuan Lu*[1,2,3]

[1] School of Information and Communication Engineering, Dalian University of Technology
[2] School of Future Technology, School of Artificial Intelligence, Dalian University of Technology
[3] NingBo Institute of Dalian University of Technology

ABSTRACT

Occluded person re-identification (ReID) is a very challenging task due to the occlusion disturbance and incomplete target information. Leveraging external cues such as human pose or parsing to locate and align part features has been proven to be very effective in occluded person ReID. Meanwhile, recent Transformer structures have a strong ability of long-range modeling. Considering the above facts, we propose a Teacher-Student Decoder (TSD) framework for occluded person ReID, which utilizes the Transformer decoder with the help of human parsing. More specifically, our proposed TSD consists of a Parsing-aware Teacher Decoder (PTD) and a Standard Student Decoder (SSD). PTD employs human parsing cues to restrict Transformer's attention and imparts this information to SSD through feature distillation. Thereby, SSD can learn from PTD to aggregate information of body parts automatically. Moreover, a mask generator is designed to provide discriminative regions for better ReID. In addition, existing occluded person ReID benchmarks utilize occluded samples as queries, which will amplify the role of alleviating occlusion interference and underestimate the impact of the feature absence issue. Contrastively, we propose a new benchmark with non-occluded queries, serving as a complement to the existing benchmark. Extensive experiments demonstrate that our proposed method is superior and the new benchmark is essential. The source codes are available at https://github.com/hh23333/TSD.

Index Terms— Occluded Person Re-identification, Vision Transformer, Part Representation, Feature Distillation.

1. INTRODUCTION

Person re-identification (ReID) is the task of retrieving the query pedestrian across multiple non-overlapping cameras. In recent years, significant improvements have been achieved with the advancement of deep learning [2, 3, 4]. However, the accuracy of person ReID under occlusion is still unsatisfactory, which is a common situation in practical scenarios.

---

*Corresponding author (zhpp@dlut.edu.cn).

Fig. 1. (a) An example with the interference of similar occlusions (from the Occluded-Duke benchmark [1]). (b) An example with similar hard negative samples and the missing part in the positive sample (from our proposed benchmark). (c) The ranking setting in our benchmark, where holistic positive samples are ignored when calculating evaluation metrics.

The difficulty of occluded person ReID is primarily due to the occlusion disturbance and misalignment caused by incomplete target information. To address these issues, many previous methods [5, 6, 7, 8, 9] leverage external cues such as human pose or parsing to locate the human body parts and align the visible ones. Recently, inspired by the strong generality in various vision tasks [10, 11, 12], some methods [13, 8, 9] propose to adopt Transformer structures to extract and disentangle deep features for occluded person ReID. For example, Wang *et al.* [9] integrate patch features and key-point information with the Transformer decoder to enhance local patch features. Jia *et al.* [14] propose a semi-attention partition method, aiming to comply consistency with human parsing while keeping resistance against noisy supervision. However, they may suffer from the part misalignment problem. In this paper, a Teacher-Student Decoder (TSD) framework is proposed to utilize the Transformer decoder and focus on the corresponding body parts with the help of human parsing. Specifically, our TSD consists of a Parsing-aware Teacher Decoder (PTD) and a Standard Student Decoder (SSD). PTD restricts the attentions to specific body parts by human parsing cues. Then, a feature distillation is employed on the outputs of PTD and SSD to ensure them closer to each other. Thereby, SSD can learn from PTD to aggregate information from corresponding body parts automatically.

