

TransReID: Transformer-based Object Re-Identification

Shuting He^{1,2*}, Hao Luo¹, Pichao Wang¹, Fan Wang¹, Hao Li¹, Wei Jiang²

¹Alibaba Group, ²Zhejiang University

{shuting_he, jiangwei_zju}@zju.edu.cn {michuan.lh, pichao.wang, fan.w, lihao.lh}@alibaba-inc.com

Abstract

Extracting robust feature representation is one of the key challenges in object re-identification (ReID). Although convolution neural network (CNN)-based methods have achieved great success, they only process one local neighborhood at a time and suffer from information loss on details caused by convolution and downsampling operators (e.g. pooling and strided convolution). To overcome these limitations, we propose a pure transformer-based object ReID framework named TransReID. Specifically, we first encode an image as a sequence of patches and build a transformer-based strong baseline with a few critical improvements, which achieves competitive results on several ReID benchmarks with CNN-based methods. To further enhance the robust feature learning in the context of transformers, two novel modules are carefully designed. (i) The jigsaw patch module (JPM) is proposed to rearrange the patch embeddings via shift and patch shuffle operations which generates robust features with improved discrimination ability and more diversified coverage. (ii) The side information embeddings (SIE) is introduced to mitigate feature bias towards camera/view variations by plugging in learnable embeddings to incorporate these non-visual clues. To the best of our knowledge, this is the first work to adopt a pure transformer for ReID research. Experimental results of TransReID are superior promising, which achieve state-of-the-art performance on both person and vehicle ReID benchmarks. Code is available at <https://github.com/heshuting555/TransReID>.

1. Introduction

Object re-identification (ReID) aims to associate a particular object across different scenes and camera views, such as in the applications of person ReID and vehicle ReID. Extracting robust and discriminative features is a crucial component of ReID, and has been dominated by CNN-based methods for a long time [19, 37, 36, 44, 42].

*This work was done when Shuting He was intern at Alibaba supervised by Hao Luo and Pichao Wang.

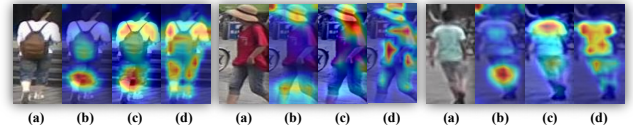


Figure 1: Grad-CAM [34] visualization of attention maps: (a) Original images, (b) CNN-based methods, (c) CNN+attention methods, (d) Transformer-based methods which captures global context information and more discriminative parts.



Figure 2: Visualization of output feature maps for 2 hard samples with similar appearances. Transformer-based methods retain backpack details on output feature maps in contrast to CNN-based methods, as noted in red boxes. For better visualization, input images are scaled to size 1024×512 .

By reviewing CNN-based methods, we find two important issues which are not well addressed in the field of object ReID. (1) Exploiting the rich structural patterns in a global scope is crucial for object ReID [54]. However, CNN-based methods mainly focus on small discriminative regions due to a Gaussian distribution of receptive fields [29]. Recently, attention modules [54, 6, 4, 48, 21, 2] have been introduced to explore long-range dependencies [45], but most of them are embedded in the deep layers and do not solve the principle problem of CNN. Thus, attention-based methods still prefer large continuous areas and are hard to extract multiple diversified discriminative parts (see Figure 1). (2) Fine-grained features with detail information are also important. However, the downsampling operators (e.g. pooling and strided convolution) of CNN reduce spatial resolution of output feature maps, which greatly affect the discrimination ability to distinguish objects with similar appearances [37, 27]. As shown in Figure 2, the details of the backpack are lost in CNN-based feature maps, making it difficult to differentiate the two people.

Recently, Vision Transformer (ViT) [8] and Data-

efficient image Transformers (DeiT) [40] have shown that pure transformers can be as effective as CNN-based methods on feature extraction for image recognition. With the introduction of multi-head attention modules and the removal of convolution and downsampling operators, transformer-based models are suitable to solve the aforementioned problems in CNN-based ReID for the following reasons. (1) The multi-head self-attention captures long range dependencies and drives the model to attend diverse human-body parts than CNN models (e.g. thighs, shoulders, waist in Figure 1). (2) Without downsampling operators, transformer can keep more detailed information. For example, one can observe that the difference on feature maps around backpacks (marked by red boxes in Figure 2) can help the model easily differentiate the two people. These advantages motivate us to introduce pure transformers in the object ReID.

Despite its great advantages as discussed above, transformers still need to be designed specifically for object ReID to tackle the unique challenges, such as the **large variations** (e.g. occlusions, diversity of poses, camera perspective) in images. Substantial efforts have been devoted to alleviating this challenge in CNN-based methods. Among them, local part features [37, 44, 20, 49, 28] and side information (such as cameras and viewpoints) [7, 61, 35, 30], have been proven to be essential and effective to enhance the feature robustness. Learning **part/stripe aggregated features** makes it robust against occlusions and misalignments [50]. However, extending the rigid stripe part methods from CNN-based methods to pure transformer-based methods may damage long-range dependencies due to global sequences splitting into several isolated subsequences. In addition, taking **side information** into consideration, such as camera and viewpoint-specific information, an invariant feature space can be constructed to diminish bias brought by side information variations. However, the complex designs for side information built on CNN, if directly applied to transformers, cannot make full use of the inherent encoding capabilities of transformers. As a result, specific designed modules are inevitable and essential for a pure transformer to successfully handle these challenges.

Therefore, we propose a new object ReID framework dubbed TransReID to learn robust feature representations. Firstly, by making several critical adaptations, we construct a strong baseline framework based on a pure transformer.

Secondly, in order to expand long-range dependencies and enhance feature robustness, we propose a **jigsaw patches module** (JPM) by rearranging the patch embeddings via shift and shuffle operations and re-grouping them for further feature learning. The JPM is employed on the last layer of the model to extract robust features in parallel with the global branch which does not include this special

operation. Hence, the network tends to extract perturbation-invariant and robust features with global context. Thirdly, to further enhance the learning of robust features, a **side information embedding** (SIE) is introduced. Instead of the special and complex designs in CNN-based methods for utilizing these non-visual clues, we propose a unified framework that effectively incorporates non-visual clues through learnable embeddings to alleviate the data bias brought by cameras or viewpoints. Taking cameras for example, the proposed SIE helps address the vast pairwise similarity discrepancy between inter-camera and intra-camera matching (see Figure 6). SIE can also be easily extended to include any non-visual clues other than the ones we have demonstrated.

To our best knowledge, we are the first to investigate the application of pure transformers in the field of object ReID. The contributions of the paper are summarised:

- We propose a strong baseline that exploits the pure transformer for ReID tasks for the first time and achieve comparable performance with CNN-based frameworks.
- We design a *jigsaw patches module* (JPM), consisting of shift and patch shuffle operation, which facilitates perturbation-invariant and robust feature representation of objects.
- We introduce a *side information embeddings* (SIE) that encodes side information by learnable embeddings, and is shown to effectively mitigate the bias of learned features.
- The final framework TransReID achieves state-of-the-art performance on both person and vehicle ReID benchmarks including MSMT17[46], Market-1501[55], DukeMTMC-reID[33], Occluded-Duke[31], VeRi-776[24] and VehicleID[23].

2. Related Work

2.1. Object ReID

The studies of object ReID have been mainly focused on person ReID and vehicle ReID, with most state-of-the-art methods based on the CNN structure. A popular pipeline for object ReID is to design suitable loss functions to train a CNN backbone (e.g. ResNet [14]), which is used to extract features of images. The cross-entropy loss (ID loss) [56] and triplet loss [22] are most widely used in the deep ReID. Luo *et al.* [27] proposed the BNNeck to better combine ID loss and triplet loss. Sun *et al.* [36] proposed a unified perspective for ID loss and triplet loss.

