Uncover the Body: Occluded Person Re-identification via Masked Image Modeling

Kunlun Xu[1,2], Yuxin Peng[1,2], and Jiahuan Zhou[1,2(✉)]

[1] Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China
[2] National Key Laboratory for Multimedia Information Processing, Peking University, Beijing 100871, China
jiahuanzhou@pku.edu.cn

Abstract. Person re-identification (ReID) has attracted tremendous attention and achieved significant progress on holistic data where the whole body of a pedestrian is completely presented. However, in a more realistic scenario where pedestrians are partially occluded, the discriminative ability of the existing ReID methods is severely limited since the visual information of the pedestrians becomes noisy and unreliable. To alleviate this issue, current solutions mostly pay more attention to visible body parts for extracting fine-grained features. Nevertheless, different occluded parts on different images of the same pedestrian always result in inaccuracy matching. In this paper, we propose an Uncover the Body Network (UBN) which exhibits the ability to remove the occlusion and attempt to restore the full body of a pedestrian. The proposed UBN can alleviate the noise brought by occlusions and extract more robust feature representations. To achieve this, we propose a MIM (Masked Image Modeling) based method for its powerful representation of partial images to the whole. Instead of randomly masking the images, we propose a Mask Prediction Module (MPM) to readily locate the occluded patches, and an occlusion-guided masking strategy is adopted to facilitate the learning. Extensive experimental results on both the occluded and holistic ReID benchmarks have demonstrated the superiority of UBN against the state-of-the-art approaches.

Keywords: Occluded Person Re-identification · Masked Image Modeling · Retrieval.

1 Introduction

Person Re-Identification (ReID) has played an important role in many practical computer vision tasks such as video surveillance [1], forensic tracking [2], and so on. Over the past years, most of ReID methods [3–6] concentrated on processing holistic data where the whole body of a pedestrian is completely visible. However, in a more realistic scenario where pedestrians are partially occluded by various obstacles, the discriminative ability of the existing ReID methods is

severely limited since the visual information of the pedestrians becomes noisy and unreliable, leading to deteriorated performance for occluded ReID [7, 8].

To mitigate the influence of occlusions, various occluded ReID methods [8–11], have been proposed which can be roughly categorized into two groups: keypoint-based methods and feature pyramid-based ones. The former group [9, 11] focuses on extracting informative features from the visible keypoint parts estimated by off-the-shelf pose estimation models. The latter group [8, 10] aims to extract multi-scale features from both the query and gallery images to alleviate the influence of occlusions. However, both groups rely on the visible person region matching across query and gallery images, which is sensitive to the occlusion distribution between different images.

In this paper, by thoroughly exploring the visible parts of a pedestrian image, we propose to suppress the adverse effects of occlusions by uncovering the occluded parts. Motivated by the recent Masked Image Modeling (MIM) research [12], deep networks have exhibited the superior ability to recover factual visual information only based on the remaining visible parts, even if the proportion of the visible parts is small. Therefore, we propose an Uncover the Body Network (UBN) which exhibits the ability to remove the occlusion and attempt to restore the full body of a pedestrian. A novel Occlusion-aware Mask Prediction Module (MPM) and a Masked Image Reconstruction Module (MIR) are designed accordingly where the MPM can automatically generate mask maps to determine the dropping patches with respect to the occlusion obstacles. Moreover, the proposed MPM utilizes learnable embedding to replace masked patches for mask map generation. As for MIR, it takes the aforementioned masked images as inputs to reconstruct the holistic person images.