Meanwhile, existing occluded person ReID datasets only use occluded samples as queries. As shown in Figure 1(a),
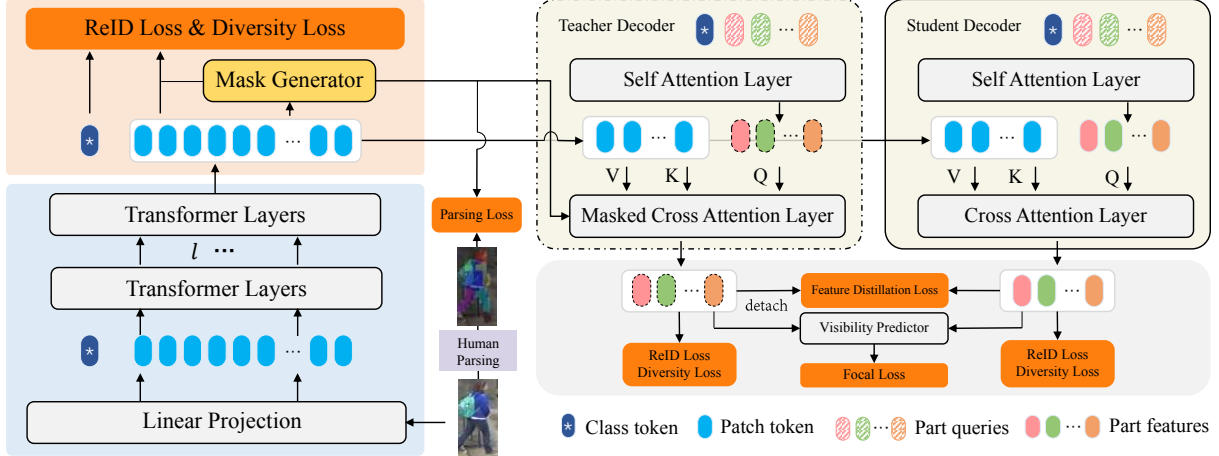
Fig. 2. Overview of our proposed approach.

the main concern is to eliminate the interference of similar occlusions. Therefore, it may underestimate the impact of the feature absence problem on evaluation metric. To solve these problems, we propose a new benchmark for occluded person ReID. As depicted in Figure 1(c), the query is a holistic sample and the positive holistic samples are ignored in the ranking list. As shown in Figure 1(b), the query and the hard negative samples have similar lower bodies, while the lower body of the positive sample is occluded. In this case, merely alleviating the noise caused by occlusion is insufficient. The misalignment problem caused by the missing local information also needs to be addressed, thereby it can be reflected in evaluation metric. Therefore, our benchmark can serve as a supplement to existing evaluation settings, and reflect the ReID performance under occlusion more comprehensively.

In summary, the main contributions of this work are three-fold: (1) A Teacher-Student Decoder (TSD) framework is proposed to incorporate the human parsing information into the Transformer for occluded person ReID. (2) A new benchmark is introduced to better evaluate the performance of occluded person ReID methods. (3) Extensive experiments on existing benchmarks and the proposed benchmark demonstrate the superiority of our method.

2. METHODOLOGY

Our proposed framework is illustrated in Figure 2. Our method adopts a pre-trained ViT [2] as the encoder to extract feature representations from the input image X. The global feature $F^g \in \mathbb{R}^D$ is the final output of the class token and $F^{pt} \in \mathbb{R}^{N \times D}$ is the final patch embedding. Here, N is the number of patches and D is the feature dimension. We elaborate on key components in the following sections.

2.1. Teacher-Student Decoder

Our proposed TSD consists of two key components, i.e., a Parsing-aware Teacher Decoder (PTD) and a Standard Student Decoder (SSD). They have the same structure and share

parameters. The only difference is that the teacher decoder takes external binary masks $M \in R^{P \times N}$ as inputs and restricts the discriminative regions for cross-attention. P is the number of parts. $M(p, i) = 1$ if the i-th location belongs to the p-th part and $M(p, i) = 0$ vice versa.

We define P semantic queries $Q^{em} = [q_1, q_2, ..., q_P]$ by a set of learnable embedding, each of which corresponds to a body part. Then they are concatenated with the class embedding and fed into the standard self-attention layer as:

$$[q^g, Q^{em}] = \text{MHSA}([F^g, Q^{em}]). \quad (1)$$

The MHSA is the Multi-Head Self-Attention. The reason for incorporating the class token is that it helps the part queries better aggregate the features of the target pedestrian by the following cross attention layer. In addition, the SSD directly employs the standard cross-attention as:

$$X^s = Softmax(QK^T/\sqrt{D})V, \quad (2)$$

where $Q = \varphi(Q^{em}) \in R^{P \times D}$ is the query embedding under linear projection φ. $K = \phi(F^{pt}) \in R^{N \times D}$ and $V = \psi(F^{pt}) \in R^{N \times D}$ are the image features under the transformations ϕ and ψ, respectively. Finally, the output of the student decoder can be obtained by:

$$F^{sd} = \text{FFN}(\text{LN}(X^s)), \quad (3)$$

where FFN and LN denote the feed-forward network and layer normalization [15], respectively. The PTD employs masked attentions via:

$$X_p^t = Softmax(H_p + Q_p K^T)V, \quad (4)$$

where X_p^t is the p-th part feature, Q_p is the p-th query embedding and $H_p \in R^N$ is defined by:

$$H_p(i) = \begin{cases} 0, & \text{if } M(p, i) = 1, \\ -\infty, & \text{otherwise.} \end{cases} \quad (5)$$