Fine-grained Features. Fine-grained features have been learned to aggregate information from different part/region. The fine-grained parts are either automatically generated by roughly horizontal stripes or by semantic parsing. Methods like PCB [37], MGN [44], AlignedReID++ [28], SAN

[32], *etc.*, divide an image into several stripes and extract local features for each stripe. Using parsing or keypoint estimation to align different parts or two objects has also been proven effective for both person and vehicle ReID [25, 30, 47, 31].

Side Information. For images captured in a cross-camera system, large variations exist in terms of pose, orientation, illumination, resolution, *etc.* caused by different camera setup and object viewpoints. Some works [61, 7] use side information such as camera ID or viewpoint information to learn invariant features. For example, Camera-based Batch Normalization (CBN) [61] forces the image data from different cameras to be projected onto the same subspace, so that the distribution gap between inter- and intra- camera pairs is largely diminished. Viewpoint/Orientation-invariant feature learning [7, 60] is also important for both person and vehicle ReID.

2.2. Pure Transformer in Vision

The Transformer model is proposed in [41] to handle sequential data in the field of natural language processing (NLP). Many studies also show its effectiveness for computer-vision tasks. Han *et al.* [11] and Salman *et al.* [18] have surveyed the application of the Transformer in the field of computer vision.

Pure Transformer models are becoming more and more popular. For example, Image Processing Transformer (IPT) [3] takes advantage of transformers by using large scale pre-training and achieves the state-of-the-art performance on several image processing tasks like super-resolution, denoising and de-raining. ViT [8] is proposed recently which applies a pure transformer directly to sequences of image patches. However, ViT requires a large-scale dataset to pretrain the model. To overcome this shortcoming, Touvron *et al.* [40] propose a framework called DeiT which introduces a teacher-student strategy specific for transformers to speed up ViT training without the requirement of large-scale pretraining data.

3. Methodology

Our object ReID framework is based on transformer-based image classification, but with several critical improvements to capture robust feature (Sec. 3.1). To further boost the robust feature learning in the context of transformer, a jigsaw patch module (JPM) and a side information embeddings (SIE) are carefully devised in Sec. 3.2 and Sec. 3.3. The two modules are jointly trained in an end-to-end manner and shown in Figure 4.

3.1. Transformer-based strong baseline

We build a transformer-based strong baseline for object ReID, following the general strong pipeline for object ReID [27, 44]. Our method has two main stages, *i.e.*, feature

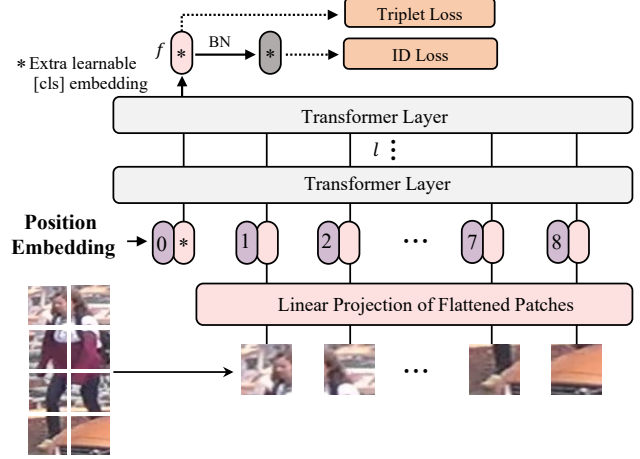


Figure 3: Transformer-based strong baseline framework (a non-overlapping partition is shown). Output [cls] token marked with * is served as the global feature f . Inspired by [27], we introduce the BNNeck after the f .

extraction and supervision learning. As shown in Figure 3. Given an image $x \in \mathbb{R}^{H \times W \times C}$, where H , W , C denote its height, width, and number of channels, respectively, we split it into N fixed-sized patches $\{x_p^i | i = 1, 2, \dots, N\}$. An extra learnable [cls] embedding token denoted as x_{cls} is prepended to the input sequences. The output [cls] token serves as a global feature representation f . Spatial information is incorporated by adding **learnable position embeddings**. Then, the input sequences fed into transformer layers can be expressed as:

$$\mathcal{Z}_0 = [x_{cls}; \mathcal{F}(x_p^1); \mathcal{F}(x_p^2); \dots; \mathcal{F}(x_p^N)] + \mathcal{P}, \quad (1)$$

where \mathcal{Z}_0 represents input sequence embeddings and $\mathcal{P} \in \mathbb{R}^{(N+1) \times D}$ is position embeddings. \mathcal{F} is a linear projection mapping the patches to D dimensions. Moreover, l transformer layers are employed to learn feature representations. The limited receptive field problem of CNN-based methods is addressed, because all transformer layers have a global receptive field. There are also no downsampling operations, so the detailed information is preserved.

Overlapping Patches. Pure transformer-based models (*e.g.* ViT, DeiT) split the images into non-overlapping patches, losing local neighboring structures around the patches. Instead, we use a **sliding window** to generate patches with overlapping pixels. Denoting the step size as S , size of the patch as P (*e.g.* 16), then the shape of the area where two adjacent patches overlap is $(P - S) \times P$. An input image with a resolution $H \times W$ will be split into N patches.

$$N = N_H \times N_W = \lfloor \frac{H + S - P}{S} \rfloor \times \lfloor \frac{W + S - P}{S} \rfloor \quad (2)$$

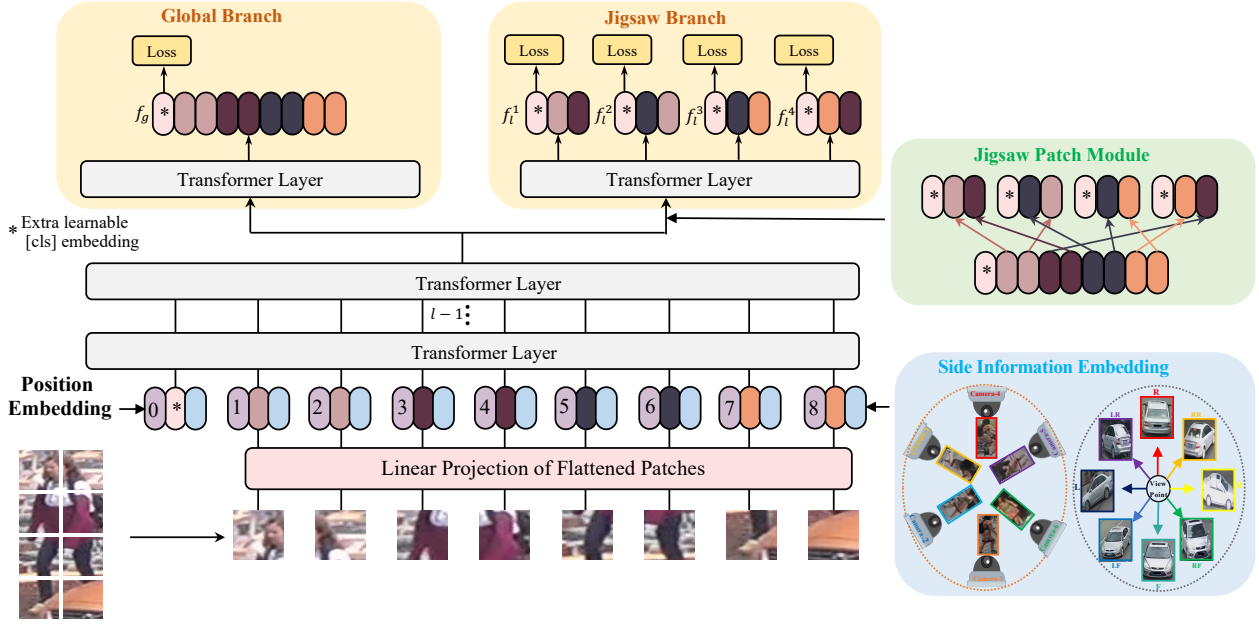


Figure 4: Framework of proposed TransReID. Side Information Embedding (light blue) encodes non-visual information such as camera or viewpoint into embedding representations. It is input into transformer encoder together with patch embedding and position embedding. Last layer includes two independent transformer layers. One is standard to encode global feature. The other contains the Jigsaw Patch Module (JPM) which shuffles all patches and regroups them into several groups. All these groups are input into a shared transformer layer to learn local features. Both global feature and local features contribute to ReID loss.

where $\lfloor \cdot \rfloor$ is the floor function and S is set smaller than P . N_H and N_W represent the numbers of splitting patches in height and width, respectively. The smaller S is, the more patches the image will be split into. Intuitively, more patches usually bring better performance with the cost of more computations.

