6. Appendix

In the appendix, we give the details of datasets used in our experiments and show additional experimental results.

6.1. Details of Datasets

Table 5 shows the details of the datasets used in our experiments. Note that all the models in our experiments are trained on the Pitts30k-train dataset and tested on the other datasets.

Visualization of Feature Embeddings. We visualize the feature embeddings of some images in the Pitts30k dataset computed by our TransVLAD with Vgg-16 in 2-D using the t-SNE method. As shown in Figure 8, images taken from the same places are mostly embedded in nearby 2D positions although their lighting and perspective are different.

Table 5. Details of datasets used in our experiments.

Dataset	Gallery Images	Query Images
Pitts30k-train	10,000	7,416
Pitts30k-val	10,000	7,068
Pitts30k-test	10,000	6,816
Pitts250k-train	91,464	7,824
Pitts250k-val	78,648	7,608
Pitts250k-test	83,952	8,280
TokyoTM-val	49,056	7,186
Tokyo 24/7	75,984	315
Oxford 5k	5,063	55
Paris 6k	6,412	220
Holidays	991	500

6.2. Additional Results with Different CNN Backbones

To verify whether the deeper CNN backbones can further improve the performance of our TransVLAD with CNN backbones on geo-localization datasets, we maintain the same TransVLAD module and choose Vgg-19, ResNet-101 and ResNet-152 as the CNN backbones for the extra training and compare with the existing results on the geo-localization benchmarks.

We plot the Precision-Recall curves for each CNN backbone in Figure 9 and report the detailed comparison of recalls at N top retrievals in Table 6. From the results, we can observe our models with deeper CNN backbones have less than 1% improvement on most geo-localization datasets. On the challenging Tokyo 24/7 dataset, the improvements by utilizing deeper CNN backbones are more obvious. For example, our TransVLAD with ResNet-152 achieves 91.7% rank-5 recall, up to 1.9% accuracy improvement against our TransVLAD with Vgg-16. By analyzing the performance of our TransVLAD with different CNN backbones, it can be concluded that deeper CNN backbones can indeed improve the generalization ability at the cost of heavy model complexity. However, the improvement of model accuracy by adopting deeper CNN backbones is limited.

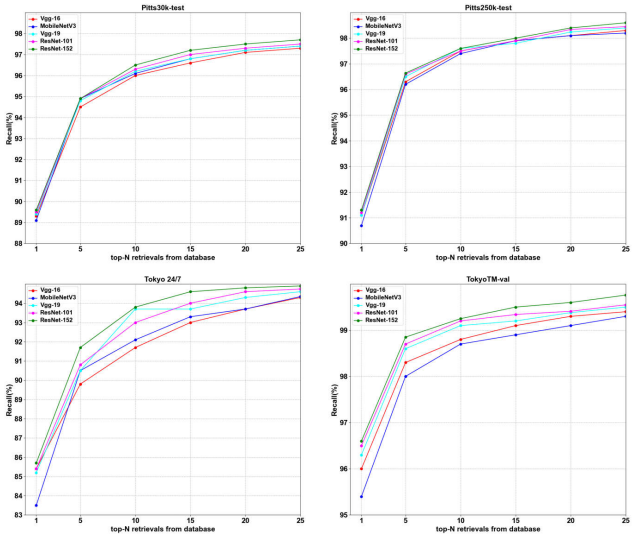


Figure 9. Comparison of recalls at N top retrievals from database with different CNN backbones.

6.3. Comparison of the Comprehensive Performance with MobileNetV3

Besides the accuracy of our proposed model, we also consider the running speed and the storage size in practical geo-localization tasks. We further evaluate the comprehensive performance for our TransVLAD with MobileNetV3 on the Pitts30k-test dataset by using different dimensions of cluster centers (c-dim) and dimensions of output vectors (o-dim). The running speed is tested on the whole Pitts30k-test dataset on a single GeForce GTX 1080 GPU and the Intel Xeon E5-2620 v4 @ 2.10GHz CPU. The storage size consists of the parameters of the model and the feature vectors of the gallery images in the Pitts30k-test dataset.

