

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

In this paper, we explore the Vision Transformer (ViT), a pure transformer-based model, for the object re-identification (ReID) task. With several adaptations, a strong baseline ViT-BoT is constructed with ViT as backbone, which achieves comparable results to convolution neural networks- (CNN-) based frameworks on several ReID benchmarks. Furthermore, two modules are designed in consideration of the specialties of ReID data: (1) It is super natural and simple for Transformer to encode non-visual information such as camera or viewpoint into vector embedding representations. Plugging into these embeddings, ViT holds the ability to eliminate the bias caused by diverse cameras or viewpoints. (2) We design a Jigsaw branch, parallel with the Global branch, to facilitate the training of the model in a two-branch learning framework. In the Jigsaw branch, a jigsaw patch module is designed to learn robust feature representation and help the training of transformer by shuffling the patches. With these novel modules, we propose a pure-transformer framework dubbed as TransReID, which is the first work to use a pure Transformer for ReID research to the best of our knowledge. Experimental results of TransReID are superior promising, which achieve state-of-the-art performance on both person and vehicle ReID benchmarks.

1. Introduction

Object re-identification (ReID) is a challenging task including person ReID and vehicle ReID, which aims to identify all images of a target object in a cross-camera system. CNN-based methods [16, 23, 31, 37] have achieved great success in both person and vehicle ReID lately. In the recent years, as transformer becomes popular for language modeling, it has also shown promising performance in numerous computer-vision applications [9, 15]. On the one hand, ‘CNN + Transformer’ becomes a popular paradigm for computer vision [1, 10, 2, 30, 18, 17, 41]; on the other

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

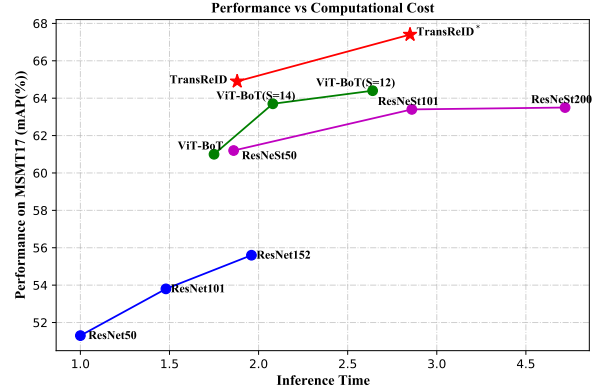


Figure 1: The comparison among TransReID, ViT-BoT, ResNet and ResNeSt on MSMT17. The computational cost of ResNet50 is taken as the baseline for inference time comparison.

hand, pure-transformer [27, 3, 7, 34] is attracting more and more attention.

The Vision Transformer (ViT) [7] is the first work to show that a pure transformer architecture can be directly applied for image classification, by treating an image as a sequence of patches. ViT achieves impressive performance with large-scale pretraining datasets. By cutting the image into patches, transformers are capable of globally attending all the patches at every layer, making the spatial correspondence weaker between the input and intermediate features. However, in ReID, the spatial alignment is critical for feature learning [32, 26]. This naturally raises the question whether ViT can be fine-tuned for tasks that need more spatial alignment than image recognition does. We adopt the ViT model as backbone for feature extraction, and by making several adaptations, a reasonably strong baseline, named as ViT-BoT, is constructed inspired by BagofTricks (BoT) [23]. ViT-BoT achieves comparable performance with CNN-based backbones including ResNet [12], ResNeSt [42] and MGN [37], etc. It demonstrates the potential that transformer-based model can be a peer backbone for ReID.

Different from image classification, ReID data usually

include some non-visual clues such as cameras and viewpoints, which have been verified effective in CNN-based works [51, 5, 50]. However, it usually requires special design for CNN-based methods to incorporate these useful information. For example, Camera-based Batch Normalization (CBN) [46] modifies the BN layers to learn camera-related knowledge; Viewpoint-Aware Network (VANet) [5] designs a viewpoint-aware metric learning approach for similar viewpoint and different viewpoints in two feature spaces to learn viewpoint-invariant features. Instead of treating them with various special designs, we propose a unified framework to simply incorporate different kinds of non-visual clues (side information) to learn invariant features in the context of transformer. Compared with these CNN-based models, these side information can be easily encoded through vector embedding in the transformer, and the module is introduced as the Side Information Embedding (SIE).