Position Embeddings. As the image resolution for ReID tasks may be different from the original one in image classification, the position embedding pretrained on ImageNet cannot be directly loaded here. Therefore, a **bilinear 2D interpolation** is introduced to help handle any given input resolution. Similar to ViT, the position embedding is also learnable.

Supervision Learning. We optimize the network by constructing ID loss and triplet loss for global features. The ID loss \mathcal{L}_{ID} is the cross-entropy loss without label smoothing. For a triplet set $\{a, p, n\}$, the triplet loss \mathcal{L}_T with soft-margin is shown as follows:

$$\mathcal{L}_T = \log \left[1 + \exp \left(\|f_a - f_p\|_2^2 - \|f_a - f_n\|_2^2 \right) \right] \quad (3)$$

3.2. Jigsaw Patch Module

Although transformer-based strong baseline can achieve impressive performance in object ReID, it utilizes information from the entire image for object ReID. However, due to challenges like occlusions and misalignments, we may only have partial observation of an object. Learning fine-grained local features such as striped

features has been widely used for CNN-based methods to tackle these challenges.

Suppose the hidden features input to the last layer are denoted as $\mathcal{Z}_{l-1} = [z_{l-1}^0; z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^N]$. To learn fine-grained local features, a straightforward solution is splitting $[z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^N]$ into k groups in order which concatenate the shared token z_{l-1}^0 and then feed k feature groups into a shared transformer layer to learn k local features denoted as $\{f_l^j | j = 1, 2, \dots, k\}$ and f_l^j is the output token of j -th group. But it may not take full advantage of global dependencies for the transformer because each local segment only considers a part of the continuous patch embeddings.

To address the aforementioned issues, we propose a jigsaw patch module (JPM) to shuffle the patch embeddings and then re-group them into different parts, each of which contains several random patch embeddings of an entire image. In addition, extra perturbation introduced in training also helps improve the robustness of object ReID model. Inspired by ShuffleNet [53], the patch embeddings are shuffled via a shift operation and a patch shuffle operation. The sequences embeddings \mathcal{Z}_{l-1} are shuffled as follow:

- **Step1: The shift operation.** The first m patches (except for [cls] token) are moved to the end, i.e. $[z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^N]$ is shifted in m steps to become $[z_{l-1}^{m+1}, z_{l-1}^{m+2}, \dots, z_{l-1}^N, z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^m]$.
- **Step2: The patch shuffle operation.** The shifted

patches are further shuffled by the patch shuffle operation with k groups. The hidden features become $[z_{l-1}^{x_1}, z_{l-1}^{x_2}, \dots, z_{l-1}^{x_N}]$, $x_i \in [1, N]$.

With the shift and shuffle operation, the local feature f_l^j can cover patches from different body or vehicle parts which means that the local features hold global discriminative capability.

As shown in Figure 4, paralleling with the jigsaw patch, another global branch which is a standard transformer encodes Z_{l-1} into $Z_l = [f_g; z_l^1, z_l^2, \dots, z_l^N]$, where f_g is served as the global feature of CNN-based methods. Finally, the global feature f_g and k local features are trained with \mathcal{L}_{ID} and \mathcal{L}_T . The overall loss is computed as follow:

$$\mathcal{L} = \mathcal{L}_{ID}(f_g) + \mathcal{L}_T(f_g) + \frac{1}{k} \sum_{j=1}^k (\mathcal{L}_{ID}(f_l^j) + \mathcal{L}_T(f_l^j)) \quad (4)$$

During inference, we concatenate the global feature and local features $[f_g, f_l^1, f_l^2, \dots, f_l^k]$ as the final feature representation. Using f_g only is a variation with lower computational cost and slight performance degradation.

3.3. Side Information Embeddings

After obtaining fine-grained feature representations, features are still susceptible to camera or viewpoint variations. In other words, the trained model may easily fail to distinguish the same object from different perspectives due to scene-bias. Therefore, we propose a Side Information Embedding (SIE) to ~~incorporate the non-visual information~~, such as cameras or viewpoints, into embedding representations to learn invariant features.

Inspired by position embeddings which encode positional information adopting learnable embeddings, we plug **learnable 1-D embeddings** to retain side information. Particularly, as illustrated in Figure 4, SIE is inserted into the transformer encoder together with patch embeddings and position embeddings. In specific, suppose there are N_C camera IDs in total, we initialize learnable side information embeddings as $\mathcal{S}_C \in \mathbb{R}^{N_C \times D}$. If camera ID of an image is r , then its camera embeddings can be denoted as $\mathcal{S}_C[r]$. Different from the position embeddings which vary between patches, camera embeddings $\mathcal{S}_C[r]$ are the same for all patches of an image. In addition, if viewpoint of the object is available, either by a viewpoint estimation algorithm or human annotations, we can also encode the viewpoint label q as $\mathcal{S}_V[q]$ for all patches of an image where $\mathcal{S}_V \in \mathbb{R}^{N_V \times D}$ and N_V represents the number of viewpoint IDs.

Now comes the problem about how to integrate two different types of information. A trivial solution might be directly adding the two embeddings together like $\mathcal{S}_C[r] + \mathcal{S}_V[q]$. However, it might make the two embeddings counteract each other due to redundant or adversarial

information. We propose to **encode the camera and viewpoint jointly** as $\mathcal{S}_{(C,V)} \in \mathbb{R}^{(N_C \times N_V) \times D}$.

Finally, the input sequences with camera ID r and viewpoint ID q are fed into transformer layers as follows:

$$\mathcal{Z}'_0 = \mathcal{Z}_0 + \lambda \mathcal{S}_{(C,V)}[r * N_k + q], \quad (5)$$

where \mathcal{Z}_0 is the raw input sequences in Eq. 2 and λ is a hyperparameter to balance the weight of SIE. As the position embeddings are different for each patch but the same across different images, and $\mathcal{S}_{(C,V)}$ are the same for each patch but may have different values for different images. Transformer layers are able to encode embeddings with different distribution properties which can then be added directly.

Here we have only demonstrate the usage of SIE with camera and viewpoint information which are both categorical variables. In practice, SIE can be further extended to encode more kinds of information, including both categorical and numerical variables. In our experiments on different benchmarks, camera and viewpoint information is included wherever available.

4. Experiments

4.1. Datasets

We evaluate our proposed method on four person ReID datasets, Market-1501 [55], DukeMTMC-reID [33], MSMT17 [46], Occluded-Duke [31], and two vehicle ReID datasets, VeRi-776 [24] and VehicleID [23]. It is noted that, unlike other datasets, images in Occluded-Duke are selected from DukeMTMC-reID and the training/query/gallery set contains 9%/ 100%/ 10% occluded images respectively. All datasets except VehicleID provide camera ID for each image, while only VeRi-776 and VehicleID dataset provide viewpoint labels for each image. The details of these datasets are summarized in Table 1.

