Occluded Person Re-identification via Saliency-Guided Patch Transfer

Lei Tan[1*], Jiaer Xia[1*], Wenfeng Liu[1], Pingyang Dai[1], Yongjian Wu[2], Liujuan Cao[1†]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, China.
[2]Tencent Youtu Lab, China.
{tanlei, xiajiaer, wenfengliu}@stu.xmu.edu.cn, {pydai, caoliujuan}@xmu.edu.cn, littlekenwu@tencent.com

Abstract

While generic person re-identification has made remarkable improvement in recent years, these methods are designed under the assumption that the entire body of the person is available. This assumption brings about a significant performance degradation when suffering from occlusion caused by various obstacles in real-world applications. To address this issue, data-driven strategies have emerged to enhance the model's robustness to occlusion. Following the random erasing paradigm, these strategies typically employ randomly generated noise to supersede randomly selected image regions to simulate obstacles. However, the random strategy is not sensitive to location and content, meaning they cannot mimic real-world occlusion cases in application scenarios. To overcome this limitation and fully exploit the real scene information in datasets, this paper proposes a more intuitive and effective data-driven strategy named Saliency-Guided Patch Transfer (SPT). Combined with the vision transformer, SPT divides person instances and background obstacles using salient patch selection. By transferring person instances to different background obstacles, SPT can easily generate photo-realistic occluded samples. Furthermore, we propose an occlusion-aware Intersection over Union (OIoU) with mask-rolling to filter the more suitable combination and a class-ignoring strategy to achieve more stable processing. Extensive experimental evaluations conducted on occluded and holistic person re-identification benchmarks demonstrate that SPT provides a significant performance gain among different ViT-based ReID algorithms on occluded ReID.

Introduction

Person re-identification (ReID) aims to match individuals across non-overlapping camera views, and it has extensive applications in security and surveillance systems. In recent years, impressive advances have been made in this area, spanning academia and industry (Ye et al. 2021; Fu et al. 2022; Zhang and Wang 2023; Tan et al. 2023). However, most generic person re-identification methods operate under the assumption that the entire body of a person is visible, which has motivated researchers to explore methods that can effectively integrate information from all body parts (Sun et al. 2018; Zhu et al. 2020). Despite this, the assumption of a fully visible body often neglects the presence of occlusions, which are common in practical applications. Inevitable occlusions caused by different obstacles in real application scenarios make the generic ReID methods suffer significant performance degradation. Consequently, there is a growing need to explore occluded person re-identification to address the corresponding issue.

(a) Occlusion caused by the obstacle (b) Occlusion caused by another person

Figure 1: Examples of Saliency-Guided Patch Transfer on Occluded-Duke. Here, we show two types of occluded samples: occlusion caused by obstacles and occlusion caused by the person. Owing to its sensitivity to content and location, SPT can effectively enhance the robustness of ViT algorithms in occluded person re-identification.

To fulfill the demand for occlusion cases, both model-driven (Qian et al. 2018; Tan et al. 2022; Wang et al. 2022a) and data-driven (Huang et al. 2018; Chen et al. 2021; Wang et al. 2022b) strategies have emerged. The model-driven approach places particular emphasis on the alignment strategy, utilizing body cues provided by additional networks or weakly supervised modules to achieve high performance in occluded ReID. However, the limited availability of occlusion data has prevented these modules from fully unleashing their potential. Meanwhile, data-driven strategies (Huang et al. 2018; Chen et al. 2021; Wang et al. 2022b) have made progress in occluded ReID by constructing images with occlusion to enhance the network's robustness. Following the random erasing paradigm (Zhong et al. 2020), these strate-

*These authors contributed equally.
†Corresponding author

gies typically adopt randomly generated or hand-selected noises to serve as obstacles that obscure randomly selected image regions. While these random-based strategies facilitate the learning of an occlusion-aware feature representation, their lack of sensitivity to content and location makes it difficult for them to imitate real occlusion cases in application scenarios.

In this paper, we aim to explore a more intuitive and realistic data-driven strategy to meet the demand for occlusion cases. However, achieving such a strategy presents two main challenges. The first challenge is how to construct the content of obstacles, while the second challenge is where to place these obstacles in fresh samples. Erasing-based strategies have typically addressed both challenges through random strategies but have shown limited improvement in real-world applications as previously discussed. With regard to the first challenge, it is worth noting that the ReID dataset already contains numerous real-world occlusion cases such as cars or trees. As shown in Fig 1, if we can isolate these occlusion cases from the original image, we can easily transfer them to other person instances. Furthermore, the new sample will not be subject to domain gap issues since the obstacles are sourced from the original training set. Additionally, dividing person instances and occlusions from the images can provide extra knowledge to locate the person instances, thereby facilitating the resolution of the second challenge.