Table 7. Comparison of the comprehensive performance for our TransVLAD with MobileNetV3 on the Pitts30k-test dataset.

Method	Model Setting		Pitts30k-test				
	c-dim	o-dim	R@1	R@5	R@10	time(s)	storage(MB)
Our-MobileNetV3	1024	4096	**89.1**	**94.9**	**96.1**	511.3	1266.7
	512	4096	**89.1**	94.2	95.8	484.2	742.5
	256	4096	88.6	94.3	95.9	464.6	480.4
	256	2048	88.3	94.3	95.8	436.4	270.3
	256	1024	87.9	94.2	95.8	424.2	165.3
	256	512	87.4	93.9	95.6	414.6	112.8
	128	4096	88.2	94.4	95.8	357.6	339.5
	128	2048	88.0	94.4	95.8	343.9	196.5
	128	1024	87.7	94.1	95.7	335.8	125
	128	512	87.4	94.0	95.7	**328.2**	**89.2**

From the results in Table 7, we can state that the running speed and storage size of our TransVLAD with MobileNetV3 can be greatly reduced by decreasing the c-dim and o-dim. More specifically, our TransVLAD with MobileNetV3 can achieve 328.2s running time and 89.2MB storage size on the Pitts30k-test dataset at the cost of 1.7% rank-1 Recalls compared to the best accuracy.