Dataset	Object	#ID	#image	#cam	#view
MSMT17	Person	4,101	126,441	15	-
Market-1501	Person	1,501	32,668	6	-
DukeMTMC-reID	Person	1,404	36,441	8	-
Occluded-Duke	Person	1,404	36,441	8	-
VeRi-776	Vehicle	776	49,357	20	8
VehicleID	Vehicle	26,328	221,567	-	2

Table 1: Statistics of datasets used in the paper.

4.2. Implementation

Unless otherwise specified, all person images are resized to 256×128 and all vehicle images are resized to 256×256 . The training images are augmented with random horizontal flipping, padding, random cropping and random erasing [57]. The batch size is set to 64 with 4 images per ID. SGD optimizer is employed with a momentum of 0.9 and

Backbone	Inference Time	MSMT17		VeRi-776	
		mAP	R1	mAP	R1
ResNet50	1x	51.3	75.3	76.4	95.2
ResNet101	1.48x	53.8	77.0	76.9	95.2
ResNet152	1.96x	55.6	78.4	77.1	95.9
ResNeSt50	1.86x	61.2	82.0	77.6	96.2
ResNeSt200	3.12x	63.5	83.5	77.9	96.4
DeiT-S/16	0.97x	55.2	76.3	76.3	95.5
DeiT-B/16	1.79x	61.4	81.9	78.4	95.9
ViT-B/16	1.79x	61.0	81.8	78.2	96.5
ViT-B/16 _{s=14}	2.14x	63.7	82.7	78.6	96.4
ViT-B/16 _{s=12}	2.81x	64.4	83.5	79.0	96.5

Table 2: Comparison of different backbones. Inference time is represented by comparing each model to ResNet50 as only relative comparison is necessary. All the experiments were carried out on the same machine for fair comparison. **ViT-B/16 is regarded as the baseline model and abbreviated as Baseline in the rest of this paper.**

the weight decay of $1e-4$. The learning rate is initialized as 0.008 with cosine learning rate decay. Unless otherwise specified, we set $m = 5, k = 4$ and $m = 8, k = 4$ for person and vehicle ReID datasets, respectively.

All the experiments are performed with one Nvidia Tesla V100 GPU using the PyTorch toolbox¹ with FP16 training. The initial weights of ViT are pre-trained on ImageNet-21K and then finetuned on ImageNet-1K, while the initial weights of DeiT are trained only on ImageNet-1K.

Evaluation Protocols. Following conventions in the ReID community, we evaluate all methods with Cumulative Matching Characteristic (CMC) curves and the mean Average Precision (mAP).

4.3. Results of Transform-based Baseline

In this section, we compare CNN-based and transformer-based backbones in Table 2. To show the trade-off between computation and performance, several different backbones are chosen. DeiT-small, DeiT-Base, ViT-Base denoted as DeiT-S, DeiT-B, ViT-B, respectively. ViT-B/16_{s=14} means ViT-Base with patch size 16 and step size $S = 14$ in overlapping patches setting. For a comprehensive comparison, inference time consumption of each backbone is included as well.

We can observe a large gap in model capacity between the ResNet series and DeiT/ViT. DeiT-S/16 is a little bit better in performance and speed compared to ResNet50. DeiT-B/16 and ViT-B/16 achieve similar performance with ResNeSt50 [51] backbone, with less inference time than ResNeSt50 (1.79x vs 1.86x). When we reduce the step size of the sliding window S , the performance of the Baseline can be improved while the inference time is also increasing. ViT-B/16_{s=12} is faster than ResNeSt200 (2.81x vs 3.12x)

¹<http://pytorch.org>

Backbone	#groups	MSMT17		VeRi-776	
		mAP	R1	mAP	R1
Baseline	-	61.0	81.8	78.2	96.5
+JPM	1	62.9	82.5	78.6	97.0
+JPM	2	62.8	82.1	79.1	96.4
+JPM	4	63.6	82.5	79.2	96.8
+JPM w/o rearrange	4	63.1	82.4	79.0	96.7
+JPM w/o local	4	63.5	82.5	79.1	96.6

Table 3: The ablation study of jigsaw patch module. ‘w/o rearrange’ means the patch features are split into parts without rearrange including shift and shuffle operation. ‘w/o local’ means we evaluate the global feature without concatenating local features.

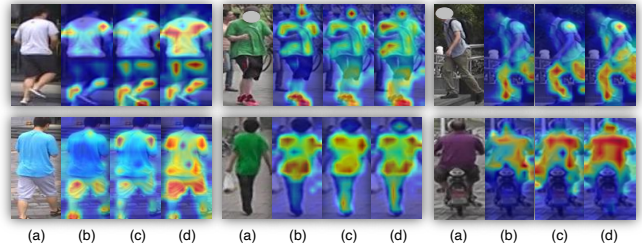


Figure 5: Grad-CAM visualization of attention maps. (a) Input images, (b) Baseline, (c) JPM w/o rearrange, (d) JPM.

and performs slightly better than ResNeSt200 on ReID benchmarks. Therefore, ViT-B/16_{s=12} achieves better speed-accuracy trade-off than ResNeSt200. In addition, we believe that DeiT/ViT still have lots of room for improvement in terms of computational efficiency.

4.4. Ablation Study of JPM

The effectiveness of the proposed JPM module is validated in Table 3. JPM provides +2.6% mAP and +1.0% mAP improvements compared to baseline on MSMT17 and VeRi-776, respectively. Increasing the number of groups k can improve the performance while slightly increasing inference time. In our experiment, $k = 4$ is a choice to trade off speed and performance. Comparing JPM and JPM w/o rearrange, we can observe that the shift and shuffle operation helps the model learn more discriminative features with +0.5% mAP and +0.2% mAP improvements on MSMT17 and VeRi-776, respectively. It is also observed that, if only the global feature f_g is used in inference stage (still trained with full JPM), the performance (denoted as ‘w/o local’) is nearly comparable with the version of full set of features, which suggests us to only use the global feature as an efficient variation with lower storage cost and computational cost in the inference stage. The attention maps visualized in Figure 5 show that JPM with the rearrange operation can help the model learn more global context information and more discriminative parts, which makes the model more robust to perturbations.

Method	Camera	Viewpoint	MSMT17		VeRi-776	
			mAP	R1	mAP	R1
Baseline			61.0	81.8	78.2	96.5
+ $S_C[r]$	✓		62.4	81.9	78.7	97.1
+ $S_V[q]$		✓	-	-	78.5	96.9
+ $S_{(C,V)}$	✓	✓	-	-	79.6	96.9

Table 4: Ablation study of SIE. Since the person ReID datasets do not provide viewpoint annotations, viewpoint information can only be encoded in VeRi-776.

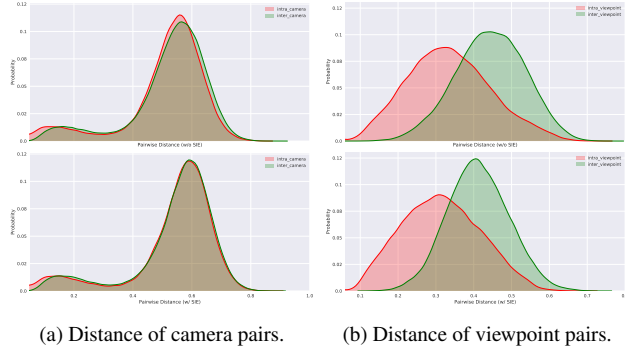


Figure 6: We visualize the distance distributions of different camera pairs and viewpoint pairs on VeRi-776. (a) inter-camera and intra-camera distance distribution. (b) inter-viewpoint and intra-viewpoint distance distribution.