To tackle the challenges of constructing content and placing obstacles in fresh samples, we propose the Saliency-Guided Patch Transfer (SPT), which uses a transformer structure that has shown impressive performance in handling occlusion cases (Naseer et al. 2021; He et al. 2022). The SPT is an online model that can be integrated into networks. It starts with a Salient Patch Selection (SPS) module that divides image samples into two subsets: an identity set and an occlusion set. Drawing inspiration from dynamic networks (Rao et al. 2021; Meng et al. 2022), we use a decision matrix to evaluate patch weights in an end-to-end manner. Once we obtain the identity and occlusion sets, the second challenge is to effectively use these subsets to generate fresh samples. While a random combination strategy may seem reasonable, it can result in poor alignment and produce unsatisfactory results. To overcome this challenge, we use an Occlusion-Aware Intersection over Union (OIoU) with a mask rolling strategy to select the most suitable identity-occlusion pairs during each batch. The OIoU encourages SPS to select a candidate sample that can provide efficient occlusion and has a similar scale to the target sample. Furthermore, since SPT uses a hard mask, it is only sufficient to retain the most important parts of the background, while some identity features (partial body or significant background) remain in the residual patches. Roughly utilizing the general training loss will consider residual information as negative parts and lead to sub-optimal results. To deal with this nuisance, we utilize a class ignoring training strategy for more stable training.

We summarize the main contributions as follows:

- We introduce the Saliency-Guided Patch Transfer (SPT) for occluded person re-identification, which is a fresh online data-driven strategy with more realistic content and more specific location.
- To better exploit the power of SPT, we introduce an Occlusion-Aware IoU with mask rolling to filter the most suitable identity-occlusion combinations and a class-ignoring strategy for more stable training.
- We incorporate SPT into the ViT-based algorithms and show their significant performance improvement on occluded person re-identification benchmarks.

Related Works

With the revolution of deep learning, computer vision is dominated by deep learning strategies (Wu et al. 2021; Zhang et al. 2023; Chen et al. 2023; Peng et al. 2021). The mainstream approaches to solving the occluded ReID can be divided into the model-driven method and the data-driven method. Model-driven approaches focus on developing alignment strategies to avoid misalignment between occluded samples. Zhuo *et al*. (Zhuo et al. 2018) firstly introduce an extra occluded/non-occluded binary classification task to distinguish the occluded images from holistic ones, while Miao *et al*. (Miao et al. 2019) propose the Pose-Guided Feature Alignment (PGFA) method that utilizes fine-grid pose landmarks to separate visible part information from occlusion noise. Wang *et al*. (Wang et al. 2020) address occlusion by using high-order relations and a human-topology graph to pass information from visible to invisible nodes, thereby alleviating the influence of obstacles. Li *et al*. (Li et al. 2021) initially utilize the transformer structure and use prototypes to disentangle the fine-grid body part without the help of an extra network in order to achieve satisfying performance. Meanwhile, data-driven approaches have also advanced occluded ReID by analyzing occlusion cases and proposing data augmentation methods. Huang *et al*. (Huang et al. 2018) propose adversarially occluded samples for data augmentation, while Chen *et al*. (Chen et al. 2021) combined a prior guided hand-crafted occlusion augmentation scheme with an attention mechanism to capture visible body parts precisely. Wang *et al*. (Wang et al. 2022b) further improves the pipeline by proposing the feature erasing and diffusion network, reaching a satisfying performance. However, most of the previous data-driven methods follow the trend of random erasing (Zhong et al. 2020), which ignores the discrepancy in both content and location between the generated and real occlusion images. Therefore, in this work, we attempt to bridge the discrepancy with SPT and address the occlusion issue through the data-driven method in a more realistic way.

Approach

Overall Framework

As the obstacles generated by erasing-based data-driven methods (Zhong et al. 2020; Huang et al. 2018; Chen et al. 2021; Wang et al. 2022b) exhibit significant differences in both content and location when compared to real-world occlusion cases, we propose the Saliency-Guided Patch Transfer (SPT) method to address this issue by fully leveraging real scenarios present in the datasets to generate occlusion samples. The framework of SPT consists of two phases, as illustrated in Figure 2. The first phase involves salient patch