Our proposed UBN can readily mitigate the above issues in existing occluded ReID methods. On the one hand, our UBN can benefit the keypoint-based approaches by recovering the occluded parts, which will enhance the ability of pose estimation models and enrich the keypoints for discriminative feature extraction. On the other hand, the proposed UBN will benefit the feature pyramid-based methods by eliminating the adverse influence of occlusion obstacles which can enhance the discriminative ability of the obtained features via recovering reasonable appearance information. Extensive experimental results have demonstrated that our UBN achieves state-of-the-art performance on various ReID benchmarks, exceeding the latest baselines by a large margin. To sum up, our contributions are three-fold:

- A novel occluded ReID model named Uncover the Body Network (UBN) is proposed which consists of a Mask Prediction Module (MPM) and a Masked Image Reconstruction Module (MIR). Thus, the MPM generates mask maps to automatically decide the occluded parts in images for masking and the MIR takes the masked images as inputs to reconstruct reasonable holistic person images.
- To facilitate MPM and MIR learning, four mask supervision strategies are readily designed. Both the subjective and objective evaluation results demon-

strate that all strategies could simultaneously promote the ReID performance and the mask prediction results.
– The conducted extensive experimental experiments on various ReID benchmarks have demonstrated that our UBN achieves state-of-the-art performance against the latest baselines by a large margin.

2 Related Work

2.1 Holistic Person ReID

Person ReID aims to identify the same pedestrian captured by different cameras at different locations and different time. Most existing ReID methods [4, 13, 6] focus on holistic data where the whole body of a pedestrian is clearly presented. Although performs well on holistic data, they suffer serious performance degradation when they are applied to partial and occluded person images, which indeed appear frequently in a more realistic application scenario. Different from them, our proposed UBN can not only tackle the holistic ReID task, but also makes a breakthrough in the scenario with heavy occlusions.

2.2 Occluded Person ReID

Existing occluded ReID methods either utilize human pose [9, 11] or extract the feature pyramid [8, 10] to facilitate part-level person matching. However, these methods can not accurately predict the occluded body parts which is not coherent with how humans tackle occlusion scenarios. and most of them are sensitive to fine-grained extra cues and are not robust to variation of occlusions. Compared with the aforementioned methods, In contrast, our proposed UBN manages to remove the occlusion obstacles and uncover the full body. Therefore, a more robust and discriminative feature representation can be obtained.

2.3 Masked Image Modeling

Recently, Masked Image Modeling (MIM) [14–18, 12] becomes an effective self-supervised pre-training manner to provide initial weights with strong representation capacity for downstream tasks [15]. The recent works MAE [17] and SimMIM [12] have demonstrated that with partial visible patches, an image with complete structural information could be successfully reconstructed. This motivates us that the occluded body parts of a pedestrian could be reconstructed based on the remaining visible parts, which could lead to a complete feature representation for ReID. However, existing MIM approaches aim to recover all the patches in the original images, which means that the occluded patches will keep being occluded in the reconstructed images. Therefore, it's vital to redesign the pipeline that guides the network to reconstruct the human body preferentially.

3 Methodology

Given a query set $Q = \{q_1, q_2, ..., q_n\}$ and a galley set $G = \{g_1, g_2, ..., g_m\}$, the goal of ReID is to compute the match scores of each image in Q and G. For the sake of convenience, we reasonably assume that the width and height of a query or galley image is W and H respectively.

3.1 Algorithm Overview

The overall pipeline of our proposed UBN is demonstrated in Figure 1. Our model mainly contains three crucial parts including an Occlusion-aware Mask Prediction Module (MPM), a Masked Image Reconstruction Module (MIR), and a TransReID-based feature extractor. More specifically, MPM processes the original occluded images by generating the occlusion-aware mask maps. Then, MIR aims to recover the complete visual appearance according to MPM's input images and predicted masks. Finally, a TransReID-based feature extractor module is adopted to leverage the recovered contextual information from MIR to facilitate feature learning.