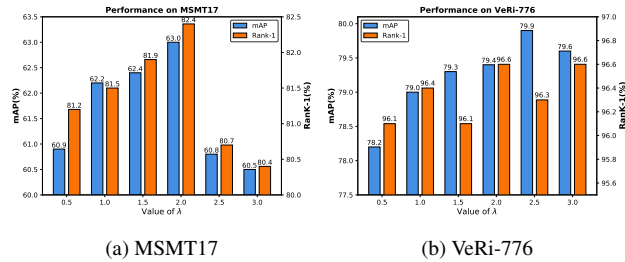


Figure 7: Impact of the hyper-parameter λ .

4.5. Ablation Study of SIE

Performance Analysis. In Table 4, we evaluate the effectiveness of the SIE on MSMT17 and VeRi-776. MSMT17 does not provide viewpoint annotations, so the results of SIE which only encode camera information are shown for MSMT17. VeRi-776 not only have a camera ID of each image, but is also annotated with 8 different viewpoints according to vehicle orientation. Therefore, the results are shown with SIE encoding various combinations of camera ID and/or viewpoints information.

When SIE encodes only the camera IDs of images, the model gains 1.4% mAP and 0.1% rank-1 accuracy improvements on MSMT17. Similar conclusion can be made on VeRi-776. Baseline obtains 78.5% mAP when SIE encodes viewpoint information. The accuracy increases to 79.6% mAP when both camera IDs and viewpoint labels

are encoded at the same time. If the encoding is changed to $S_C[r] + S_V[q]$, which is sub-optimal as discussed in Section 3.3, we can only achieve 78.3% mAP on VeRi-776. Therefore, the proposed $S_{(C,V)}$ is a better encoding manner.

Visualization of Distance Distribution. As shown in Figure 6, the distribution gaps with cameras and viewpoints variations are obvious in Figure 6a and Figure 6b, respectively. When we introduce the SIE module into Baseline, the distribution gaps between inter-camera/viewpoint and intra-camera/viewpoint are reduced, which shows that the SIE module weakens the negative effect of the scene-bias caused by various cameras and viewpoints.

Ablation Study of λ . We analyze the influence of weight λ of the SIE module on the performance in Figure 7. When $\lambda = 0$, Baseline achieves 61.0% mAP and 78.2% mAP on MSMT17 and VeRi-776, respectively. With λ increasing, the mAP is improved to 63.0% mAP ($\lambda = 2.0$ for MSMT17) and 79.9% mAP ($\lambda = 2.5$ for VeRi-776), which means the SIE module now is beneficial for learning invariant features. Continuing to increase λ , the performance is degraded because the weights for feature embedding and the position embedding are weakened.

4.6. Ablation Study of TransReID

Finally, we evaluate the benefits of introducing JPM and SIE in Table 5. For the Baseline, JPM and SIE improve the performance by +2.6%/+1.0% mAP and +1.4%/+1.4% mAP on MSMT17/VeRi-776, respectively. With these two modules used together, TransReID achieves 64.9% (+3.9%) mAP and 80.6% (+2.4%) mAP on MSMT17 and VeRi-776, respectively. The experimental results show the effectiveness of our proposed JPM, SIE, and the overall framework.

Method	JPM	SIE	MSMT17		VeRi-776	
			mAP	R1	mAP	R1
Baseline	×	×	61.0	81.8	78.2	96.5
	✓	×	63.6	82.5	79.2	96.8
	×	✓	62.4	81.9	79.6	96.9
TransReID	✓	✓	64.9	83.3	80.6	96.9

Table 5: The ablation study of TransReID.

4.7. Comparison with State-of-the-Art Methods

In Table 6, our TransReID is compared with state-of-the-art methods on six benchmarks including person ReID, occluded ReID and vehicle ReID.

Person ReID. On MSMT17 and DukeMTMC-reID, TransReID* (DeiT-B/16) outperforms the previous state-of-the-art methods by a large margin (+5.5%/+2.1% mAP). On Market-1501, TransReID* (256×128) achieves comparable performance with state-of-the-art methods especially on

Backbone	Method	Size	MSMT17		Market1501		DukeMTMC		Occluded-Duke		Method	VeRi-776		VehicleID	
			mAP	R1	mAP	R1	mAP	R1	mAP	R1		mAP	R1	R1	R5
CNN	CBN [⊙] [61]	256×128	42.9	72.8	77.3	91.3	67.3	82.5	-	-	PRReID[13]	72.5	93.3	72.6	88.6
	OSNet [58]	256×128	52.9	78.7	84.9	94.8	73.5	88.6	-	-	SAN[32]	72.5	93.3	79.7	94.3
	MGN [44]	384×128	52.1	76.9	86.9	95.7	78.4	88.7	-	-	UMTS [16]	75.9	95.8	80.9	87.0
	RGA-SC [54]	256×128	57.5	80.3	88.4	96.1	-	-	-	-	VANet [⊙] [7]	66.3	89.8	83.3	96.0
	SAN [17]	256×128	55.7	79.2	88.0	96.1	75.7	87.9	-	-	SPAN [⊙] [5]	68.9	94.0	-	-
	SCSN [6]	384×128	58.5	83.8	88.5	95.7	79.0	91.0	-	-	PGAN [52]	79.3	96.5	78.0	93.2
	ABDNet [4]	384×128	60.8	82.3	88.3	95.6	78.6	89.0	-	-	PVEN [⊙] [30]	79.5	95.6	84.7	97.0
	PGFA [31]	256×128	-	-	76.8	91.2	65.5	82.6	37.3	51.4	SAVER [19]	79.6	96.4	79.9	95.2
	HOReID [43]	256×128	-	-	84.9	94.2	75.6	86.9	43.8	55.1	CFVMNet [38]	77.1	95.3	81.4	94.1
	ISP [59]	256×128	-	-	88.6	95.3	80.0	89.6	52.3	62.8	GLAMOR[39]	80.3	96.5	78.6	93.6
DeiT-B/16	Baseline	256×128	61.4	81.9	86.6	94.4	78.9	89.3	53.1	60.6	Baseline	78.4	95.9	83.1	96.8
	TransReID [⊙]	256×128	63.9	82.7	88.0	94.7	81.2	90.1	55.6	62.8	TransReID [⊙]	80.6	96.8	84.6	97.4
	TransReID [⊙]	384×128	65.5	83.5	88.1	94.9	81.3	90.2	-	-	TransReID [⊙]	81.2	96.8	-	-
	TransReID ^{*⊙}	256×128	66.2	84.3	88.4	95.0	81.9	91.1	58.1	66.4	TransReID ^{*⊙}	81.4	96.8	85.2	97.6
	TransReID ^{*⊙}	384×128	66.3	84.5	88.5	95.1	82.1	91.1	-	-	TransReID ^{*⊙}	82.3	97.1	-	-
ViT-B/16	Baseline	256×128	61.0	81.8	86.8	94.7	79.3	88.8	53.1	60.5	Baseline	78.2	96.5	82.3	96.1
	TransReID [⊙]	256×128	64.9	83.3	88.2	95.0	80.6	89.6	55.7	64.2	TransReID [⊙]	79.6	97.0	83.6	97.1
	TransReID [⊙]	384×128	66.6	84.6	88.8	95.0	81.8	90.4	-	-	TransReID [⊙]	80.6	96.9	-	-
	TransReID ^{*⊙}	256×128	67.4	85.3	88.9	95.2	82.0	90.7	59.2	66.4	TransReID ^{*⊙}	80.5	96.8	85.2	97.5
	TransReID ^{*⊙}	384×128	69.4	86.2	89.5	95.2	82.6	90.7	-	-	TransReID ^{*⊙}	82.0	97.1	-	-

Table 6: Comparison with state-of-the-art methods. DukeMTMC denotes the DukeMTMC-reID benchmark. The star * in the superscript means the backbone is with a sliding-window setting. Results are shown for person ReID datasets (left) and vehicle ReID datasets (right). Only the small subset of VehicleID is used in this paper. [⊙] and [⊙] indicate the methods are using camera IDs and viewpoint labels, respectively. [⊙] means both are used. Viewpoint and camera information are used wherever available. Best results for previous methods and best of our methods are labeled in bold.

mAP. Our method also shows superiority when compared with methods which also integrate camera information like CBN [61].