For better training of ViT-BoT, we design a new branch called Jigsaw branch. It is in the last layer of ViT-BoT, parallel with the standard Global branch. Even though the Global branch encodes global information of all patches, only a few discriminative patches make the main contribution. Thus, in the Jigsaw branch, a new module named as jigsaw patch module (JPM) is designed, inspired by these stripe based methods [32, 37, 24, 28]. Different from those CNN-based methods, where the stripes are collected as local continuous patches, the patches in JPM are shuffled to form new larger patches. There are two pros for this design. Firstly, the shuffled patches push the model to learn a robust representation that is invariant to perturbations; secondly, with shuffled patches, the newly formed patches contain global information, making it easier to make decision of ID recognition. Combining the novel designed SIE and JPM modules, we propose the final model architecture termed as TranReID. As shown in Figure 1, TransReID achieves great Speed-accuracy tradeoff.

In summary, the contributions of this paper are as follows:

- 1) We propose a pure transformer framework to ReID tasks for the first time, and construct a strong baseline ViT-BoT with several adaptations. ViT-BoT achieves comparable performance with state-of-the-art CNN-based frameworks.

- 2) The Side Information Embedding (SIE) is introduced as a unified framework to encode various kinds of side information for the object ReID. It is demonstrated that SIE can reduce the feature bias caused by different cameras or object viewpoints.

- 3) Jigsaw Patches Module (JPM) is proposed to take advantage of stripe-based ideas. With a shuffle operation, the JPM facilitates the training for a better and more robust feature representation in a two-branch learning framework.

- 4) The TransReID achieves state-of-the-art performance on both person and vehicle ReID benchmarks including MSMT17, Market-1501, DukeMTMC-reID, Occluded-Duke, VeRi-776 and VehicleID.

2. Related Work

2.1. Object ReID

The studies of object ReID mainly focus on person ReID and vehicle ReID. The current state-of-the-art methods are mostly based on the CNN structure. Summarizing these works, we can find that representation learning, local features and invariant features are critical for successful ReID, and the corresponding related works are presented as below.

Representation Learning. A popular pipeline for object ReID is to design suitable loss functions to train a CNN backbone (e.g. ResNet [12]), which is used to extract features of images. The loss functions can be categorized into classification-based loss and metric loss. For classification-based loss (a.k.a. ID loss), Zheng *et al.* [47] proposed the ID-discriminative embedding (IDE) to train the ReID model as image classification and it is fine-tuned from the ImageNet [6] pre-trained models. Different from ID loss, metric loss regards the ReID task as a clustering or ranking problem. The most widely used metric loss is the triplet loss [19]. Luo *et al.* [23] proposed the BNNeck to better combine ID loss and triplet loss. Sun *et al.* [31] proposed a unified perspective for ID loss and triplet loss.

Local Features. It learns fine-grained features to aggregate different part/region features. The fine-grained parts are either automatically generated by roughly horizontal stripes or by semantic parsing. Methods like PCB [32], MGN [37], AlignedReID++ [24], SAN [28], 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 [22, 25, 39, 26].

Invariant Features. In a cross-camera system, there exists pose, orientation, illumination, resolution variances caused by different camera setup and object viewpoints. Some works [51, 5] use side information such as camera ID or viewpoint information to learn invariant features. For example, Camera-based Batch Normalization (CBN) [51] forces the image data from different cameras to be projected onto the same subspace, so that the distribution gap between any camera pair is largely diminished. Viewpoint/Orientation-invariant feature learning [5, 50] is also important for both person and vehicle ReID.

2.2. Transformer in Vision

The Transformer model is proposed in [35] 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.* [9] and Salman *et al.* [15] have surveyed the application of the Transformer in the field of computer vision.

The Transformer model is initially used to handle sequential features extracted by CNN models for the videos. Girdhar *et al.* [8] use a variant of transformer architecture to aggregate contextual cues in a video relevant to a particular person. The Transformer models are then extended to some popular computer-vision tasks including image processing [2], object detection [1], semantic segmentation [46, 41], object tracking [30], etc. For example, Image Processing Transformer (IPT) [2] 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. The detection transformer (DETR) [1] redesigns the framework of object detection, which is a simple and fully end-to-end object detector.

Pure Transformer models are becoming more and more popular. ViT [7] 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.* [34] 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. We extend ViT to object ReID task and show its effectiveness in this paper.

3. Methodology

In this section, we make some adaptations for ReID task in Sec. 3.1 and the newly constructed baseline is named as ViT-BoT. Based on ViT-BoT, we propose the TransReID framework in Sec. 3.2 which includes a Side Information Embedding (SIE) module and a Jigsaw Patch Module (JPM).