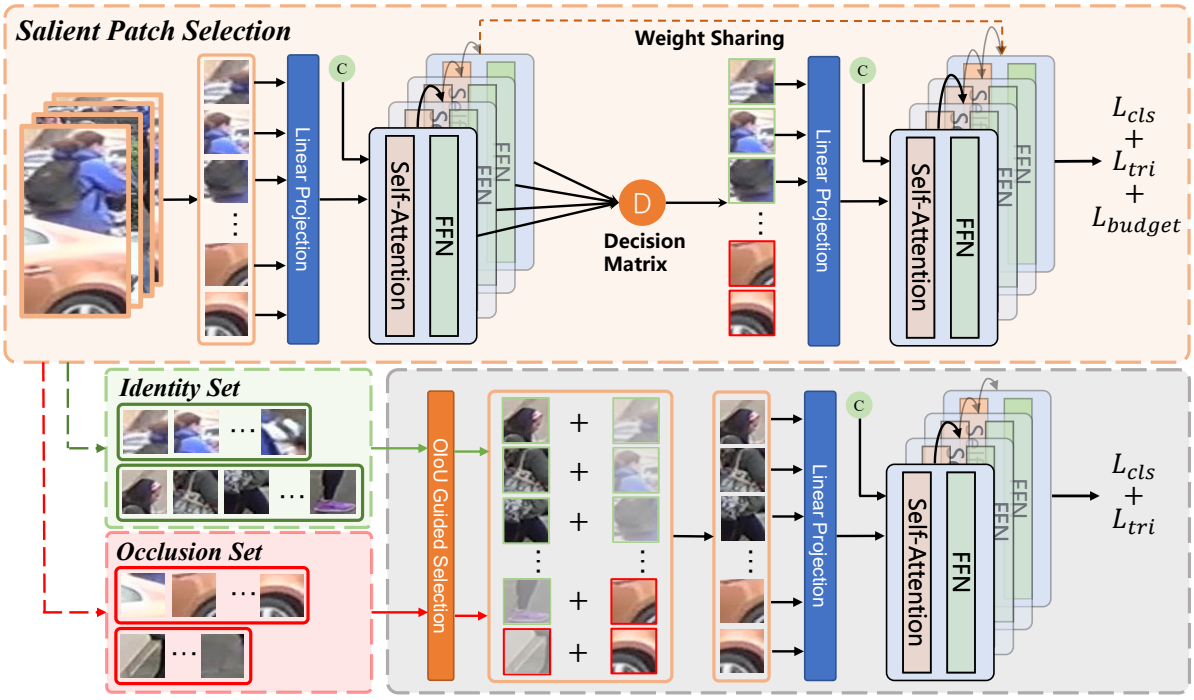

Figure 2: The framework of proposed Saliency-Guided Patch Transfer (SPT). SPT initiates with the Salient Patches Selection (SPS) module, which aims to provide the salient patch mask and segregate the entire image in a batch into two sets, namely Identity Set and Occlusion Set, by utilizing a decision matrix. These two sets are then recombined with a specific probability, guided by OIoU, to generate high-quality samples for training the final model. Herein, the symbol c denotes the class token.

selection (SPS), which draws inspiration from dynamic networks (Rao et al. 2021; Meng et al. 2022) to select the most discriminative patches to activate throughout the network. During this phase, the SPS divides patches into two subsets, namely the identity set containing the target person instances, and the occlusion set consisting of obstacles such as non-identity persons, cars, and so on. In the second phase, the two subsets are reorganized with a certain probability p during training. Salient patches from the target sample in the identity set are then transferred to occlusion patches under the Occlusion-Aware Intersection over Union (OIoU) with a mask rolling strategy to simulate occlusions.

Salient Patch Selection

The primary challenge in transferring occlusion from one image to another is to distinguish between person instances and background occlusions. This enables us to manipulate the instances and synthesize various occlusion scenarios effectively. Drawing inspiration from dynamic networks (Rao et al. 2021; Meng et al. 2022) that can spontaneously select salient patches, we incorporate a salient patch selection module (SPS) into our network. This module encourages the network to remove background patches while preserving the most important patches. It should be noted that SPS operates solely on the training set, eliminating the need to accommodate new data, and thereby improving its efficiency.

The salient patch selection module (SPS) is primarily composed of a decision matrix that employs token representations to generate the final choice via a fully connected layer. It is widely acknowledged within ViT-based structures that token representations from deeper blocks possess a superior ability to estimate their own value (Meng et al. 2022), while the same is challenging for shallow blocks. As a result, during SPS training, we conducted two times of inferences for one-step optimization, as input patches should be determined by the deeper blocks to ensure effective decision-making by the module.

Formally, for a ViT with L blocks, in the first inference, given the input to the l_{th} block without the class token as z_l, the patch mask M is given as:

$$M = \sigma(W^p Concat(z_0, z_1, z_2, \ldots, z_L)). \quad (1)$$

Herein, σ refers to the sigmoid function and W^p denotes the decision matrix which aims to determine the salient patches. For the second inference, we aim to remove most of the noise background. With the patch mask M, input image and class token z^{cls} after patchify layer will be shown as:

$$Z_0 = [z_0^{cls}; M \odot z_0]. \quad (2)$$

Occlusion-Aware IoU

After constructing the identity and occlusion sets, the most intuitive generation strategy is to randomly select one identity sample and one occlusion sample to combine. Although this strategy is simple and efficient, it inevitably causes several bad cases in which the person may go beyond the scope.

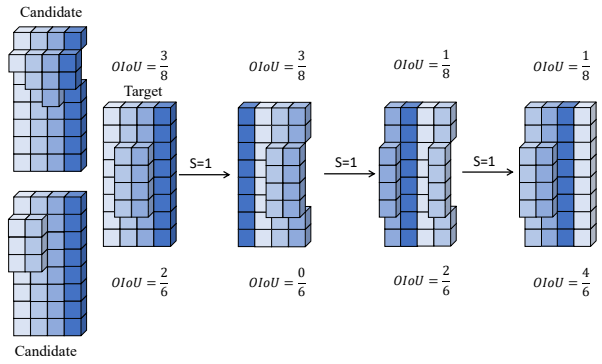

Figure 3: The motivation for Mask Rolling strategy (MR). '**S**' refers to the rolling stride. The MR aims to select combinations with similar person scales for better quality. Therefore, MR horizontally rolls the input instance mask with stride = 1 and calculates the maximum OIoU with the candidate instance mask. To illustrate, we present an example where two candidates show the same OIoU result. It is apparent that the bottom candidate, which is more similar in scale to the target, yields a larger maximum OIoU result after mask rolling.