Occluded ReID. ISP implicitly uses human body semantic information through iterative clustering and HOReID introduces external pose models to align body parts. TransReID (DeiT-B/16) achieves 55.6% mAP with a large margin improvement (at least +3.3% mAP) compared to aforementioned methods, without requiring any semantic and pose information to align body parts, which shows the ability of TransReID to generate robust feature representations. Furthermore, TransReID* improves the performance to 58.1% mAP with the help of overlapping patches.

Vehicle ReID. On VeRi-776, TransReID* (DeiT-B/16) reaches 82.3% mAP surpassing GLAMOR by 2.0% mAP. When only using viewpoint annotations, TransReID* still outperforms VANet and SAVER on both VeRi-776 and VehicleID. Our method achieves state-of-the-art performance about 85.2% Rank-1 accuracy on VehicleID.

DeiT vs ViT vs CNN. TransReID* (DeiT-B/16) reaches competitive performance with existing methods under a fair comparison (ImageNet-1K pre-training). Extra results of our methods with ViT-B/16 are also reported in Table 6 for further comparison. DeiT-B/16 achieves similar performance with ViT-B/16 for shorter image patch sequences. When the number of input patches is increasing, ViT-B/16 reaches better performance than DeiT-B/16,

which shows ImageNet-21K pre-training provides ViT-B/16 better generalization capability. Although CNN-based methods mainly report performance with the ResNet50 backbone, they may include multiple branches, attention modules, semantic models, or other modules that increase computational consumption. We have conducted a fair comparison on inference speed between TransReID* and MGN [44] on the same computing hardware. Compared with MGN, TransReID* is 4.8% faster in speed. Therefore, TransReID* can achieve more promising performance under comparable computation to most of CNN-based methods.

5. Conclusion

In this paper, we investigate a pure transformer framework for the object ReID task, and propose two novel modules, *i.e.*, jigsaw patch module (JPM) and side information embedding (SIE). The final framework TransReID outperforms all other state-of-the-art methods by a large margin on several popular person/vehicle ReID datasets including MSMT17, Market-1501, DukeMTMC-reID, Occluded-Duke, VeRi-776 and VehicleID. Based on the promising results achieved by TransReID, we believe the transformer has great potential to be further explored for ReID tasks. Based on the rich experience gained from CNN-based methods, it is in prospect that more efficient transformer-based networks can be designed with better representation power and less computational cost.

References

- [1] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Toward transformer-based object detection. *arXiv preprint arXiv:2012.09958*, 2020. 12
- [2] Binghui Chen, Weihong Deng, Jiani Hu, Jiani Hu, and Jiani Hu. Mixed high-order attention network for person re-identification. In *ICCV*, 2019. 1
- [3] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *CVPR*, 2021. 3
- [4] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abdnnet: Attentive but diverse person re-identification. In *ICCV*, 2019. 1, 8
- [5] Tsai-Shien Chen, Chih-Ting Liu, Chih-Wei Wu, and Shao-Yi Chien. Orientation-aware vehicle re-identification with semantics-guided part attention network. In *ECCV*, pages 330–346. Springer, 2020. 8
- [6] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Saliency-guided cascaded suppression network for person re-identification. In *CVPR*, June 2020. 1, 8
- [7] Ruihang Chu, Yifan Sun, Yadong Li, Zheng Liu, Chi Zhang, and Yichen Wei. Vehicle re-identification with viewpoint-aware metric learning. In *ICCV*, pages 8282–8291, 2019. 2, 3, 8
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, and Xiaohua et al. Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1, 3
- [9] Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*, 2019. 11
- [10] Angela Fan, Pierre Stock, Benjamin Graham, Edouard Grave, Rémi Gribonval, Hervé Jégou, and Armand Joulin. Training with quantization noise for extreme model compression. *arXiv e-prints*, pages arXiv–2004, 2020. 11
- [11] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020. 3
- [12] Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. *arXiv preprint arXiv:1803.01719*, 2018. 11
- [13] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. Part-regularized near-duplicate vehicle re-identification. In *CVPR*, pages 3997–4005, 2019. 8
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2
- [15] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661. Springer, 2016. 11
- [16] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In *AAAI*, volume 34, pages 11165–11172, 2020. 8
- [17] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11173–11180, 2020. 8
- [18] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 3
- [19] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa. The devil is in the details: Self-supervised attention for vehicle re-identification. In *ECCV*, pages 369–386. Springer, 2020. 1, 8
- [20] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, pages 384–393, 2017. 2
- [21] Wei Li, Xiatian Zhu, Xiatian Gong, Shaogang, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 1
- [22] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE TIP*, 26(7):3492–3506, 2017. 2
- [23] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, pages 2167–2175, 2016. 2, 5
- [24] Xinchun Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *ICME*, pages 1–6. IEEE, 2016. 2, 5
- [25] Xinchun Liu, Wu Liu, Jinkai Zheng, Chenggang Yan, and Tao Mei. Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification. In *ACMMM*, pages 907–915, 2020. 3
- [26] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 11
- [27] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, pages 0–0, 2019. 1, 2, 3, 11
- [28] Hao Luo, Wei Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, and Chi Zhang. Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognition*, 94:53–61, 2019. 2
- [29] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NeurIPS*, pages 4905–4913, 2016. 1
- [30] Dechao Meng, Liang Li, Xuejing Liu, Yadong Li, Shijie Yang, Zheng-Jun Zha, Xingyu Gao, Shuhui Wang, and Qingming Huang. Parsing-based view-aware embedding network for vehicle re-identification. In *CVPR*, pages 7103–7112, 2020. 2, 3, 8
- [31] 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. 2, 3, 5, 8