3.1. ViT-BoT

The framework of ViT-BoT is shown in Figure 2. Since the original training details of ViT are not directly applicable for ReID task, several detailed adaptations are made to achieve the strong baseline ViT-BoT.

Overlapping Patches. As a preprocessing step, ViT splits the images into N non-overlapping patches, leaving the local neighboring structures around the patches not well preserved. Instead, we use a sliding window to generate patches with overlapping pixels. Assuming that the step size of the sliding window is S pixels, size of the patch is

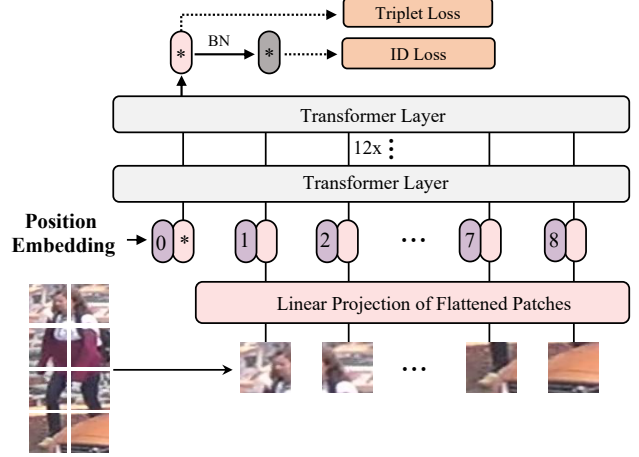


Figure 2: The framework of ViT-BoT. The preprocessing of the input image is same with ViT (a non-overlapping partition is shown). The final output token is used as the global feature.

$P = 16$ pixels and the shape of the area where two adjacent patches overlap is $(P - S) \times P$. To summarize, given an input image, after resizing to a fixed resolution $H \times W$, will be split into N patches.

$$N = N_H \times N_W = \lfloor \frac{H + S - P}{S} \rfloor \times \lfloor \frac{W + S - P}{S} \rfloor \quad (1)$$

where $\lfloor \cdot \rfloor$ is the floor function. The larger the overlapping area, the more patches the image will be split into. More patches usually can bring better performance and in the meanwhile cause more computations. A trade-off between performance and computational cost needs to be made here. For better distinguishment, ViT-BoT _{$s=12$} means the image is split with $S = 12$, and for case $S = P$ (the non-overlapping version) the subscript is omitted for simplicity.

Position Embedding. The position embedding ρ_i encodes the position information of the i -th patch p_i , which is important for the Transformer Encoder to encode spatial information. The parameters of ViT pre-trained on ImageNet are loaded to facilitate training. However, as the image resolution for ReID task is different from the one in ViT, the position embedding pretrained on ImageNet cannot be directly imported here. Therefore, a bilinear interpolation is introduced to the position embedding at runtime to help ViT-BoT to handle any given input size and shape. Similar to ViT, the position embedding of ViT-BoT is also learnable.

Feature Learning. Given an image split into a sequence of patches, a learnable embedding (i.e. class token) is prepended to the embeddings of patches, and the class