Intersection over Union (IoU) is the most popular metric for comparing the similarity between two arbitrary shapes in various computer vision tasks. Therefore, we attempt to introduce the Intersection over Union (IoU) in this part. Generally, given two arbitrary shapes from the identity set as input instance mask M_i from input sample I_i and candidate instance mask M_j from candidate sample I_j, the IoU can be attained by:

$$IoU(M_i, M_j) = \frac{M_i \cap M_j}{M_i \cup M_j}. \tag{3}$$

The Intersection over Union (IoU) is commonly used and its application can help in avoiding bad cases in SPT. However, using the standard IoU to guide the combination of the identity set and occlusion set has several drawbacks. Firstly, the original IoU treats the input sample and candidate samples equally. Given similar IoU results, it would be preferable for an input sample I_i to select a candidate sample I_j with a smaller area to fabricate a better-occluded sample. Secondly, the original IoU is not sensitive to differences in scale. Generally, selecting a candidate sample I_j with a similar scale to input sample I_i can avoid mismatch and large-scale differences in person instances and obstacles.

Hence, we propose an Occlusion-Aware IoU (OIoU) to overcome the above drawback and select the most suitable combinations. Firstly, we use the M_j instead of the $M_i \cup M_j$ in the denominator in IoU as:

$$OIoU(M_i, M_j) = \frac{M_i \cap M_j}{M_j}. \tag{4}$$

Compared to the standard IoU, such a replacement only considers the relationship between the $M_i \cap M_j$ and M_j. In this way, we can avoid the SPT selecting a large M_j which cannot bring efficient occlusion on the target sample.

With respect to the second limitation, it is apparent that instances with similar scales can yield significantly higher OIoU results if they are well-aligned, whereas those with substantial differences in scale cannot. To address this, we propose a rolling strategy (MR) in the OIoU calculation to differentiate between samples with significant scale differences. Specifically, as illustrated in Fig 3, the top and bottom candidates show the same OIoU result as the target sample. However, if we horizontally roll the M_i with a small stride and consider the maximum OIoU result between the rolled M_i and its candidates, the bottom candidate, which has a similar scale, demonstrates stronger competitiveness than the upper one.

With OIOU, the SPT processing for the sample M_i can be formulated as:

$$\begin{aligned} &Z_0^i = [z_0^{cls}; M_j \odot z_0^i + (1 - M_j) \odot z_0^j;], \\ &with \quad OIoU(M_i, M_j) \geq \alpha_1 \\ &and \quad Max(OIoU(Roll(M_i), M_j)) \geq \alpha_2, \end{aligned} \tag{5}$$

herein, the α_1 and α_2 denote the lower bound to control the candidate selection.

Loss Function and Optimization

As shown in Figure 2, the SPT processing contains two phases of training. For the Salient Patch Selection (SPS) module, we aim to encourage the network to weigh the importance of each patch. Inspired by the AdaViT(Meng et al. 2022), we utilize an extra budget loss to control the usage of the patches for each batch as:

$$\mathcal{L}_{budget} = \frac{1}{D} \sum_{d=1}^{D} M_i^d - \beta, \tag{6}$$

where D refers to the number of image patches. $M_i^d \in (0, 1)$ refers to the decision result of d_{th} patch in an input instance mask M_i. The β denotes the usage budgets in terms of the percentage patches to reserve.

By combining the budget loss $\mathcal{L}_{budget}$ with the widely used softmax loss $\mathcal{L}_{cls}^{sps}$ and triplet loss $\mathcal{L}_{tri}^{sps}$, we train the SPS end-to-end by minimizing the $\mathcal{L}_{sps}$:

$$\mathcal{L}_{sps} = \mathcal{L}_{cls}^{sps} + \mathcal{L}_{tri}^{sps} + \mathcal{L}_{budget}. \tag{7}$$

Except for the training for SPS, we just employ the softmax loss $\mathcal{L}_{cls}$ and triplet loss $\mathcal{L}_{tri}$ to train the final model as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{tri}. \tag{8}$$

Here, It is worth noting that the target and candidate samples typically usually belong to distinct identities. Due to the limited budget we use, it is only sufficient to retain the important parts of the background samples, while some identity features still remain in the residual parts. When these parts are padded into the foreground samples, they can act as strong noise. Therefore, if the labels of background samples are not ignored, these residues will be considered negative samples, causing the network to ignore the valuable information within them. As a result, this information will also be neglected in the original samples, rendering them unable

to effectively represent. Therefore, during the calculation of $\mathcal{L}_{cls}$ and $\mathcal{L}_{tri}$, we have ignored the class of the candidate sample. More specifically, for the target sample x_i with label i and the candidate sample x_j with lable j, the softmax loss $\mathcal{L}_{cls}$ for newly x_i is given as:

$$\begin{aligned} &\mathcal{L}_{cls}(x_i) = -log\frac{e^{s_i\cos(\theta_i)}}{\sum_{n=1,n\neq j}^{N} e^{s_n\cos(\theta_n)}}, \\ &with \quad s_n = \left\|W_n^T\right\|\|x_n\|, \theta_n = \langle x_i, W_n\rangle, \end{aligned} \tag{9}$$

where the W_n refers to the n_{th} line of the classification matrix W, which can also be considered as the prototype for the n_{th} class. The N refers to the number of classes. Similarly, the $\mathcal{L}_{tri}$ for newly x_i can be calculated as:

$$\begin{aligned} &\mathcal{L}_{tri}(x_i) = (\|x_i - x^{pos}\|^2 - \|x_i - x^{neg}\|^2 + m), \\ &with \quad label(x^{neg}) \neq j. \end{aligned} \tag{10}$$