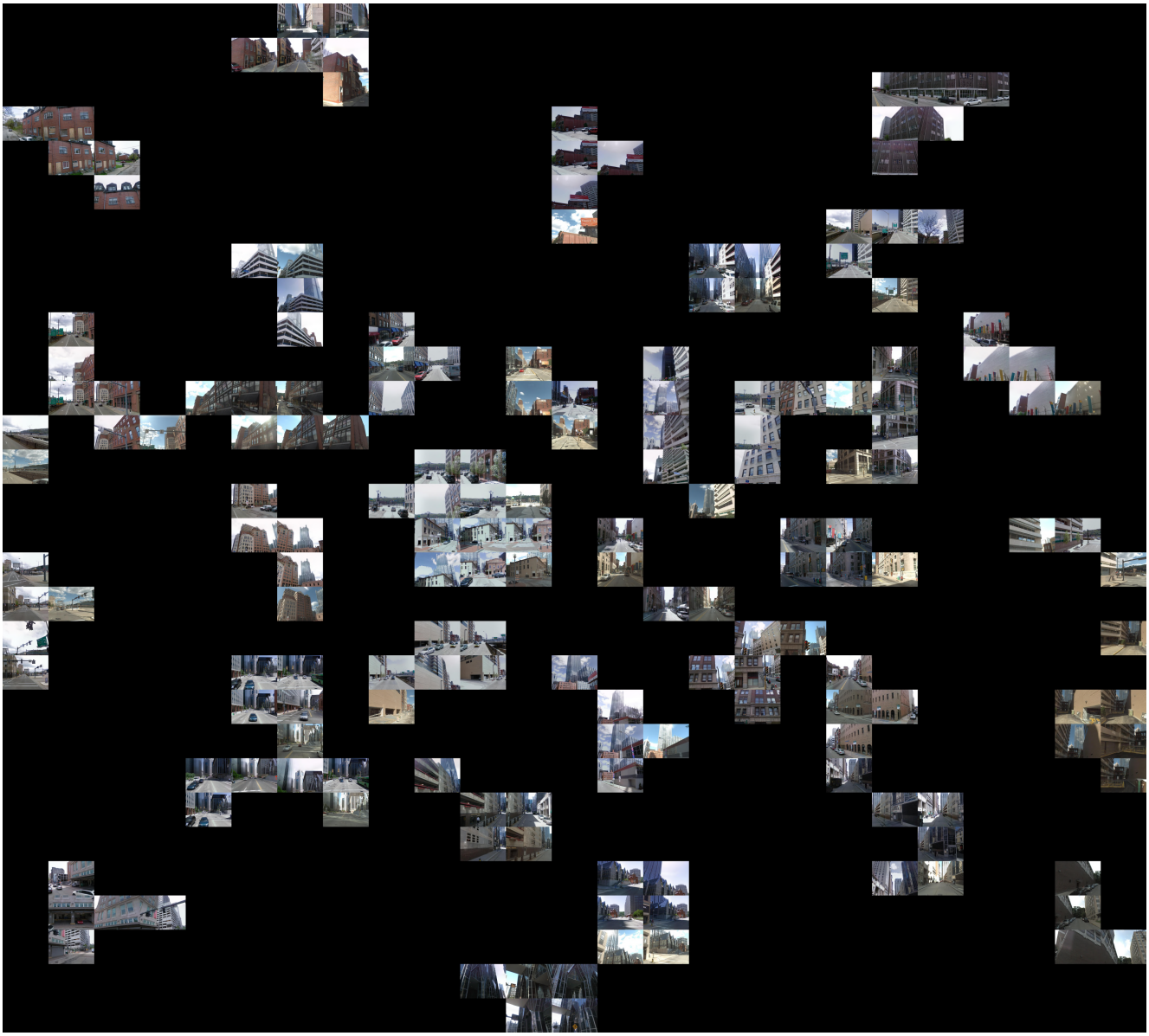


Figure 8. Visualization of feature embeddings computed by our TransVLAD with Vgg-16 using t-SNE on the Pitts30k dataset (part of images).

Table 6. Comparison of Recalls with different CNN backbones on the Pitts30k/250k-test, TokyoTM-val, Tokyo 24/7 datasets.

Method	Pitts30k-test				Pitts250k-test				TokyoTM-val				Tokyo 24/7			
	R@1	R@5	R@10	R@15	R@1	R@5	R@10	R@15	R@1	R@5	R@10	R@15	R@1	R@5	R@10	R@15
Our-Vgg16	89.3	94.5	96.0	96.6	91.1	96.3	97.5	97.9	96.0	98.3	98.8	99.1	85.4	89.8	91.7	93.0
Our-MobileNetV3	89.1	**94.9**	96.1	96.8	90.7	96.2	97.4	97.9	95.4	98.0	98.7	98.9	83.5	90.5	92.1	93.3
Our-Vgg19	89.4	94.8	96.2	96.8	91.1	96.5	**97.6**	97.8	96.3	98.6	99.1	99.2	85.2	90.5	93.7	93.7
Our-ResNet101	89.5	**94.9**	96.3	97.0	91.2	**96.6**	97.5	97.9	96.5	98.7	**99.2**	99.3	85.4	90.8	93.0	94.0
Our-ResNet152	**89.6**	**94.9**	**96.5**	**97.2**	**91.3**	**96.6**	**97.6**	**98.0**	**96.6**	**98.8**	**99.2**	**99.5**	**85.7**	**91.7**	**93.8**	**94.6**

6.4. Additional Qualitative Evaluation

To better demonstrate the performance of our TransVLAD with CNN backbones on the geo-localization task, we visualize additional attention maps of query images by our model with VGG-16, SARE and NetVLAD on both Pitts30k-test and challenging Tokyo 24/7 datasets. For generating the attention maps, we use the feature maps before