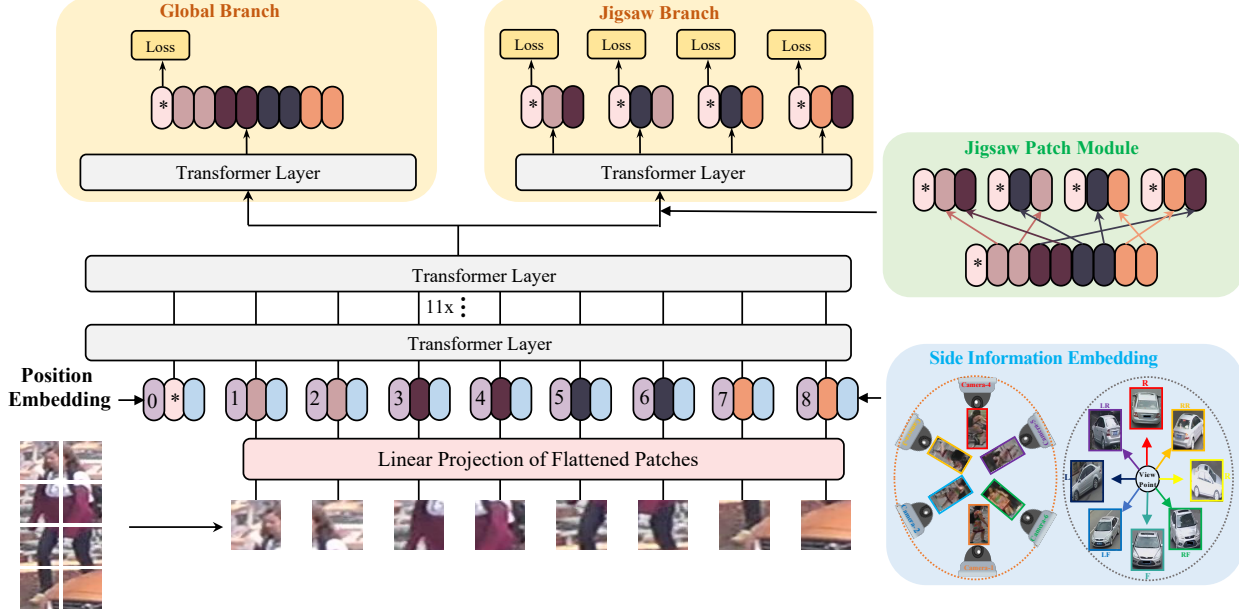


Figure 3: The framework of proposed TransReID. Side Information Embedding (light blue) encodes non-visual information such as camera or viewpoint into embedding representations. It is input into the transformer encoder together with patch embedding and position embedding. The last layer includes two independent transformer layers. One is standard to encode the global feature. The other one 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 the global feature and local features contribute to ReID loss.

token of the last encoder layer (final class token) serves as the global feature representation of an image. The final class token is denoted as f , and the remaining outputs corresponding to the input patches are denoted as $P_o = \{p_{o1}, p_{o2}, p_{o3}, \dots, p_{oN}\}$, where N is the number of total patches. Inspired by [23], we introduce the BNNeck after the final class token. The ID loss \mathcal{L}_{ID} is the cross-entropy loss without label smoothing. For a triplet set $\{a, p, n\}$, the triplet loss is the soft-margin version as follows:

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

3.2. TransReID

Although ViT-BoT can achieve impressive performance in the object ReID task, it does not take advantage of the specialties in ReID data. To make better exploration of side information and fine-grained parts, we propose the Side Information Embedding (SIE) and the Jigsaw Patch Module (JPM). With SIE and JPM, the proposed framework TransReID is presented in Figure 3.

3.2.1 Side Information Embedding

In object ReID, a challenging problem is the appearance bias caused by various cameras, viewing angles and other factors. To tackle this problem, CNN-based frameworks

usually need to modify the network structure or design specific loss functions to incorporate those non-visual clues (side information) such as camera IDs and viewpoint predictions.

Transformer model is perfectly suited here, as it can easily fuse these side information by encoding them into embedding representations. Similar to the position embedding, we can apply learnable layers to encode side information. In specific, if the camera ID of an image is C , then its camera embedding can be denoted as $\mathcal{S}(C)$. Different from the position embedding which varies between patches, camera embedding $\mathcal{S}(C)$ is the same for all patches of an image. In addition, if the viewpoint of the object is available, either by a viewpoint estimation algorithm or human annotations, we can also encode the viewpoint label V as $\mathcal{S}(V)$ for all patches of an image.

Now it comes the problem about how to integrate two different types of information. A trivial solution might be directly adding up the two embeddings $\mathcal{S}(C) + \mathcal{S}(V)$, but it might make the two embeddings canceled out by each other. We propose to encode camera ID C and viewpoint label V jointly as $\mathcal{S}(C, V)$. In other words, for C_N camera IDs and V_N viewpoint labels, $\mathcal{S}(C, V)$ has a total of $C_N \times V_N$ different values. Finally, the input embedding of i -th patch

is as follows:

$$E^i = \mathcal{F}(p_i) + \rho_i + \lambda \mathcal{S}(C, V) \quad (3)$$

where \mathcal{F} is the linearly projection to learn feature embedding and λ is a hyperparameter to balance the weight of $\mathcal{S}(C, V)$. As the position embedding ρ_i is different for each patch but the same across different images, and $\mathcal{S}(C, V)$ is the same for each patch but may have different values for different images, TransReID is able to learn two different embeddings which can then be added directly. The whole input embeddings is $[E_0; E_1, E_2, \dots, E_N]$, where E_0 is the class token.

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

3.2.2 Jigsaw Patch Module

We change the last layer of ViT-BoT to two parallel branches which learn global features and local features with two independent Transformer layers. Suppose the hidden features input to the last layer are denoted as $Z_{l-1} = [z_{l-1}^0; z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^N]$. The Global branch is a standard transformer which encodes Z_{l-1} into $Z_l = [f_g; z_l^1, z_l^2, \dots, z_l^N]$, where f_g is viewed as the global feature of CNN-based methods. To learn fine-grained part-level features, a straightforward solution one would like to try 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 transformer layer to learn k local features denoted as $\{f_l^1, f_l^2, \dots, f_l^k\}$. f_l^k is the output token of k -th group. Two recent works [40, 13] show that a token embedding is mainly determined by its nearby tokens. Therefore, an group of nearby patches embedding mainly observe a limited continuous area.

