TransReID: Transformer-based Object Re-Identification (Supplementary Material)

Shuting He[1,2*] Hao Luo[2] Pichao Wang[2] Fan Wang[2] Hao Li[2] Wei Jiang[1†]
[1]Zhejiang University [2]Alibaba Group
{shuting_he,jiangwei_zju}@zju.edu.cn {michuan.lh,pichao.wang,fan.w,lihao.lh}@alibaba-inc.com

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 1 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 [6]. 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 [9] to transformer-based models will cause a significant drop in performance. AdamW [8] 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 [7] 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 [4, 5]. 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 1. 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 2 and Table 3 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 2, 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 3. (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 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

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

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 1: 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 [7], 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 2: 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 3: 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.

has a strong correlation with the corresponding patch. ViT-FRCNN [2] 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,

(a) Original **(b) w/o rearrange** **(c) w/ rearrange**

Figure 1: 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.

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 $[f_l^1, f_l^2, ..., f_l^k]$ ($k = 4$ in our

cases) by JPM module in Figure 1. Brighter region means higher corresponding weights. Several observations can be made from Figure 1: (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 sub-sequences. 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. Performances of Light-weight Transformer-based and CNN-based Methods

In order to further verify the effectiveness of the proposed method, we compare commonly used light-weight CNN backbones in ReID practical applications, such as e.g., moblenet-v3, shufflenet to pure transformer with comparable parameters in Table 4. we conduct light-weight CNN-based methods based on BoT and DeiT-Tiny (ImageNet-1K pretrained) based on our transformer-based baseline with fair experimental settings. "DeiT-Tiny" shows better performance compared to the other two light-weight CNN-based methods with comparable parameters and speed (inference time per image).

Backbone	#params	Speed	MSMT17		DukeMTMC	
			mAP	R1	mAP	R1
DeiT-Tiny	5.72M	9.6ms	45.8	68.5	72.2	85.3
MobileNet-v3	5.48M	10.1ms	39.4	64.4	67.4	81.1
ShuffleNet-v2	7.40M	9.2ms	42.4	66.5	71.0	84.1

Table 4: Comparison of light-weight transformer-based and CNN-based methods.

D. Comparisons of Model Complexity

We analyze model complexity between TransReID and MGN[11] in the main paper. Here, we give more detailed experimental results in Table 5. TransReID* (ViT/16(s=12)) has comparable parameters with some powerful approaches with ResNet50 backbone (*e.g*. MGN[11], ABD-Net[3] and DSA-reID[1]). TransReID achieves a better trade-off between accuracy and speed under the same settings.

E. 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 2 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.

Method	#params	Speed	MSMT17		DukeMTMC	
			mAP	R1	mAP	R1
TransReID*	102.9M	15.6ms	69.4	86.2	82.6	90.7
MGN	68.8M	16.2ms	-	-	78.4	88.7
ABD-Net	69.2M	14.9ms	60.8	82.3	78.6	89.0
DSA-reID	187.8M	34.7ms	-	-	74.3	86.2

Table 5: Comparisons of model complexity among different state-of-the-art methods.

References

[1] Densely semantically aligned person re-identification. In *CVPR*, 2019. 3

[2] 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. 2

[3] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, 2019. 3

[4] Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*, 2019. 1

[5] 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. 1

[6] Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. *arXiv preprint arXiv:1803.01719*, 2018. 1

[7] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661. Springer, 2016. 1, 2

[8] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 1

[9] 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

[10] 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. 4

[11] 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. 3

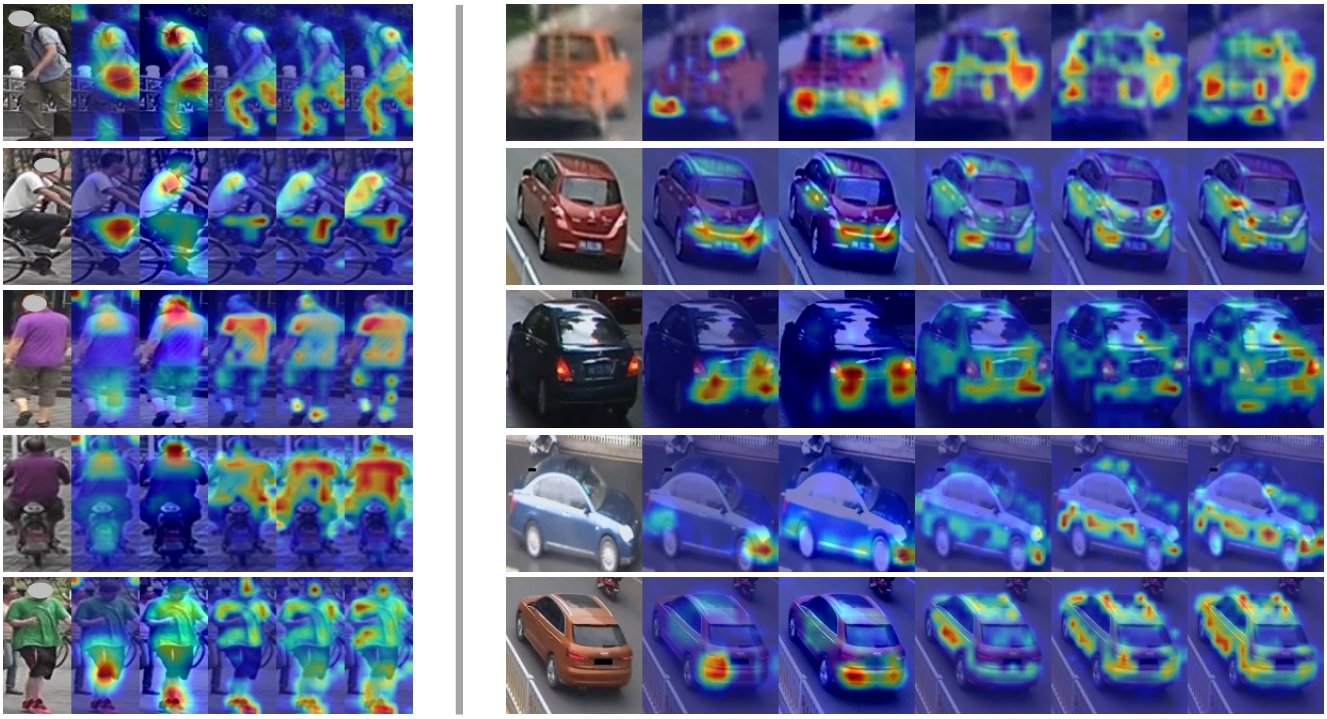

Figure 2: Grad-CAM [10] visualization of attention maps. (a) Original images, (b) CNN-based methods, (c) CNN+Attention methods, (d) Transformer-based baseline, (e) TransReID w/o rearrange, (f) TransReID. Faces in the images are masked for anonymization.