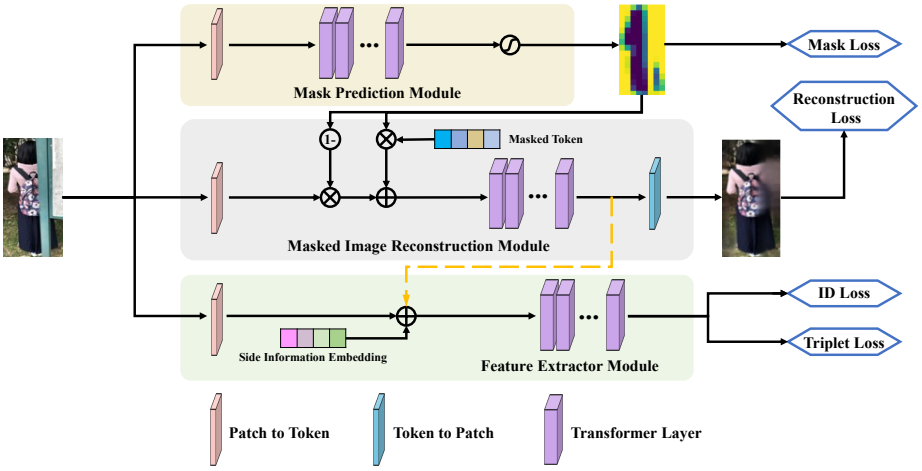

Fig. 1. The overall architecture of our proposed Uncover the Body Network (UBN).

3.2 MPM: Occlusion-Aware Mask Prediction Module

Recall existing MIM methods [17, 12, 16], all of them randomly mask image patches or pixels and then force the model to reconstruct the masked areas. In occluded ReID, we expect to leverage the visible person parts to predict the occluded body parts. To do so, we propose a Mask Prediction Module (MPM) to learn the occlusion-aware mask maps according to the global context. As illustrated in Figure 1, the proposed MPM consists of a patch-to-token layer along with multiple transformer layers [19, 3] in which the ultimate transformer will generate token-level mask predictions. The patch-to-token operation follows the standard ViT [19] that divides the given image into non-overlapping patches and then maps each patch into a token vector representation, after which the positional information would be added.

3.3 MIR: Masked Image Reconstruction Module

Based on the predicted occlusion-aware mask maps from MPM, our UBN will further recover the occluded patches to complete the holistic person images. Therefore, a Masked Image Reconstruction Module (MIR) is designed for holistic

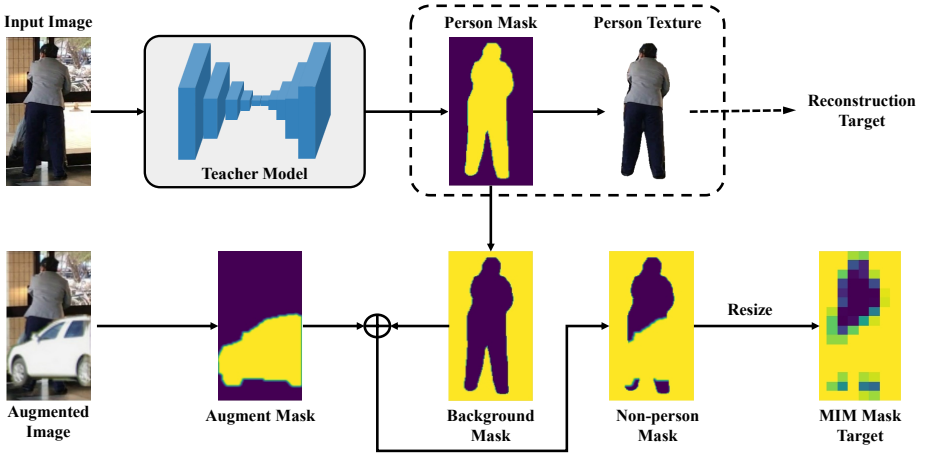

Fig. 2. An example of reconstruction target and MIR mask target generation. The image is from the holistic PRID dataset. The teacher model is a pre-trained instance segmentation model. Person mask is the biggest person mask in the image. Person texture is extracted from the image according to the person mask. Augmented image is generated by pasting a external object and it serves as the input image of MPM and MIR. The Augment mask is the mask map of pasting area. The background mask is the opposite of person mask. Non-person mask is the pixel-wise binary sum of the Augment mask and background mask. MIR mask target is sub-sampled from Non-person mask in order to fit the resolution of tokens.

person generation which mainly contains three components: MIR Input, MIR Encoder and MIR Head.

MIR Input: Given a token t_i generated from the original image, we generate a MIM input token t_i' by

$$t_i' = Mask_i * t^m + (1 - Mask_i) * t_i, \tag{1}$$

where $Mask_i$ is the Mask score of ith token and t^m is the learnable mask token. When $Mask_i$ is 0 or 1, Equation (1) is equivalent to the mask operation of SimMIM and MaskFeat in which the original token information is totally kept or dropped. When $Mask_i \in (0, 1)$, it denotes ith token partly contribute to person construction.