To address this issue, we propose the Jigsaw Patch Module (JPW) to shuffle patches before grouping them. The shuffle operation is implemented by a shift operation and a patch shuffle operation inspired by ShuffleNet [43] as follows:

- **Step1: The shift operation.** The first m patches (we recommend $m < H_N$) are moved to the end. $[z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^N]$ 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 is further shuffled by the patch shuffle operation with k groups. The hidden features become $[z_{l-1}^{x1}, z_{l-1}^{x2}, \dots, z_{l-1}^{xN}]$, $x_i \in [1, N]$.

We divide the shuffled features into k groups according to the previous description. JPM encodes them into k local features $\{f_l^1, f_l^2, \dots, f_l^k\}$ by a shared transformer. With the shuffle operation, the local feature f_l^k can cover patches from different body parts or vehicle parts. 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:

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

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.

4. Experiments

4.1. Datasets

We evaluate our proposed method on four person ReID datasets, Market-1501 [45], DukeMTMC-reID [29], MSMT17 [38], Occluded-Duke [26], and two vehicle ReID datasets, VeRi-776 [21] and VehicleID [20]. 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 dataset provides 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 with 10 pixels, random cropping and random erasing [48]. The batch size is set to 64 with 4 images per ID. SGD optimizer is employed with the weight decay of $1e-4$. The learning rate is initialized as 0.01 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, in this paper.

Backbone	Training 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	4.72x	63.5	83.5	77.9	96.4
ViT-BoT	1.75x	61.0	81.8	78.2	96.5
ViT-BoT _{s=14}	2.08x	63.7	82.7	78.6	96.4
ViT-BoT _{s=12}	2.64x	64.4	83.5	79.0	96.5

Table 2: Comparison of ViT-BoT and BoT with 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

All the experiments are performed with one Nvidia Tesla V100 GPU using the PyTorch toolbox¹ with FP16 training.

Evaluation Protocols. Following conventions in ReID community, we evaluate all methods with Cumulative Matching Characteristic (CMC) curves and the mean Average Precision (mAP). All the experimental results are performed under the setting of single query.

4.3. Results of ViT-BoT Baseline

In this section, ViT-BoT is compared with BoT to mainly demonstrate its effectiveness as a baseline. To show the trade-off between computation and performance, several different backbones are chosen for BoT, and different choices of step size S to form overlapping patches are also presented. For a comprehensive comparison, in addition to the performance on ReID benchmarks, inference time consumption of all backbones is included as well.

As shown in Table 2, BoT with larger backbone consistently achieve better performance on both MSMT17 and VeRi-776. There exist a large gap in model capacity between the ResNet series and ViT. However, ViT-BoT achieves similar performance with ResNeSt50 [42] backbone on MSMT17 and VeRi-776, with less inference time than ResNeSt50 (1.75x vs 1.86x). When we reduce the step size of the sliding window s , the performance of the ViT-BoT can be improved while the inference time is also increasing. ViT-BoT_{s=12} is faster than ResNeSt200 (2.64x vs 4.75x), and ViT-BoT_{s=12} performs slightly better than ResNeSt200 on ReID benchmarks. Therefore, ViT-BoT_{s=12} achieves better speed-accuracy trade-offs than ResNeSt200. In addition, we believe that ViT still has lots of room for improvement in terms of computational efficiency.

¹<http://pytorch.org>

Model	Cam	View	MSMT17		VeRi-776	
			mAP	R1	mAP	R1
ViT-BoT	✓		61.0	81.8	78.2	96.5
			62.4	81.9	78.7	97.1
	✓	✓	-	-	78.5	96.9
ViT-BoT _{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 3: Ablation Study of SIE. Since the person ReID datasets do not provide viewpoint annotations, viewpoint information can only be encoded in VeRi-776.

4.4. Ablation Study of SIE

Performance Analysis. In Table 3, we evaluate the effectiveness of the SIE on MSMT17 and VeRi-776. MSMT17 doesn't provide viewpoint annotations, so results of SIE only encoding camera information are shown for MSMT17. VeRi-776 does not only have 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 info.