- [32] Jingjing Qian, Wei Jiang, Hao Luo, and Hongyan Yu. Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification. *Measurement Science and Technology*, 31(9):095401, 2020. 3, 8
- [33] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35. Springer, 2016. 2, 5
- [34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 1, 13
- [35] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, June 2019. 2
- [36] Yifan Sun, Changmao Cheng, Yuhang Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6398–6407, 2020. 1, 2
- [37] 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. 1, 2
- [38] Ziruo Sun, Xiushan Nie, Xiaoming Xi, and Yilong Yin. Cfmnet: A multi-branch network for vehicle re-identification based on common field of view. In *ACMMM*, pages 3523–3531, 2020. 8
- [39] Abhijit Suprem and Calton Pu. Looking glamorous: Vehicle re-id in heterogeneous cameras networks with global and local attention. *arXiv preprint arXiv:2002.02256*, 2020. 8
- [40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 2, 3
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 6000–6010, 2017. 3
- [42] Guangcong Wang, Jian-Huang Lai, Wenqi Liang, and Guangrun Wang. Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification. In *CVPR*, pages 10568–10577, 2020. 1
- [43] 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. 8
- [44] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACMMM*, pages 274–282, 2018. 1, 2, 3, 8
- [45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 1
- [46] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 2, 5
- [47] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 420–428. ACM, 2017. 3
- [48] Bryan (Ning) Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. Second-order non-local attention networks for person re-identification. In *ICCV*, October 2019. 1
- [49] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE TIP*, 28(6):2860–2871, 2019. 2
- [50] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI*, 2021. 2
- [51] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 6
- [52] Xinyu Zhang, Rufeng Zhang, Jiewei Cao, Dong Gong, Mingyu You, and Chunhua Shen. Part-guided attention learning for vehicle re-identification. *IEEE TITS*, 2020. 8
- [53] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pages 6848–6856, 2018. 4
- [54] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, June 2020. 1, 8
- [55] 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. 2, 5
- [56] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM TOMM*, 14(1):13, 2018. 2
- [57] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, volume 34, pages 13001–13008, 2020. 5
- [58] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3702–3712, 2019. 8
- [59] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. *ECCV*, 2020. 8
- [60] Zihui Zhu, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Xing Sun, and Weishi Zheng. Aware loss with angular regularization for person re-identification. In *AAAI*, volume 34, pages 13114–13121, 2020. 3
- [61] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *ECCV*, pages 140–157. Springer, 2020. 2, 3, 8

Appendix

A. More Experimental Results

A.1. Study on Transformer-based Strong Baseline

A transformer-based strong baseline with a few critical improvements has been introduced in Section 3.1 of the main paper. In this section, hyper-parameters and the settings for training such a baseline model will be analyzed in detail. Ablation studies are shown in Table 7 for performance on MSMT17 and Veri-776 with different variations of the training settings.

Initialization and hyper-parameters. For our experiments, we initialize the pure transformer with ViT or DeiT ImageNet pre-trained weights and we initialize the weights for the SIE with a truncated normal distribution [12]. Compared with ViT, DeiT is more sensitive to hyper-parameter settings. For the training of DeiT, we use a learning rate of 0.05 on MSMT17 and a high random erasing probability with 0.8 on each dataset to avoid overfitting. Other hyper-parameters settings are the same with ViT.

Optimizer. Transformers are sensitive to the choice of the optimizer. Directly applying Adam optimizer with the hyper-parameters commonly used in ReID community [27] to transformer-based models will cause a significant drop in performance. AdamW [26] is a commonly used optimizer for training transformer-based models, with much better performance compared with Adam. The best results are actually achieved by SGD in our experiments.

Network Configuration. Position embeddings incorporate crucial spatial information which provides a significant boost in performance and is one of the key ingredients of our proposed training procedure. Without the position embeddings, the performance decreases by 38.6% mAP and 10.2% mAP on MSMT17 and VeRi-776,

respectively.

Introducing stochastic depth [15] can boost the mAP performance by about 1%, and it has also been proved to facilitate the convergence of transformer, especially for those deep ones [9, 10]. Regarding other regularization methods, adding either drop out or attention drop out will result in performance drop. In our experiments, we set all the probability of regularization methods as 0.1.

Loss Function. Different choices of loss functions have been compared in the bottom section of Table 7. The soft version of triplet loss provides 0.7% mAP improvement on MSMT17 compared with the regular triplet loss. Introducing label smoothing is harmful to performance, even though it has been a widely adopted trick. Therefore, the best combination for loss functions is soft triplet loss and cross entropy loss without label smoothing.

A.2. More Ablation Studies of JPM and SIE

In the main paper, we have demonstrated the effectiveness of using JPM and SIE based on the Baseline (ViT-B/16). More results about JPM and SIE are shown in Table 8 and Table 9 respectively, with the Baseline ViT-B/16_{s=12}, which is supposed to have better feature representation ability and higher performance than ViT-B/16. From Table 8, we observe that: (1) The proposed JPM performs better with the rearrange schemes, indicating that the shift and patch shuffle operation help the model learn more discriminative features which are robust against perturbations. (2) The JPM module provides a consistent performance improvement over the baselines, no matter the baseline is ViT-B/16 or the stronger ViT-B/16_{s=12}, demonstrating the effectiveness of the proposed JPM.

Similar conclusions can be made from Table 9. (1) We make better use of the viewpoint and camera information so that they are complementary with each other and combining them leads to the best performance. (2) Introducing SIE

Method	OPT	PE	SP	DO	ADO	STL	LS	MSMT17		VeRi-776	
								mAP	R1	mAP	R1
ViT-B/16 Baseline	SGD	✓	✓	✗	✗	✓	✗	61.0	81.8	78.2	96.5
Optimizer	Adam	✓	✓	✗	✗	✓	✗	37.4 (-24.6)	60.2 (-21.6)	65.8 (-12.4)	91.7 (-4.8)
	AdamW	✓	✓	✗	✗	✓	✗	60.6 (-0.4)	81.7 (-0.1)	78.0 (-0.2)	96.5 (-0.0)
Network Configuration	SGD	✗	✓	✗	✗	✓	✗	22.4 (-38.6)	38.3 (-43.5)	68.0 (-10.2)	92.8 (-3.7)
	SGD	✓	✗	✗	✗	✓	✗	59.9 (-1.1)	80.2 (-1.6)	77.2 (-1.0)	96.1 (-0.4)
	SGD	✓	✓	✓	✗	✓	✗	60.0 (-1.0)	80.7 (-1.1)	78.0 (-0.2)	96.3 (-0.2)
	SGD	✓	✓	✗	✓	✓	✗	58.0 (-3.0)	78.8 (-3.0)	74.3 (-3.9)	94.9 (-1.6)
Loss Function	SGD	✓	✓	✗	✗	✗	✗	60.3 (-0.7)	81.3 (-0.5)	77.5 (-0.7)	95.6 (-0.9)
	SGD	✓	✓	✗	✗	✓	✓	59.8 (-1.2)	80.4 (-1.4)	77.4 (-0.8)	96.5 (-0.0)

Table 7: Ablation study about training settings on MSMT17 and VeRi-776. The first row corresponds to the default configuration employed by our transformer-based strong baseline (ViT-B/16 as default backbones). The symbols ✓ and ✗ indicate that the corresponding setting is included or excluded, respectively. mAP(%) and R1(%) accuracy scores are reported. The abbreviations OPT, PE, SP, DO, ADO, STL, LS denote Optimizer, Position Embedding, Stochastic Depth [15], Drop Out, Attention Drop Out, Soft Triplet Loss, Label Smoothing, respectively.

Backbone	#groups	MSMT17		VeRi-776	
		mAP	R1	mAP	R1
Baseline (ViT-B/16)	-	61.0	81.8	78.2	96.5
+JPM	1	62.9	82.5	78.6	97.0
+JPM	2	62.8	82.1	79.1	96.4
+JPM	4	63.6	82.5	79.2	96.8
+JPM w/o rearrange	4	63.1	82.4	79.0	96.7
+JPM w/o local	4	63.5	82.5	79.1	96.6
Baseline (ViT-B/16 _{s=12})	-	64.4	83.5	79.0	96.5
+JPM	4	66.5	84.8	80.0	97.0
+JPM w/o rearrange	4	66.1	84.5	79.6	96.7
+JPM w/o local	4	66.3	84.5	79.8	96.8

Table 8: Detailed ablation study of jigsaw patch module (JPM). ‘w/o rearrange’ means the patch sequences are split into subsequences without rearrangement. ‘w/o local’ means we evaluate the global feature without concatenating local features.