MIR Encoder Similar to MAE, MaskFeat and SimMIM, we also use transformer layers as the basic module of feature encoder. The transformer parameters follow the default setting of ViT-B [20, 19].

MIR Head We use a simple linear layer as MIR Head to map the tokens from the encoder into RGB pixels.

3.4 FEM: TransReID-based Feature Extractor Module

Once the occluded person images are recovered by the proposed MPM and MIR modules, a TransReID-based [3] Feature Extractor Module (FEM) is utilized to obtain the final feature representations for the given images. FEM is an efficient

and effective backbone that adapts the ViT to the ReID task to construct the data stream from the original occluded person images to the final embedding. As shown in Figure 1. the proposed FEM firstly utilizes a patch-to-token layer to map the original image into several tokens T_{reid}. Then the MIR token and Side Information Embedding (SIE) are added to T_{reid} to supplement more information such as holistic human body and camera IDs. The addition weights of MIR token and SIE are λ_{mir} and λ_{sie} respectively. Moreover, the fused token is fed to l transformer layers which will eventually extract a discriminative and robust feature representation for cross-image similarity calculation. Therefore, based on FEM, our proposed method can readily leverage the holistic information recovered by MIR to supplement occluded information.

3.5 Model Training

The MPM and MIR together formulate a complete pipeline for recovering the helpful information of the occluded parts. Therefore, the key issue here is how to guide MPM to identify which token should be dropped and meanwhile promote MIR to generate a holistic person image without occlusions. To tackle the above issue, both the MPM and MIM are pre-trained using holistic person images and then the whole model with all the three proposed modules will be jointly trained to accomplish effective Person ReID.

MPM and MIR Pre-training Given an input holistic person image x, our method automatically generates an occluded image x^a by introducing an external object to x as the occlusion obstacle. To guide the model to reconstruct a holistic person without occlusions, we adopt an instance segmentation model, *i.e.* CBNetV2 [21], as the teacher model to generate person mask $M_p = \mathbb{R}^{H \times W}$. We use M_p to get the visible person areas and adopt these areas as the reconstruction target. Given original image x and reconstructed image x', the Reconstruction loss is calculated by:

$$L_R = ||(x' - x) * M_p||. \tag{2}$$

To accomplish MPM learning, we provide four kinds of supervision strategies:

1) Non-person supervision. Given person mask M_p, the background mask M_{bk} can be calculated simply by $M_{bk} = 1 - M_p$. Besides, after implementing occlusion augmentation, the augmentation mask M_{aug} could also be obtained. Therefore, the non-person mask M_{non-p} could be calculated by $M_{non-p} = M_p + M_{aug}$. The Non-person supervision takes M_{non-p} resized to $H/K \times W/K$ as the target, where K is the patch size of each token. The core idea of this supervision is to drop the patches not containing the person and keep the patches containing the person.

2) Occlusion supervision. The occlusion mask M_{occ} is calculated by $M_{occ} = M_p \times M_{aug}$. The Occlusion supervision takes M_{occ} resized to $H/K \times W/K$ as the target and the key idea is to drop the occluded patches.

3) All-drop supervision. In this setting, the target is a mask map with each element set to 1, which means that no image information should be kept. However, since MIR needs vital image patches, such as those containing human body parts, to reconstruct a holistic person, it will drive MPM to output a lower mask

score in those areas. Eventually, the learned mask map will have larger values in occluded and background patches, and smaller values in visible body patches.

4) All-keep supervision. This setting is the opposite strategy of 3). The target is a mask map with each element set to 0, which means that all image information should be kept. In such condition, MIR will drive MPM to drop the patches harmful to holistic person reconstruction. Eventually, the learned mask map will have bigger values at occlusion patches and smaller valves at the visible body patches and background patches.