2661

The output of the teacher decoder can be obtained by:

$$F^{td} = \text{FFN}(\text{LN}(X^t)), \qquad (6)$$

then the feature distillation loss is employed to ensure each student query focus on a specific body part:

$$L_{fd} = \frac{1}{P} \sum_i (1 - Sim(F_i^{sd}, F_i^{td})), \qquad (7)$$

where $Sim(\cdot)$ denotes the cosine similarity. In addition, L_{fd} may lead the decoder to extract identical features. Thus, a diversity loss is employed on the part features:

$$L_{div} = \frac{1}{P(P-1)} \sum_{i=1, i \neq j}^{P} \sum_{j=1}^{P} Sim(F_i^{td}, F_j^{td}). \qquad (8)$$

2.2. Mask Generation

The mask can be obtained by directly employing human parsing methods [16]. However, it suffers from domain gaps and the lack of end-to-end training leads to generic masks not suitable for ReID. Therefore, we propose a mask generator with both a body part prediction objective and a ReID objective as in [6]. The mask generator consists of a fully-connected layer with parameters $G \in R^{(P+1) \times D}$ followed by a softmax layer. Thus, the part heatmaps M can be obtained by:

$$M = Softmax(F^{pt}G^T). \qquad (9)$$

To generate the semantic mask as the input of PTD, we perform the $argmax$ operation on M.

2.3. Overall Training and Inference Procedure

Training Procedure: To optimize the proposed framework, the overall loss is formulated as:

$$L = L_{ce}(F^g) + L_{tri}(F^g) + L_{ce}(F_c^{sd}) + L_{tri}^p(F^{sd}) + L_{ce}(F_c^{td})$$
$$+ L_{tri}^p(F^{td}) + L_m + L_{fd} + L_{div} + L_v, \qquad (10)$$

where L_{ce} is the cross-entropy loss with the BNNeck trick [17]. L_{tri} is the triplet loss [18]. F_c^{sd} and F_c^{td} are produced by concatenating the P part features along the channel dimension. L_{tri}^p is the part average triplet loss [6]. L_m is used to train the mask generation as in [6]:

$$L_m = L_{ce}(F_c^{part}) + L_{tri}^p(F^{part}) + L_{pa}(M) \qquad (11)$$

$$F_i^{part} = \frac{\sum_{l=0}^{N} M(l)F^{pt}(l)}{\sum_{l=0}^{N} M(l)} \qquad (12)$$

$$F_c^{part} = Concat(F_1^{part}, F_2^{part}, ..., F_P^{part}). \qquad (13)$$

Here, L_{pa} is a cross-entropy loss with label smoothing. L_v is the focal loss [19] to supervise the visibility v_p by a binary classifier upon F^{sd}:

$$L_v = \begin{cases} -\alpha(1 - v_p)^\gamma \log(v_p), & \text{if } \hat{v}_p = 1, \\ -(1-\alpha)v_p^\gamma \log(1 - v_p), & \text{otherwise,} \end{cases} \qquad (14)$$

where $\hat{v}_p$ is the label of the p-th part visibility defined by the human parsing as in [6], α and γ are parameters to balance positive vs. negative and hard vs. easy samples, respectively. **Inference Procedure:** The mask generator and teacher decoder are not needed in the inference procedure. The visibility-based part-to-part matching strategy [6, 5] is adopted to calculate the distance of the query q and gallery sample g:

$$d = \frac{\sum_{i \in \{g, 1, ..., P\}} (v_i^q \cdot v_i^g \cdot dist(F_i^q, F_i^g))}{\sum_{i \in \{g, 1, ..., P\}} (v_i^q \cdot v_i^g)}. \qquad (15)$$

Here, $\{F_i | i \in 1, ..., P\}$ is the part feature from the student decoder and $dist$ is the Euclidean distance. Finally, the ranking list is achieved by sorting the distance.

3. NEW BENCHMARK

To thoroughly measure the ReID performance under occlusion, we introduce the Re-Occluded-Duke dataset, which is a reorganized version of the Occluded-Duke dataset [1].