Method	Camera	View	MSMT17		VeRi-776	
			mAP	R1	mAP	R1
Baseline (ViT-B/16)	✗	✗	61.0	81.8	78.2	96.5
	✓	✗	62.4	81.9	78.7	97.1
	✗	✓	-	-	78.5	96.9
	✓	✓	-	-	79.6	96.9
Baseline (ViT-B/16 _{s=12})	✗	✗	64.4	83.5	79.0	96.5
	✓	✗	65.9	84.1	79.4	96.4
	✗	✓	-	-	79.3	97.0
	✓	✓	-	-	80.3	96.9

Table 9: Detailed ablation study of side information embeddings (SIE). Experiments of viewpoint information are only conducted on VeRi-776 as the person ReID datasets do not provide viewpoint annotations. The symbols ✓ and ✗ indicate that the corresponding information is included or excluded.

provides consistent improvement over the baselines of either ViT-B/16 or ViT-B/16_{s=12}.

B. Analysis on Rearranging Patches in JPM

Although transformers can capture the global information in the image very well, a patch token still has a strong correlation with the corresponding patch. ViT-FRCNN [1] shows that the output embeddings of the last layer can be reshaped as a spatial feature map that includes location information. In other words, if we directly divide the original patch embeddings into k parts, each part may only consider a part of the continuous patch embeddings. Therefore, to better capture the long-range dependencies, we rearrange the patch embeddings and then re-group them into different parts, each of which contains several random patch embeddings of an entire image. In this way, the JPM module help to learn robust features with improved discrimination ability and more diversified coverage.

To verify the above point, we visualize the learned attention of local features $[f_l^1, f_l^2, \dots, f_l^k]$ ($k = 4$ in our cases) by JPM module in Figure 8. Brighter region means

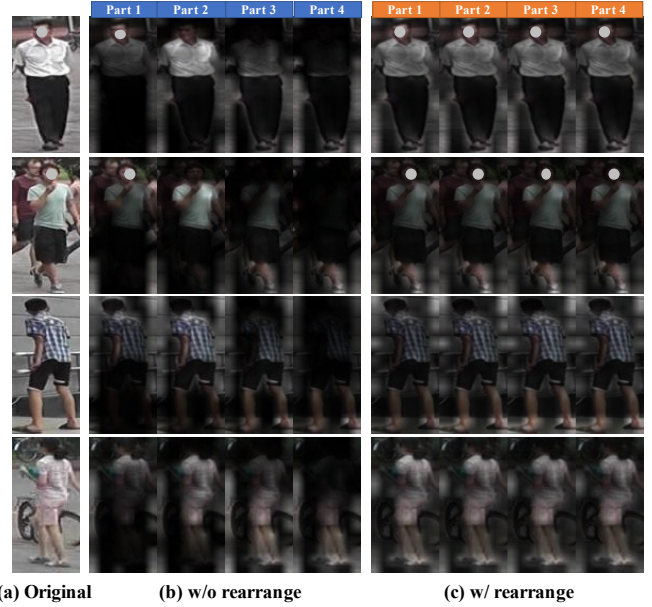


Figure 8: Visualization of the learned attention masks for local features by JPM module. Higher weight results in higher brightness of the region. Note that we visualize the learned attention weights which are averaged among attention heads in the last layer. Faces in the images are masked for anonymization.

higher corresponding weights. Several observations can be made from Figure 8: (1) The attention learned by the ‘JPM w/o rearrange’ tends to focus on limited receptive fields (i.e. the range of the corresponding patch sequences) due to global sequences being split into several isolated subsequences. For example, ‘Part 1’ mainly pays attention to the head of a person, and ‘Part 4’ is mainly focused around the bottom area. (2) In contrast, ‘JPM w/ rearrange’ is able to capture long-range dependencies and each part has attention responses across the whole image because it is forced to extend its scope to the whole image through the rearranging operation. (3) According to the superior ReID performance and the intuitive visualization of rearranging effect, JPM is proved to not only capture more details at finer granularities but also learn robust and discriminative representations in the global context.

C. More Visualization of Attention Maps

In the main paper, we use Grad-CAM to visualize the gradient responses of our schemes, CNN-based methods, and CNN+attention methods. Following the similar setup, Figure 9 shows more visualization results, with the similar conclusion that transformer-based methods capture global context information and more discriminative parts, which are further enhanced in our proposed TransReID for better performance.

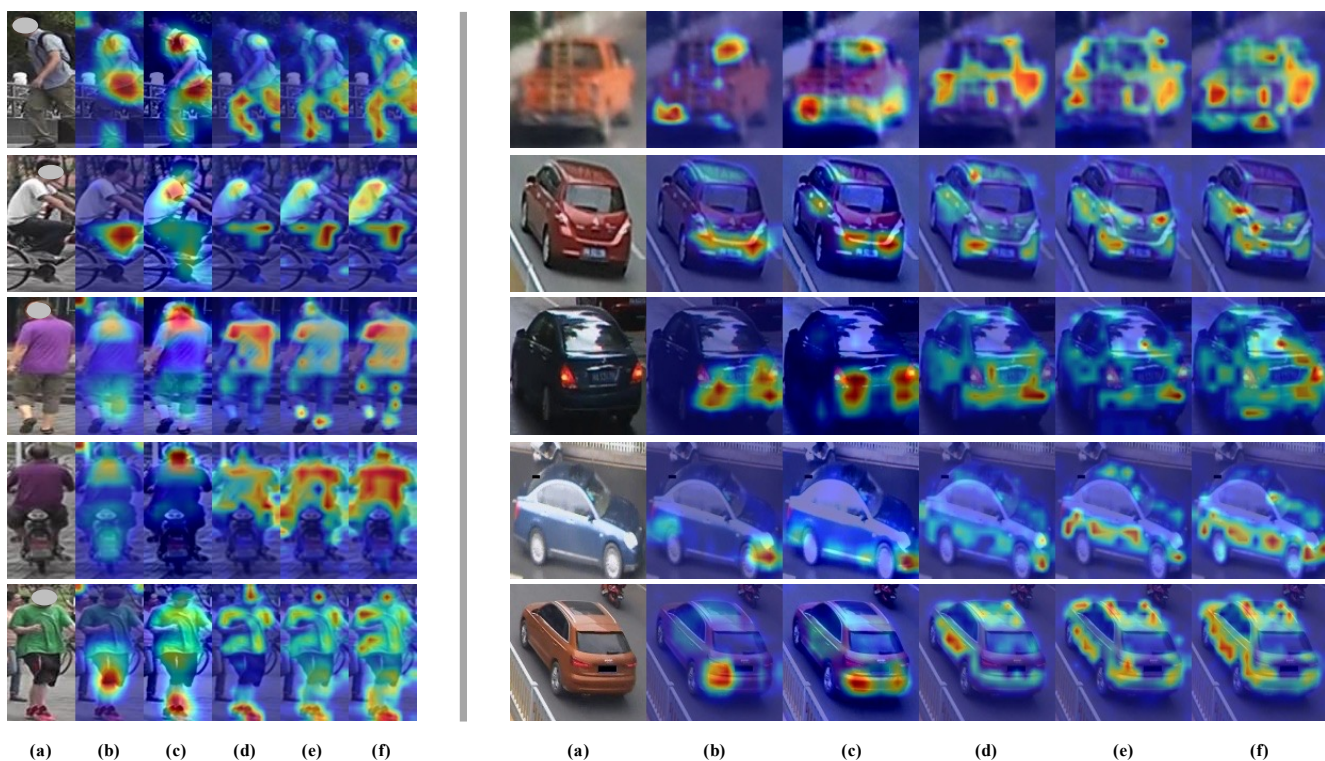


Figure 9: Grad-CAM [34] visualization of attention maps. (a) Original images, (b) CNN-based methods, (c) CNN+Attention methods, (d) Transformer-based baseline, (e) TransReID w/o rearrange, (f) TransReID. Faces in the images are masked for anonymization.