Figure 2 illustrates an example of the reconstruction target and target generation pipeline for Non-person supervision. For the above four supervision strategies, we adopt Binary Cross Entropy Loss to calculate Mask Loss L_m. The performance of each strategy is discussed in the **Ablation Studies**.

The overall loss of MPM and MIR Pre-training is calculated by:

$$L_{pre} = L_R + \lambda_1 L_m, \tag{3}$$

where λ_1 is super parameter to balance the loss weight of MPM and MIR. In our experiments, λ_1 is set to 1 for strategy 1) and 2), and for strategy 3) and 4), λ_1 is set to 0.001.

Joint Training The Joint Training procedure aims to train a RRID model robust to occlusion. Given a wild occluded image, as illustrated in Figure 1 we pass the image through MPM, MIR, and FEM successively. For supervision, we only adopt ID loss and triplet loss at the last layer of the FEM. The ID loss refers to cross entropy loss, which calculated by

$$L_{ID} = -y_i \log \left(\frac{\exp\left(W_i f_i\right)}{\sum_{j=1}^{ID_s} \exp\left(W_j f_j\right)} \right), \tag{4}$$

where y is the ground truth, f is the extracted feature, W is a linear projection matrix.

The triplet loss is soft-margin loss which minimizes the gap between positive samples and maximizes the gap between positive and negative samples. Functionally, the loss can be presented as:

$$L_{Tri} = log(1 + e^{\|f_a - f_p\|_2^2 - \|f_a - f_n\|_2^2}) \tag{5}$$

$$L = L_{ID} + L_{Tri}, \tag{6}$$

where $\langle$a,p,n$\rangle$ is a triplet set $\langle$anchor, positive sample, negative sample$\rangle$.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets We conduct all experiments on four ReID datasets, including two holistic datastes Market1501 [22] and DukeMTMC-reID [23] as well as two occluded datasets Occluded-DukeMTMC [9] and Occluded-REID [7].

Evaluation Metrics To perform a fair comparison with existing methods, all experiments follow the common evaluation settings in person ReID methods. The Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) are adopted to evaluate the performance. All experiments are performed in the single query setting.

4.2 Implementation

In our experiments, images are resized to 256×128. During MPM and MR pre-training, we collect the images from three holistic datasets: Market1501, DukeMTMC-reID and MSMT17. The basic pre-training augmentation includes Horizontal-Flipping, Random-Crop, Random-Rotation, and Colorjitter. Besides, we also adopt occlusion augmentation which randomly pasted external objects on the image. The external objects are extracted from MS-COCO valset [24]. We use four Nvidia 3090 GPU to pre-train MPM and MIR for 400 epochs.

During Person ReID training, we first augment the training images by random horizontal flipping, padding, random cropping and random erasing. The batch size is set to 64 and each one concludes 16 identities. And we train the whole network for 160 epochs using the SGD optimizer with a momentum of 0.9 and weight decay of 1e-4. And the initial learning rate is 0.008 with cosine learning rate decay. We embed the MIR-feature into the TransReID baseline using the token-level feature fusion approach and the weight of SIE(λ_s) and the weight of MIR token embedding (λ_m) are respectively set to 3.0 and 1.0. Furthermore, PRID training experiments are conducted on one Nvidia 3090 GPU.

Table 1. Performance comparison with state-of-the-art methods on four datasets, including Occluded-DukeMTMC (O-Duke), Occluded-REID (O-REID), Market1501 and DukeMTMC-reID (DukeMTMC).

	O-Duke		O-REID		Market1501		DukeMTMC	
Method	mAP	R1	mAP	R1	mAP	R1	mAP	R1
PCB [25]	33.7	42.6	38.9	41.3	77.4	92.3	66.1	81.8
RE [26]	30.0	40.5	-	-	71.3	87.1	62.4	79.3
FD-GAN [27]	-	40.8	-	-	77.7	90.5	64.5	80.0
DSR [28]	30.4	40.8	62.8	72.8	75.6	91.3	68.7	82.4
SFR [29]	32.0	42.3	-	-	81.0	93.0	71.2	84.8
FRR [30]	-	-	68.0	78.3	86.6	95.4	78.4	88.6
ISP [31]	52.3	62.8	-	-	88.6	95.3	80.0	89.6
PGFA [9]	37.3	51.4	-	-	76.8	91.2	65.5	82.6
HOReID [10]	43.8	55.1	70.2	80.3	84.9	94.2	75.6	86.9
OAMN [8]	46.1	62.6	-	-	79.8	92.3	72.6	86.3
PAT [32]	53.6	64.5	72.1	**81.6**	88.0	95.4	78.2	88.8
TransReID [3]	55.7	64.2	67.3	70.2	88.2	95.0	80.6	89.6
Ours(UBN)	**57.3**	**65.2**	**74.8**	79.9	**88.6**	**95.5**	**81.1**	**90.2**