Here, $\|\cdot\|^2$ indicate the Euclidean distance, the x^{pos} and x^{neg} refer to the hardest positive and negative sample of x_i, and m refers to the extra margin.

Experiment

Datasets and Experimental Setting

Occluded-Duke (Miao et al. 2019) is a large-scale dataset for occluded person re-identification. The training set consists of 15,618 images of 702 persons. The query set consists of 2,210 images of 519 persons and the gallery set consists of 17,661 images of 1,110 persons. **Occluded-REID** (Zhuo et al. 2018) consists of 2000 images from 200 persons captured by mobile cameras. Each person in this dataset has 5 whole-body images and 5 occluded person images. Owing to its limited scale, previous works (Gao et al. 2020) only consider the Occluded-REID as the testing set. Therefore, in this paper, we inherit this setting and employ the Market-1501 (Zheng et al. 2015) as training set. **Market-1501** (Zheng et al. 2015) is a holistic ReID dataset captured from 6 cameras. It includes 12,936 training images of 751 persons as the training set, 3,368 images of 750 persons as the query, and 19,732 images of 750 persons as the gallery. **DukeMTMC-reID** (Zheng, Zheng, and Yang 2017) contains 36,441 images of 1,812 persons captured by eight cameras, in which 16,522 images of 702 identities are used as the training set, 2,228 images and 16,522 images of 702 persons that do not appear in the training set are used as the query and gallery respectively.

Evaluation Protocol. To verify comparison with other methods, we adopt the Cumulative Matching Characteristic (CMC) and mean Average Precision (*m*AP) as evaluation metrics and follow the evaluation settings provided by existing occluded methods (Wang et al. 2020; Gao et al. 2020).

Implementation details. The vanilla ViT model pretrained on the ImageNet(Deng et al. 2009) is selected as the baseline method. Meanwhile, we also conduct experiments on the state-of-the-art ViT-based occluded ReID method DPM (Tan et al. 2022). We resize all input images to 256×128 and employ commonly used data augmentation techniques, including horizontal flipping, padding, and random cropping. During the training phase, we train the SPS in the first 50 epochs and then freeze the SPS to train the network for 120 epochs. Each mini-batch with 64 images, including 16 identities and 4 images per identity. We employ the SGD optimizer and initialize the learning rate as 0.008 with cosine learning rate decay. We set β as 0.3 and 0.5 when training for occluded-Duke and occluded-ReID, respectively. For controlling the minimized OIoU, we set α_1 as 0.5. Regarding α_2, we set it as 0.1 and apply a ranking strategy during each batch. Specifically, only the OIoU of the sample after mask rolling in the top 10% will be considered the final candidate for a specific target sample. All experiments are implemented using PyTorch on a single Nvidia 3090 Ti.

Setting	R-1	R-5	R-10	*m*AP
ViT-Base				
Baseline	60.7	77.0	82.5	53.0
Baseline + SPT (Random)	65.3	80.4	85.4	55.8
Baseline + SPT (OIoU)	66.6	81.6	86.7	56.5
Baseline + SPT (OIoU + MR)	68.6	82.8	87.5	57.4

Table 1: Ablation study of SPT on Occluded-Duke. MR refers to the mask rolling strategy.

Setting	R-1	R-10	R-20	*m*AP
ViT-Base ($\beta = 0.3$)				
ViT-SPT ($p = 0.1, \beta = 0.3$)	67.1	81.4	86.3	56.5
ViT-SPT ($p = 0.2, \beta = 0.3$)	68.6	82.8	87.5	57.4
ViT-SPT ($p = 0.3, \beta = 0.3$)	67.3	82.0	86.8	56.4
ViT-Base ($p = 0.2$)				
ViT-SPT ($p = 0.2, \beta = 0.2$)	65.0	80.8	85.9	55.5
ViT-SPT ($p = 0.2, \beta = 0.3$)	68.6	82.5	87.5	57.4
ViT-SPT ($p = 0.2, \beta = 0.4$)	64.1	79.6	84.6	53.8

Table 2: Performance on Occluded-Duke under different hyper-parameter settings.

Ablation Study

In this subsection, we construct different SPT variants to show the quantitative improvement of each design and show all the results in Table 1. Compared to the vanilla ViT model, adding the SPT by randomly combining two samples from the identity set and occluded set can provide a significant performance. It is natural to ask why the SPT can make sense under a random combination strategy. We deem that it benefits from the patchify strategy in the vision transformer, which largely decreases the image size. Under this condition, even if we employ a random selection strategy, the target and candidate samples can still show competitive IoU. Of course, with the guidance of OIoU, the SPT processing becomes more stable and controllable. After adding the OIoU, the performance has increased in both rank-1 accuracy and *m*AP as 1.3% and 0.7%. Furthermore, the performance gains extra 2.0% and 0.9% via the mask rolling strategy which demonstrates that the scale and shape also matter in the SPT processing.