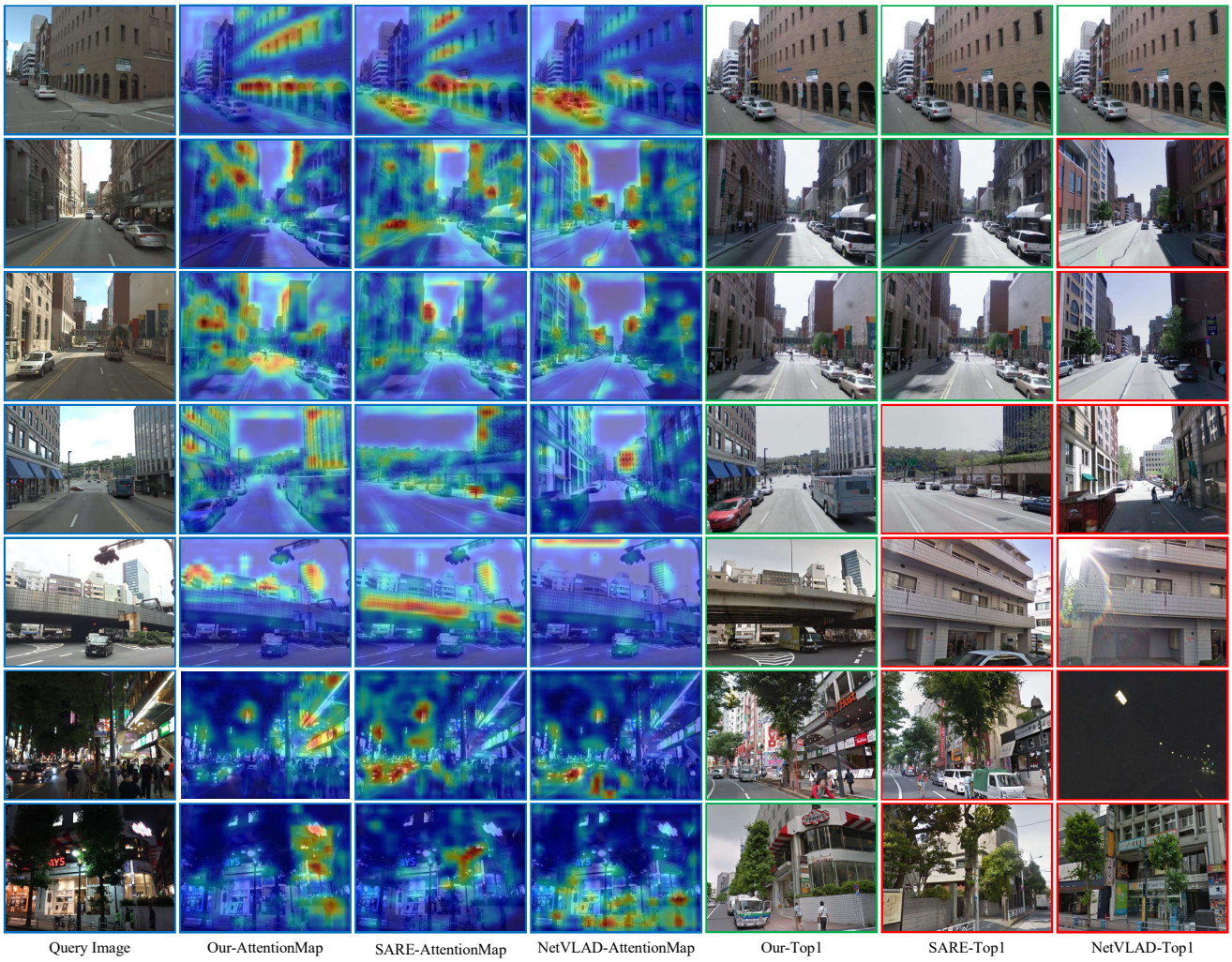
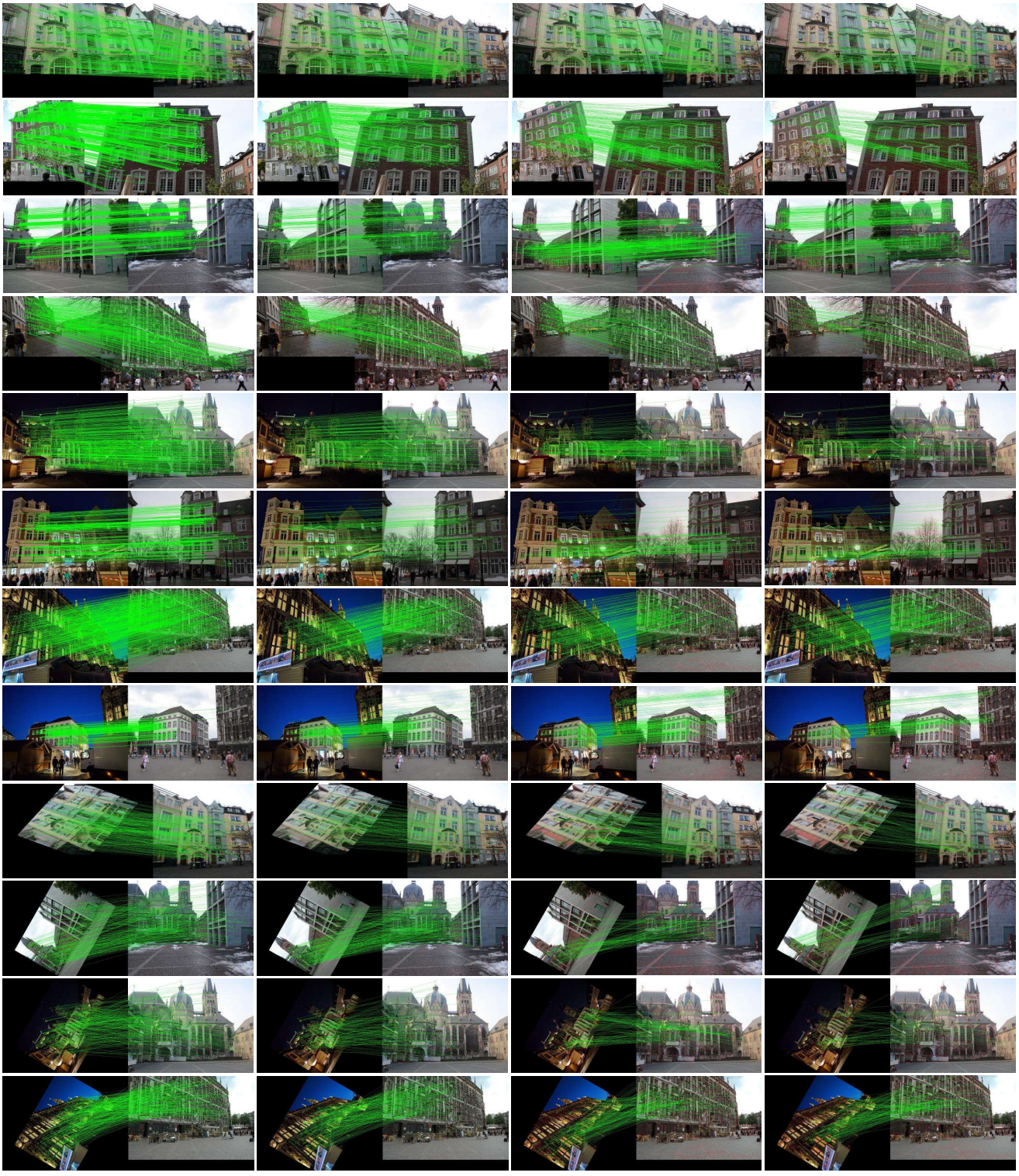