4.3 Comparison with State-of-the-art Methods

As shown in Table 1, we compare our method with the state-of-the-art approaches on four datasets, including Market-1501, DukeMTMC-reID, Occluded-DukeMTMC, and Occlude-REID. The results demonstrate we can achieve excellent results on both occluded and holistic datasets.

On Occluded Datasets On the challenging Occluded-DukeMTMC, UBN achieves state-of-the-art performance with an mAP of 57.3% and R1 of 65.2%. This performance surpasses classical hand-crafted methods such as PCB [25] and FD-GAN [27], which use keypoints of pedestrians, by 23.6% and 22.6%, respectively. The performance also surpasses PGFA [9], a famous pose-guided approach based on a CNN backbone, by 20.0% and 13.8%. Moreover, the UBN model outperforms the transformer-based PAT by 3.7%/0.7% and TransReID by 1.6%/1.0%.

As for Occluded-REID, which can only be used for testing, we evaluate on it with our model trained on Market1501. The model achieved an mAP of 74.8% and R1 of 79.9%, which outperforms most recent methods. Notably, the achieved mAP of 74.8% on this dataset is state-of-the-art performance. However, as transformer-based architectures struggle to obtain strong generalization results on small training sets [33], the UBN model failed to achieve the highest R1 accuracy on this dataset.

On Holistic Datasets On Market-1501, we obtain 0.4% (mAP) and 0.5% (R1) improvement over TransReID. On DukeMTMC, we also achieve 0.5% (mAP) and 0.6% (R1) improvement. As far as we know, we are the first to achieve R1 over 90% on DukeMTMC-reID among approaches focusing on occluded PRID. Therefore, UBN is also capable of promoting model learning on holistic datasets, albeit with smaller improvements compared to the occluded ones. This demonstrates that the uncovering occlusion design can benefit overall PRID performance.

4.4 Ablation Studies

Ablation Study of Mask Supervision Strategies In Section (3.5), we introduce four mask supervision strategies: non-person supervision, occlusion supervision, all-drop supervision and all-keep supervision. In Table 2, we conduct comparison experiments of our four strategies (line 2-5) with the baseline TransReID (line 1). The results show that all supervision strategies are effective and could promote the performance consistently and significantly.

Table 2. Performance on the baseline and different mask supervision strategies.

	Occ-Duke			
Setting	mAP	R1	R5	R10
TransReID	55.7	64.2	-	-
non-person supervision	57.3	65.2	79.0	84.2
occlusion supervision	57.5	64.3	79.2	84.5
all-drop supervision	57.2	64.9	78.7	83.8
all-keep supervision	57.2	65.1	78.8	83.6

Table 3. Performance analysis of different fusion approaches. Comparing with the baseline TransReID, all fusion approaches would effectively promote performance. Best performance would be achieved once token-level fusion is adopted.

	Occ-duke			
Fusion Approaches	mAP	R1	R5	R10
TransReID	55.7	64.2	-	-
input-level	56.7	64.4	79.0	83.0
token-level	57.3	65.2	79.0	84.2
semantic-level	57.3	65.0	78.7	83.8