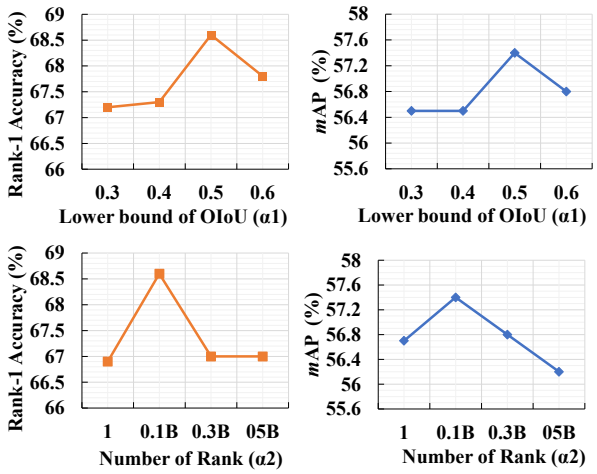

Figure 4: Analysis of parameters α_1 and α_2 in Eq. 5. Herein, the B refers to the bacth size. The optimal performance reaches when $\alpha_1 = 0.5$ and $\alpha_1 = 0.1$.

Method	ViT-Base	ViT-SPT	w/ IOU	w/o class ignore
Rank-1	60.7	68.6	64.8	65.2
*m*AP	53.0	57.4	54.3	54.4

Table 3: Performance on Occluded-Duke under different settings.

Discriminative Evaluations

Comparison between different settings. As shown in Table 3, we compare the performance between using IoU and OIoU under the same setting. While using IoU outperforms vanilla ViT, its results are significantly lower than using OIoU. This discrepancy can be attributed to the pursuit of high IoU values, which tends to encourage the selection of similar candidate samples to the target. On the contrary, OIoU prefers to select candidates with a small area, which largely simplifies the construction of efficient occluded conditions, leading to a larger improvement in performance. Also, we show the necessity of using the class ignoring in the training phase. As we mentioned before, without class ignoring, the residual information will be ignored and lead to sub-optimal results. Quantitatively, compared to the general training strategy, adding class ignoring has improved the performance by 3.4% in rank-1 and 3.0% in *m*AP.

Impact of the hyper-parameters p and β. We introduce a parameter p in SPT to determine whether a given sample should be processed. The results of our experiments, shown in Table 2, indicate that the peak performance is achieved when p is set to 0.2. Generally, increasing p may lead to more occlusion cases in the dataset, which can significantly alter the distribution of the original data. Moreover, we also discuss the hyper-parameter β, which controls the budget of SPS in Eq. 6. Setting a small value of β can help to identify the most discriminative patches in the identity set, but this may also increase the residuals in the occluded set. Conversely, a large value of β can decrease the residuals in the occluded set, but it may also increase the noise in the identity set. Our experiments in Table 2 demonstrate that setting β to 0.3 achieves the best overall performance.

Method	**Occluded-Duke** R-1	*m*AP	**Occluded-REID** R-1	*m*AP
Part Bilinear	36.9	-	-	-
PCB	42.6	33.7	41.3	38.9
FD-GAN	40.8	-	-	-
ISP	62.8	52.3	-	-
TransReID†*	66.4	59.2	-	-
DSR	40.8	30.4	72.8	62.8
PGFA	51.4	37.3	-	-
PVPM+Aug	-	-	70.4	61.2
HOReID	55.1	43.8	80.3	70.2
OAMN	62.6	46.1	-	-
Part-Label	62.2	46.3	81.0	71.0
PAT*	64.5	53.6	81.6	72.1
FED*	68.1	56.4	86.3	79.3
ViT*	60.7	53.0	81.2	76.7
ViT-SPT*	68.6	57.4	86.8	**81.3**
DPM†*	71.4	61.8	85.5	79.7
DPM-SPT†*	**74.7**	**63.0**	**87.8**	81.1

Table 4: Comparison with previous state-of-the-art methods on Occluded-Duke and Occluded-REID. The symbol $*$ denotes the methods that employ the transformer structure. The symbol † represents methods that employ a small stride in a sliding-window setting.

Impact of the hyperparameters α_1 and α_2. In Eq. 5, we set two hyper-parameters α_1 and α_2 to guide the selection of SPS. Therefore, we conduct empirical experiments to measure the performance of the model under different hyper-parameters settings and show the result in Figure 4. The performance linearly increases when α_1 is less than 0.5. After that, continuing to increase α_1, the performance will decrease. It denotes that OIoU is a suitable means to select the proper candidate sample and could improve the effectiveness of SPS. Due to the patchify processing in ViT largely compressed the image size, even under a random selection, the identity-occlusion pairs can show a competitive IoU. Therefore, we can observe that the performance becomes stable when α_1 is less than 0.4. For α_2, we use a ranking strategy. As shown in Figure 4, the best performance is achieved when we rank the OIoU after mask rolling and select the top 10% samples as candidates.