3.1. Properties of Re-Occluded-Duke

We maintain the same training set as the Occluded-Duke dataset [1], which consists of 15,618 images from 702 identities. Then, we merge the original query and gallery sets into a new gallery set, which contains 18,001 images from 1,110 identities. For images from the identities that appear in the original query set, we manually annotate their occlusion status into non-pedestrian occlusion (NPO), non-target pedestrian occlusion (NTP) and holistic images. From these annotated images, we select all holistic images as the candidate set and randomly sample up to 5 images per identity to the query set. Finally, the query set contains 2538 holistic images from 517 identities.

3.2. New Evaluation Metrics

The Cumulative Matching Characteristic (CMC) curves and mean Average Precision (mAP) are common metrics for evaluating the performance of different person ReID methods. In order to better evaluate occluded samples, we propose several new evaluation metrics. Occ-CMC and Occ-mAP focus on the retrieval of occluded samples. They ignore the correct matches of holistic samples in the rank list. Similarly, NPO-CMC and NPO-mAP are metrics that consider non-pedestrian occlusion, while NTP-CMC and NPO-mAP are metrics that account for non-target pedestrian occlusion.

4. EXPERIMENTS

4.1. Datasets and Evaluation Metrics

We evaluate our model on the holistic DukeMTMC-reID dataset [20] and the occluded Occluded-Duke dataset [1] as well as our benchmark. We report the Rank-1 and mAP for the former two datasets and report the metrics as described in Sec. 3.2 for our benchmark.

Table 1. Comparison with other state-of-the-art methods on Occluded-Duke and DukeMTMC-reID. ∗ indicates the backbone is with an overlapping stride setting. † indicates it is reproduced by replacing the original backbone with ViT.

Method	Occluded-Duke		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
ViT-B [2]	61.5	53.5	88.8	79.3
TransReID [22]	64.2	55.7	89.6	80.6
BPBreID† [6]	66.0	56.7	90.2	80.8
PFD [9]	67.7	**60.1**	**90.6**	**82.2**
FED [23]	68.1	56.4	89.4	78.0
Ours	**70.6**	57.3	90.2	81.7
DPM* [24]	71.4	61.8	91.0	82.6
SAP* [14]	70.0	62.2	-	-
PFD* [9]	69.5	61.8	**91.2**	**83.2**
Ours *	**74.5**	**62.8**	90.8	82.8

Table 2. Comparison with other methods on our benchmark.

Method	OCC		NPO		NTP	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
VIT-B [2]	67.1	52.5	60.8	51.1	60.1	51.4
FED [23]	63.9	47.4	57.6	46.0	56.7	46.6
BPBreID* [6]	67.8	54.1	61.5	53.4	59.0	50.4
DPM [24]	69.2	53.5	62.0	50.8	63.6	53.9
PFD [9]	70.9	55.7	64.8	54.3	**64.6**	**55.2**
Ours	**71.4**	**58.7**	**68.0**	**61.5**	61.9	52.5
SAP* [14]	71.4	57.1	65.8	55.4	**65.4**	56.6
Ours *	**73.2**	**61.7**	**68.8**	**62.7**	64.9	**57.5**

4.2. Implementation Details

We adopt the ViT-B [2] as our backbone. All images are resized to 256×128. The training images are augmented with random erasing [21], padding, and random cropping for all experiments. The batch size is set to 64 with 4 images per ID. The SGD optimizer is employed with a momentum of 0.9 and the weight decay of $1e^{-4}$. The learning rate is initialized as 0.004 with a cosine learning rate decay. We train the model for 120 epochs. The number of parts P is set to 8. We adopt the same human parsing model as in [6].

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

We compare our model with other state-of-the-art methods on Occluded-Duke and DukeMTMC-reID in Table 1. All of them adopt Transformers as the backbone. Our method outperforms all the competitors on Occluded-Duke dataset. The main reason is that our TSD guides each query adaptively focus on a corresponding body part, thereby better disentangling pedestrian features. Furthermore, our method achieves competitive results on the holistic dataset DukeMTMC-reID.

Table 2 shows the result of our method and other methods on our new benchmark. It can be observed that our method can achieves superior performance, especially on the NPO samples. It is noteworthy that FED [23] performs well on the Occluded-Duke dataset. However, it is inferior on our new benchmark, even worse than the baseline method. The main reason is that it only considers the noise introduced by occlusion, but ignores the misalignment issue.

Table 3. Ablation study for the main components on Occluded-Duke and our benchmark.