Ablation Study of Fusion Approaches We discuss various approaches for integrating the MIR feature into the FEM, including input-level, token-level, and semantic-level integration. The input-level integration denotes the output of the MIR Module is directly used as the input of FEM. The token-level integration means that the MIR feature is regarded as a new embedding and added to input sequence embeddings. And the semantic-level integration refers to adding the MIR-feature to the output of FEM. As shown in Table 3, we claim that the token-level fusion method is better. The mAP/R1 of the method achieves 57.3%/65.2%, higher than the input-level one (56.7%/64.4%) and the semantic-level one (57.3%/65.0%). Since the MIR-feature is used for reconstructing the occluded pedestrian, it primarily reflects the contour, edge, shape features, and other low-level representations of the image. So it's reasonable that fusing the MIR feature and the original information at the token level is more suitable. In this way, on the one hand, the MIR-feature can provide more valuable clues for low-level information filtering out the occlusion, and on the other hand, it can make full use of the high-level semantic representational capacity of the transformer-based model.

4.5 Visualization Results

The retrieval results of TransReID [3] and our proposed UBN on Occlude-REID dataset are illustrated in Figure 3. It is obvious our UBN achieves significantly better retrieval performance compared with the TransReID, especially when serious occlusion appears.

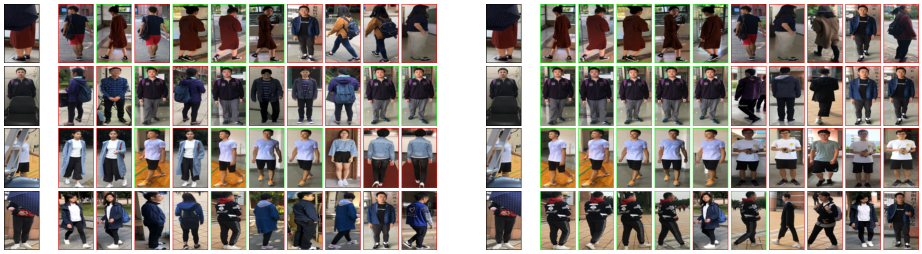

Fig. 3. Retrieval comparison of TransReID and our UBN. The images on the left are from the Occlude-REID query dataset. The images in green and red boxes indicate the correctly and wrongly retrieved instances respectively. The retrieval results are arranged from left to right in descending order of matching scores.

5 Conclusion

In this paper, we present Uncover the Body Network (UBN), a MIM-based occluded person re-identification network. UBN is inspired by recent MIM models that can recover the whole image with partly known patches. Instead of randomly masking the images, UBN uses a Mask Prediction Module (MPM) to readily locate the occluded patches and then applied them to an occlusion-guided Masked Image Reconstruction Module (MIR) to reconstruct the holistic person images. Experimental results on both the occluded and holistic ReID benchmarks have demonstrated the superiority of UBN over the state-of-the-art approaches.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (61925201, 62132001).

References

1. Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884. Springer, 2016.
2. Shiqin Wang, Xin Xu, Lei Liu, and Jing Tian. Multi-level feature fusion model-based real-time person re-identification for forensics. *Journal of Real-Time Image Processing*, 17(1):73–81, 2020.
3. Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021.
4. Xianghao Zang, Ge Li, Wei Gao, and Xiujun Shu. Learning to disentangle scenes for person re-identification. *IVC*, 116:104330, 2021.
5. Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*, pages 393–402, 2019.
6. Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *TIP*, 28(6):2860–2871, 2019.
7. Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *ICME*, pages 1–6. IEEE, 2018.
8. 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.
9. 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.
10. 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.
11. Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *CVPR*, pages 11744–11752, 2020.
12. Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022.
13. Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3702–3712, 2019.
14. Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, pages 1691–1703. PMLR, 2020.
15. Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015.
16. Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016.
17. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

18. Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, pages 14668–14678, 2022.
19. 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 preprint arXiv:2010.11929*, 2020.
20. Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
21. Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cbnetv2: A composite backbone network architecture for object detection. *arXiv preprint arXiv:2107.00420*, 2021.
22. 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.
23. Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3754–3762, 2017.
24. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.
25. 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.
26. Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, volume 34, pages 13001–13008, 2020.
27. Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *NeurIPS*, 31, 2018.
28. Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *CVPR*, pages 7073–7082, 2018.
29. Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399*, 2018.
30. 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*, pages 8450–8459, 2019.
31. Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *ECCV*, pages 346–363. Springer, 2020.
32. 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.
33. 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.