Visualization Analysis. To gain a deeper understanding of SPT, we visualize the target and candidate samples, as well as the output of SPT, according to the geometric relationship between the image and its corresponding patches. As illustrated in Figure 5, it sheds light on how SPT is capable of transferring real-world occlusion scenarios, such as trees, signs, and non-identical individuals, from person to person. This transfer enriches the occluded sample in the dataset and emphasizes the importance of occlusion training. Additionally, as all obstacles stem from the same scenarios captured by the same cameras, the generated samples

Method	Market-1501		DukeMTMC-reID	
	R-1	*m*AP	R-1	*m*AP
FD-GAN	90.5	77.7	80.0	64.5
PCB	92.3	71.4	81.8	66.1
ISP	95.3	88.6	89.6	80.0
CDNet	95.1	86.0	88.6	76.8
TransReID†*	95.2	88.9	90.7	82.0
DSR	83.6	64.3	-	-
Ad-Occluded	86.5	70.4	79.2	62.1
FPR	95.4	86.6	88.6	78.4
PGFA	91.2	76.8	82.6	65.5
HOReID	94.2	84.9	86.9	75.6
OAMN	93.2	79.8	86.3	72.6
PAT∗	95.4	88.0	88.8	78.2
FED∗	95.0	86.3	89.4	78.0
ViT∗	94.3	86.8	88.7	79.3
ViT-SPT*	94.5	86.2	89.4	79.1
DPM†∗	**95.5**	**89.7**	91.0	**82.6**
DPM-SPT†∗	**95.5**	89.4	**91.1**	82.4

Table 5: Comparison with state-of-the-art methods on Market-1501 and DukeMTMC-reID. The symbol ∗ denotes the methods that employ the transformer structure. The symbol † represents methods that employ a small stride in sliding-window setting.

exhibit minimal domain gap with the training/testing sets.

Comparison with State-of-the-art Methods

In this section, we compare the ViT-SPT with previously generic/occluded ReID methods in both occluded and holistic ReID datasets. As the vanilla ViT is not designed to address occluded ReID tasks, we also evaluate the SPT using the occluded ReID method DPM (Tan et al. 2022) to showcase its universality under different model settings.

Results on Occluded Datasets In Table 4, we evaluate the SPT-enhanced ViT and DPM models with previous state-of-the-art holistic (Sun et al. 2018; Suh et al. 2018; Ge et al. 2018; Zhu et al. 2020; He et al. 2021) and occluded (He et al. 2018; Huang et al. 2018; He et al. 2019; Miao et al. 2019; Gao et al. 2020; Wang et al. 2020; Chen et al. 2021; Yang et al. 2021; Wu et al. 2021; Wang et al. 2022b; Tan et al. 2022) ReID methods on the occluded-Duke and occluded-ReID. We observe that simply combining the SPT with the vanilla ViT model yields competitive performance on both occluded-Duke and occluded-ReID datasets. In contrast to FED (Wang et al. 2022b), which also employs a data-driven strategy, ViT-SPT achieves higher performance without any modification to the model. Since the vanilla ViT model is not intended to solve occluded ReID problems, we further combine SPT with DPM to evaluate its performance. Although DPM shows a powerful performance on both occluded-Duke and occluded-ReID datasets, adding SPT also unleashes its potential. Specifically, in occluded-Duke, the performance after adding the SPT increases 3.3% and 1.2% in Rank-1 accuracy and *m*AP. While in the occluded-ReID, the performance increases 2.3% in Rank-1 accuracy and 1.4% in *m*AP. This suggests that SPT is a generalized approach that can easily plug in the ViT-based ReID algorithms to bolster its ability to occluded cases.

Figure 5: Examples of SPT on Occluded-Duke by re-sampling the corresponding patches from the image.

Results on Holistic Datasets While SPT aims to bolster robustness in occluded cases, it is deemed unacceptable for SPT to impair performance under holistic conditions. To this end, we evaluate the performance of ViT-SPT and DPM-SPT on the holistic datasets and compare them with other holistic (Ge et al. 2018; Sun et al. 2018; Zhu et al. 2020; Li, Wu, and Zheng 2021; He et al. 2021) and occluded (He et al. 2018; Huang et al. 2018; He et al. 2019; Miao et al. 2019; Wang et al. 2020; Chen et al. 2021; Wu et al. 2021; Wang et al. 2022b; Tan et al. 2022) approaches in Table 5. From Table 5, it can be observed that compared to the vanilla ViT and DPM, the addition of SPT does not significantly compromise their performance in the holistic conditions. Since holistic ReID datasets rarely include occluded cases and SPT mainly focuses on constructing occluded samples, it exhibits limited ability to improve the performance in the holistic testing scenario. Nonetheless, SPT can be considered an effective data-driven strategy that can greatly improve model robustness against occlusions without degradation in holistic conditions.