Figure 10. Visualized additional attention maps. The attention maps of query images are generated to show the regions where the models focus. We compare the attention maps and the top-1 retrieved gallery images. Green and red borders indicate correct and incorrect retrieval results, respectively. (Best viewed in color.)

the VLAD layer. From the results in Figure 10, we can observe our model with VGG-16 focuses on the discriminative landmarks (e.g. buildings, signs) due to the global reasoning enhanced by our sparse transformer module, while the other two models incorrectly focus on changeable objects (e.g. trees, cars, pedestrians and light). The misdirection by changeable objects will result in false retrieval results, since the objects may shift or vanish from the right gallery images, or appear in the incorrect gallery images.

6.5. Generalization on Detecting and Matching Keypoints

To further verify the performance and generalization capability of our TransVLAD with CNN backbones, we adopt DFM model and replace the VGG backbone with our network trained on Pitts30k-train dataset for feature detection. We further estimate matching results on both original and rotated image pairs by detecting and matching keypoints with deep CNNs. From the matching results shown in Figure 11, we can observe the DFM with our backbone generates more dense correct matches than the other three models on both original and rotated image pairs, which can attest to the stronger generalization ability of our network.



(a) DFM with our backbone (b) DFM (c) GIFT (d) Superpoint

Figure 11. Qualitative matching results with four methods. The correct matches are drawn in green lines.