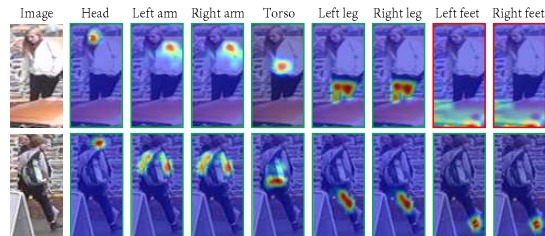| Method	Occluded-Duke		OCC	NPO	NTP
	Rank-1	mAP	mAP	mAP	mAP
Baseline	61.5	53.5	52.3	50.2	50.8
M1	59.9	51.5	49.0	47.4	48.4
M2	65.1	54.3	54.6	57.8	48.6
M3	68.3	54.9	54.5	56.8	48.2
M4	70.6	57.3	58.7	61.5	52.5

Fig. 3. Visualization of attention maps. Green boxes indicate visible predictions, while red boxes for invisible ones.

4.4. Ablation Study

In this section, we conduct ablation studies on the Occluded-Duke and our benchmark to analyze the effectiveness of each component. The results are shown in Table 3. The baseline is with the standard vision Transformer [22]. M1 denotes the pure Transformer encoder-decoder architecture. Its results are inferior to the baseline. The possible reason is that the model can not adaptively decouple features in the absence of auxiliary supervision, and the newly initialized decoder may even bring a negative effect. From M2, when the masked teacher is added, the performance is greatly improved. This indicates that the masked teacher can effectively aggregate local features and transfer the information to guide the student decoder. M3 denotes the diversity loss is further added to M2. It can be observed that the diversity loss also brings performance improvement. By replacing the fixed mask with a learnable one, M4 surpasses M3. It shows our mask generator can effectively identify distinctive regions. Furthermore, Figure 3 shows the attention maps towards each query of the decoder. Our method can successfully focus on different body parts and precisely predict the visibility. It clearly shows the effectiveness of extracting discriminative regions.

5. CONCLUSION

In this paper, a Teacher-Student Decoder (TSD) framework is proposed to incorporate the human parsing information into the Transformer structure for occluded person ReID. In addition, a new benchmark is introduced to better evaluate the ReID performance under occlusion. Extensive experiments demonstrate the superiority of our method.

6. REFERENCES

[1] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang, "Pose-guided feature alignment for occluded person re-identification," in *ICCV*, 2019, pp. 542–551.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2010.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[4] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al., "Deep high-resolution representation learning for visual recognition," *IEEE TPAMI*, vol. 43, no. 10, pp. 3349–3364, 2020.

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

[6] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi, "Body part-based representation learning for occluded person re-identification," in *WACV*, 2023, pp. 1613–1623.

[7] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *ICCV*, 2019, pp. 8450–8459.

[8] Mengxi Jia, Xinhua Cheng, Shijian Lu, and Jian Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE TMM*, 2022.

[9] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *AAAI*, 2022, vol. 36, pp. 2540–2549.

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *ECCV*. Springer, 2020, pp. 213–229.

[11] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang, "Unihcp: A unified model for human-centric perceptions," in *CVPR*, 2023, pp. 17840–17852.

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022, pp. 1290–1299.

[13] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu, "Hat: Hierarchical aggregation transformers for person re-identification," in *ACM MM*, 2021, pp. 516–525.

[14] Mengxi Jia, Yifan Sun, Yunpeng Zhai, Xinhua Cheng, Yi Yang, and Ying Li, "Semi-attention partition for occluded person re-identification," in *AAAI*, 2023, vol. 37, pp. 998–1006.

[15] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.

[16] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang, "Self-correction for human parsing," *IEEE TPAMI*, vol. 44, no. 6, pp. 3260–3271, 2020.

[17] 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*, 2019, pp. 0–0.

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

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.

[20] Zhedong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017, pp. 3754–3762.

[21] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang, "Random erasing data augmentation," in *AAAI*, 2020, vol. 34, pp. 13001–13008.

[22] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang, "Transreid: Transformer-based object re-identification," in *ICCV*, 2021, pp. 15013–15022.

[23] 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*, 2022, pp. 4754–4763.

[24] Lei Tan, Pingyang Dai, Rongrong Ji, and Yongjian Wu, "Dynamic prototype mask for occluded person re-identification," in *ACM MM*, 2022, pp. 531–540.