Conclusion

In this paper, we propose a novel data-driven method, named Saliency-Guided Patch Transfer (SPT), which leverages real-world scenes in the training set to achieve controllable occlusion construction. SPT divides the sample after patchify into an identity set and an occlusion set through salient patch selection. By recombining these two subsets, SPT can effectively exploit scene information from the dataset and produce high-quality occluded samples. Furthermore, an occlusion-aware Intersection over Union (IoU) with mask rolling and a class-ignoring training strategy is proposed to control SPT's process, ensuring stable and effective patch transfer. Consequently, SPT can be seamlessly integrated into ViT-based algorithms, resulting in significant performance improvements in occluded ReID.

Acknowledgments

This work was supported by the National Key R&D Program of China (No.2022ZD0118202), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305, and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

References

Chen, P.; Liu, W.; Dai, P.; Liu, J.; Ye, Q.; Xu, M.; Chen, Q.; and Ji, R. 2021. Occlude them all: Occlusion-aware attention network for occluded person re-id. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11833–11842.

Chen, Z.; Ding, J.; Cao, L.; Shen, Y.; Zhang, S.; Jiang, G.; and Ji, R. 2023. Category-aware Allocation Transformer for Weakly Supervised Object Localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6643–6652.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the CVPR*, 248–255.

Fu, D.; Chen, D.; Yang, H.; Bao, J.; Yuan, L.; Zhang, L.; Li, H.; Wen, F.; and Chen, D. 2022. Large-Scale Pre-training for Person Re-identification with Noisy Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2476–2486.

Gao, S.; Wang, J.; Lu, H.; and Liu, Z. 2020. Pose-guided visible part matching for occluded person ReID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11744–11752.

Ge, Y.; Li, Z.; Zhao, H.; Yin, G.; Yi, S.; Wang, X.; et al. 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *Advances in neural information processing systems*, 31.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

He, L.; Liang, J.; Li, H.; and Sun, Z. 2018. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7073–7082.

He, L.; Wang, Y.; Liu, W.; Zhao, H.; Sun, Z.; and Feng, J. 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8450–8459.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15013–15022.

Huang, H.; Li, D.; Zhang, Z.; Chen, X.; and Huang, K. 2018. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5098–5107.

Li, H.; Wu, G.; and Zheng, W.-S. 2021. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6729–6738.

Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; and Wu, F. 2021. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2898–2907.

Meng, L.; Li, H.; Chen, B.-C.; Lan, S.; Wu, Z.; Jiang, Y.-G.; and Lim, S.-N. 2022. AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12309–12318.

Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; and Yang, Y. 2019. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 542–551.

Naseer, M. M.; Ranasinghe, K.; Khan, S. H.; Hayat, M.; Shahbaz Khan, F.; and Yang, M.-H. 2021. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34: 23296–23308.

Peng, J.; Zhou, Y.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; and Ji, R. 2021. Knowledge-Driven Generative Adversarial Network for Text-to-Image Synthesis. *IEEE Transactions on Multimedia*.

Qian, X.; Fu, Y.; Xiang, T.; Wang, W.; Qiu, J.; Wu, Y.; Jiang, Y.-G.; and Xue, X. 2018. Pose-normalized image generation for person re-identification. In *Proceedings of the ECCV*, 650–667.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C.-J. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34: 13937–13949.

Suh, Y.; Wang, J.; Tang, S.; Mei, T.; and Lee, K. M. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, 402–419.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the ECCV*, 480–496.

Tan, L.; Dai, P.; Ji, R.; and Wu, Y. 2022. Dynamic Prototype Mask for Occluded Person Re-Identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 531–540.

Tan, L.; Zhang, Y.; Shen, S.; Wang, Y.; Dai, P.; Lin, X.; Wu, Y.; and Ji, R. 2023. Exploring invariant representation for visible-infrared person re-identification. *arXiv preprint arXiv:2302.00884*.

Wang, G.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; and Sun, J. 2020. 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*, 6449–6458.

Wang, T.; Liu, H.; Song, P.; Guo, T.; and Shi, W. 2022a. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2540–2549.

Wang, Z.; Zhu, F.; Tang, S.; Zhao, R.; He, L.; and Song, J. 2022b. Feature Erasing and Diffusion Network for Occluded Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4754–4763.

Wu, Q.; Dai, P.; Chen, J.; Lin, C.-W.; Wu, Y.; Huang, F.; Zhong, B.; and Ji, R. 2021. Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4330–4339.

Yang, J.; Zhang, J.; Yu, F.; Jiang, X.; Zhang, M.; Sun, X.; Chen, Y.-C.; and Zheng, W.-S. 2021. Learning To Know Where To See: A Visibility-Aware Approach for Occluded Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11885–11894.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Y.; and Wang, H. 2023. Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2153–2162.

Zhang, Y.; Yan, Y.; Li, J.; and Wang, H. 2023. MRCN: A Novel Modality Restitution and Compensation Network for Visible-Infrared Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, 1116–1124.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, 3754–3762.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random Erasing Data Augmentation. In *Proceedings of the AAAI*.

Zhu, K.; Guo, H.; Liu, Z.; Tang, M.; and Wang, J. 2020. Identity-guided human semantic parsing for person re-identification. In *European Conference on Computer Vision*, 346–363. Springer.

Zhuo, J.; Chen, Z.; Lai, J.; and Wang, G. 2018. Occluded person re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.