When SIE encodes only the camera IDs of images, ViT-BoT and ViT-BoT_{s=12} get 1.4% and 1.1% mAP improvements on MSMT17, respectively. Similar conclusion can be made on VeRi-776. ViT-BoT obtains 78.5% mAP when SIE encode viewpoint information. The accuracy increase to 79.6% mAP when both camera IDs and viewpoint labels are encoded at the same time. If the encoding is changed to $\mathcal{S}(C) + \mathcal{S}(V)$, which is sub-optimal as discussed in Section 3.2.1, ViT-BoT only can achieve 78.3% mAP on VeRi-776. Therefore, the proposed $\mathcal{S}(C, V)$ is a better encoding manner.

As shown in the bottom half of Table 3, SIE is also effective when added to a stronger baseline ViT-BoT_{s=12}. The observation is similar to the case in ViT-BoT, and the mAP with all possible information encoded can be improved to 65.9% and 80.3% on MSMT17 and VeRi-776, respectively.

Appearance Feature Bias. In order to verify that SIE can reduce the appearance bias, the distributions of pairwise similarities are visualized for pairs of inter-camera, intra-camera, inter-viewpoint, and intra-viewpoint on VeRi-776 in Figure 4. The distribution gaps between cameras and viewpoints are obvious in Figure 4a and Figure 4b, respectively. When we introduce the SIE module into ViT-BoT, the distribution gaps between inter-camera/viewpoint and intra-camera/viewpoint are reduced in Figure 4c and Figure 4d, which shows that the SIE module reduces the

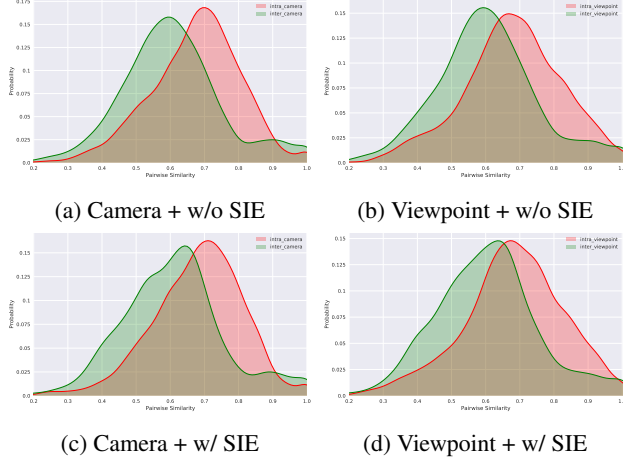


Figure 4: We visualize the distributions of inter-camera, intra-camera, inter-viewpoint and intra-viewpoint distance on VeRi-776. (a) and (c) show inter-camera and intra-camera similarities. (b) and (d) show inter-viewpoint and intra-viewpoint similarities.

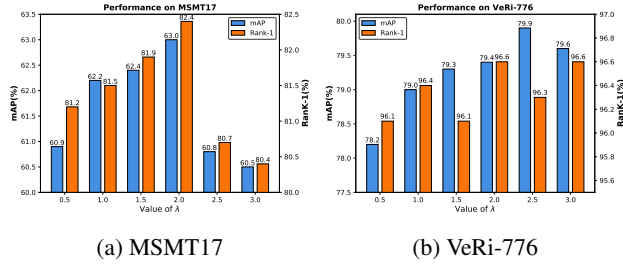


Figure 5: Impact of the hyper-parameter λ .

appearance bias caused by various cameras and viewpoints.

Ablation Study of λ . The balancing weight λ is a hyper-parameter to tune in SIE module. We analyze the influence of λ on the performance in Figure 5. When $\lambda = 0$, the 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.

4.5. Ablation Study of JPM

The effectiveness of the proposed JPM module is validated in Table 4. With the baseline ViT-BoT, JPM provides +2.6% mAP and 1.0% mAP improvements on MSMT17 and VeRi-776, respectively. Increasing the number of groups k can improve the performance while slightly increase inference time. In our experience, $k = 4$ is a choice to trade off speed and performance. Comparing

Backbone	#groups	MSMT17		VeRi-776	
		mAP	R1	mAP	R1
ViT-BoT	-	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 shuffle	4	63.1	82.4	79.0	96.7
+JPM w/o local	4	63.5	82.5	79.1	96.6
ViT-BoT _{s=12}	-	64.4	83.5	79.0	96.5
+JPM	4	66.5	84.8	80.0	97.0
+JPM w/o shuffle	4	66.1	84.5	79.6	96.7
+JPM w/o local	4	66.3	84.5	79.8	96.8

Table 4: The ablation study of jigsaw patch module. ‘w/o shuffle’ means the patch features are split into parts without shuffle. ‘w/o local’ means we evaluate the global feature without concatenating local features.

Backbone	MSMT17		VeRi-776	
	mAP	R1	mAP	R1
ViT-BoT	61.0	81.8	78.2	96.5
+SIE	62.4	81.9	79.6	96.9
+JPM	63.6	82.5	79.2	96.8
TransReID	64.9	83.3	80.6	96.9
ViT-BoT _{s=12}	64.4	83.5	79.0	96.5
+SIE	65.9	84.1	80.3	96.9
+JPM	66.5	84.8	80.0	97.0
TransReID*	67.4	85.3	81.7	97.1

Table 5: The ablation study of TransReID.

JPM and JPM w/o shuffle, we can observe the shuffle operation help 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 use only the global feature as an efficient variation with lower storage cost and computational cost in the inference stage. For the stronger baseline ViT-BoT_{s = 12}, we can observe similar conclusions, and the JPM module improve the performance by +2.1% mAP and +1.0% mAP on MSMT17 and VeRi-776, respectively.

4.6. Ablation Study of TransReID

Finally, we evaluate the benefits of introducing SIE and JPM in Table 5. For the ViT-BoT, SIE and JPM improve the performance by +1.4%/+2.6% mAP and +1.4%/+1.0% 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-

Method	Size	MSMT17		Market-1501		DukeMTMC-reID		Occluded-Duke		Method	Size	VeRi-776		VehicleID	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1			mAP	R1	R1	R5
CBN [⊙] [51]	256×128	42.9	72.8	77.3	91.3	67.3	82.5	-	-	PRReID[11]	256×256	72.5	93.3	72.6	88.6
OSNet [49]	256×128	52.9	78.7	84.9	94.8	73.5	88.6	-	-	SAN[28]	256×256	72.5	93.3	79.7	94.3
MGN [37]	384×128	52.1	76.9	86.9	95.7	78.4	88.7	-	-	UMTS [14]	256×256	75.9	95.8	80.9	87.0
RGA-SC [44]	256×128	57.5	80.3	88.4	96.1	-	-	-	-	VANet [⊙] [5]	224×224	66.3	89.8	83.3	96.0
ABDNet [4]	384×128	60.8	82.3	88.3	95.6	78.6	89.0	-	-	PVEN [⊙] [25]	256×256	79.5	95.6	84.7	97.0
PGFA [26]	256×128	-	-	76.8	91.2	65.5	82.6	37.3	51.4	SAVER [16]	256×256	79.6	96.4	79.9	95.2
HOREID [36]	256×128	-	-	84.9	94.2	75.6	86.9	43.8	55.1	CFVMNet [33]	256×256	77.1	95.3	81.4	94.1
TransReID [⊙]	256×128	64.9	83.3	88.2	95.0	80.6	89.6	55.7	64.2	TransReID [⊙]	256×256	79.6	97.0	83.6	97.1
TransReID ^{*⊙}	256×128	67.4	85.3	88.9	95.2	82.0	90.7	59.2	66.4	TransReID [⊙]	256×256	80.6	96.9	-	-
TransReID [⊙]	384×128	66.6	84.6	88.8	95.0	81.8	90.4	57.2	64.0	TransReID ^{*⊙}	256×256	80.5	96.8	85.2	97.5
TransReID ^{*⊙}	384×128	69.4	86.2	89.5	95.2	82.6	90.7	59.4	66.7	TransReID ^{*⊙}	256×256	81.7	97.1	-	-

Table 6: Comparison with state-of-the-art methods. The star * in the superscript means the backbone is ViT-BoT_{s=12}. 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 only used wherever available. Best results for previous methods and best of our methods are labeled in bold.

776, respectively. For the ViT-BoT_{s=12}, SIE and JPM can also provide similar performance boost, i.e. 67.4% (+3.0%) mAP and 81.7% (+2.6%) mAP on MSMT17 and VeRi-776, respectively. The experimental results shows the effectiveness of our proposed module SIE and JPM, and the overall framework.

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. We compare TransReID with other methods on MSMT17, Market-1501 and DukeMTMC-reID. Since the image resolution is a critical factor for model performance, we evaluate our method with two different resolutions. On MSMT17 and DukeMTMC-reID, TransReID^{*} outperforms the previous state-of-the-art ABDNet by a large margin (+8.6%/+4.0% mAP). On Market-1501, TransReID^{*} (256×128) achieves comparable performance with state-of-the-art methods especially on mAP. Our method also shows superiority when comparing with methods which also integrate camera information like CBN.

Occluded ReID. Compared to PGFA and HOREID, TransReID achieves 55.7% mAP with a large margin improvement (+11.9% mAP) on Occluded-Duke, without requiring any semantic information to align body parts, which shows the ability of TransReID to generate robust feature representations. Furthermore, TransReID^{*} improves the performance to 59.2% mAP with the help of overlapping patches.

Vehicle ReID. On VeRi-776, TransReID^{*} reaches 81.7% mAP surpassing SAVER by 2.1% mAP. When only using viewpoint annotations, TransReID^{*} still outperforms VANet and SAVER on both VeRi-776 and

VehicleID. Additionally, our method achieves state-of-the-art performance to 85.2% mAP on a larger dataset VehicleID.

5. Conclusion

In this paper, we investigate a pure transformer framework for the object ReID task. A CNN-based baseline BoT is extended to ViT-BoT with several adaption. ViT-BoT achieves comparable performance on both person and vehicle ReID benchmarks. Based on ViT-BoT, we proposed two novel modules to the Transformer framework, i.e., side information embedding (SIE) and jigsaw patch module (JPM). Experiments conducted on MSMT17, Market-1501, DukeMTMC-reID, Occluded-Duke, VeRi-776 and VehicleID with various settings validate the effectiveness of our framework TransReID. The proposed TransReID achieves state-of-the-arts on all above six benchmarks.

Even though ViT just opens the door for pure transformer based model on image classification, the promising results achieved by TransReID make us believe that the transformer has great potential for ReID. It is expected that the ViT-BoT or TransReID could be as a starting point for more research works dedicated to the transformer-based ReID framework. In the future, we plan to explore a more efficient transformer-based framework for vision tasks, especially on the trade-off of representation capability and computational cost.

References

- [1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 3

- [2] 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. *arXiv preprint arXiv:2012.00364*, 2020. 1, 3
- [3] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 1
- [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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8351–8361, 2019. 8
- [5] Ruihang Chu, Yifan Sun, Yadong Li, Zheng Liu, Chi Zhang, and Yichen Wei. Vehicle re-identification with viewpoint-aware metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8282–8291, 2019. 2, 8
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3
- [8] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019. 3
- [9] 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. 1, 3
- [10] Liang Han, Pichao Wang, Zhaozheng Yin, Fan Wang, and Hao Li. Exploiting better feature aggregation for video object detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1469–1477, 2020. 1
- [11] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3997–4005, 2019. 8
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2
- [13] Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Convbert: Improving bert with span-based dynamic convolution. *Advances in Neural Information Processing Systems*, 33, 2020. 5
- [14] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11165–11172, 2020. 8
- [15] 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. 1, 3
- [16] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa. The devil is in the details: Self-supervised attention for vehicle re-identification. In *European Conference on Computer Vision*, pages 369–386. Springer, 2020. 1, 8
- [17] Xiangyu Li, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. Transformer guided geometry model for flow-based unsupervised visual odometry. *Neural Computing and Applications*, pages 1–12, 2021. 1
- [18] Xiangyu Li, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Transactions on Cognitive and Developmental Systems*, pages 1–7, 2021. 1
- [19] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017. 2
- [20] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2167–2175, 2016. 5
- [21] Xinchun Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016. 5
- [22] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 907–915, 2020. 2
- [23] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2, 4
- [24] 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
- [25] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2020. 2, 8
- [26] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 542–551, 2019. 1, 2, 5, 8
- [27] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran.

- Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018. 1
- [28] 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. 2, 8
- [29] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 5
- [30] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 1, 3
- [31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 1, 2
- [32] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 1, 2
- [33] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3523–3531, 2020. 8
- [34] 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. 1, 3
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017. 3
- [36] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6449–6458, 2020. 8
- [37] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 1, 2, 8
- [38] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 5
- [39] 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. 2
- [40] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. In *International Conference on Learning Representations*, 2019. 5
- [41] Enze Xie, Wenjia Wang, Wenhui Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer. *arXiv preprint arXiv:2101.08461*, 2021. 1, 3
- [42] 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. 1, 6
- [43] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 5
- [44] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3186–3195, 2020. 8
- [45] Liang Zheng, Liye Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 5
- [46] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840*, 2020. 2, 3
- [47] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2018. 2
- [48] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020. 5
- [49] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712, 2019. 8
- [50] Zhihui Zhu, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Xing Sun, and Weishi Zheng. Aware loss with angular regularization for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13114–13121, 2020. 2
- [51] 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 *European Conference on Computer Vision*, pages 140–157. Springer, 2020. 2, 8