Attention-Based Spatial Guidance for Image-to-Image Translation - Supplementary

Yu Lin, Yigong Wang, Yifan Li, Yang Gao, Zhuoyi Wang, Latifur Khan
University of Texas at Dallas
800 W. Campbell Road, Richardson, Texas
yxl163430,yxw158830,yli,yxg122530,zxw151030,lkhan@utdallas.edu

1. Overview

In this supplementary, we first present the network architecture in Sec 2. In Sec 3, we studied the impact of different attention fusion stages. The qualitative result of the *Cityscape* dataset is available in Sec 4. We justify our hypothesis why the improvement over supervised *Cityscape* is limited in Sec 5. More qualitative result are provided in the remaining sections.

2. Architecture and Implementation

To be comparable with previous methods [3, 4, 8], we use 256×256 images for the *unsupervised Cityscape translation* and all object and scenery tasks, and 128×128 images for the *supervised Cityscape translation*. We applied our method on the CycleGAN [8] for all unsupervised translation tasks. In other words, we adopt the network architecture of CycleGAN as the backbone of our proposed model. In specific, we adopted the ResNet 9-blocks [2] generator and the PatchGAN [4] discriminator. This generator contains 2 down-sampling blocks, 9 residual blocks and 2 upsampling blocks. For the supervised translation, we adopted the UNet-128 [6] generator and a same PatchGAN discriminator. The PatchGAN discriminator is composed of 5 convolution layers, including normalization and ReLU layers.

Before diving into the detail of our modified discriminator, let us first describe TAM's 2-branch architecture [7] in detail. They built a very deep network with numbers of *attention block*s. Each *attention block* contains two branches: mask branch and trunk branch. Mask branch cascades the input features through a bottom-up top-down architecture that mimics human attention. Trunk branch is applied as feature processing. To build a TAM discriminator with this 2-branch architecture, we replaced the *ResBlock* by a simple convolution layer. Please note that more parameters commonly means more powerful network and the discriminator is already too strong comparing to the generator, so we have to deduce the capacity of the discriminator. In such TAM discriminator, we use the first convolution layer as feature extractor, three consecutive convolution layers for trunk branch and the last one convolution layer for classification. The mask branch is composed of two downsampling layers, two convolution layers and one upsampling layer.

Similar to prior works, we applied Instance Normalization (IN) for both generators and discriminators. In the preprocessing step, we resized the input image to 286×286 (143×143) then randomly cropping back to 256×256 (128×128). For all the unsupervised experiments, we set the weight factor of the GAN loss to 1, $\lambda_{GAN} = 1$, and the weight factor of cycle consistency to 10, $\lambda_{Cyc} = 10$, in our objective. On the other hands, we set the weight factor of the GAN loss to 1, $\lambda_{GAN} = 1$, and the weight of L1 loss to 10, $\lambda_{L1} = 100$.

Method	F0	F1	F2	F3	F4
$H \rightarrow Z$	**1.03 ± 0.35**	1.07 ± 0.41	1.29 ± 0.51	1.47 ± 0.61	1.04 ± 0.63
$Z \rightarrow H$	**3.42 ± 0.51**	3.42 ± 0.65	3.46 ± 0.60	3.88 ± 0.66	3.63 ± 0.68

Table 1. Target KID for different fusion stages on *horse2zebra*. **F0**: early fusion; **F1**: fusion before the down-sampling; **F2**: fusion after the down-sampling; **F3**: fusion at the end of 4th residual block; **F4**: fusion at the end of 9th residual block;

3. Attention Fusion Strategies

Noted that we blend the attention mask and the raw input before feed them into the generator. However, in the community, late-fusion seems to be the preferred choice (blend the attention map and input feature in the hidden space). In this section, we provide an ablation study for where to place the attention.

The result is presented in Table 1. It's easy to see that the Target FID score (Since we are doing an object translation task) doesn't change much when we plugin the attention map elsewhere. Thus we chose early fusion in our experiments.

4. Qualitative Results for *Cityscape*

The qualitative result of unsupervised *Cityscape* translation is presented in Figure 1 and Figure 2. Noted that the classical GAN model [1] is suffered from model collapse, which means it maps all the input to a same image. Another interesting observation is that even FAL [3] generates sharp looking images, the content is not fully correct. For example, look at the first row in Figure 2, the car segmentation (blue region) covers too much area compared to the ground truth.

5. Hypothesis Justification

In this section, we empirically justify our hypothesis on why we have the improvement over per-class accuracy and IoU are marginal. We conduct two additional experiments and analyze their results. Firstly, we compute the per-class statistic across the whole dataset. Figure 3 shows the *statistic frequency* of each class, which denotes the number of images (in percentage) that contain at least one specific object. For example, the *building* appears almost every image while less than 5% of images contain *trains*.

In Figure 4, we further provide the per-class *average frequency* that describes the average number of specific objects that appear per image. For instance, the bar plot told us that an image contains about 10 cars and 2.5 buildings on average. Such statistic information justifies our second hypothesis that some classes merely appear in the dataset. Thus the further improvement over per-class accuracy and IoU are prohibited (*e.g.* bus and train).

We then compute the attention intensity of each class during training. The result is presented in Figure 5. The *attention intensity* is defined as the number of pixels that have attention larger than a given threshold α (We use $\alpha = 0.5$ in this experiment). In other words, we assume a pixel is crucial if its corresponding attention value is larger than α. We propose that a specific class is aware by the discriminator if at least 50% of its pixel is crucial. For example, a 128×128 image contains 512 *car* pixels, then *car* is attended if at least 256 pixels have attention value greater than α.

Based on the figure, we discover that the discriminator may focus on some small classes (*e.g.* rider and terrain). The per-class accuracy and IoU are affected if the generator tries too hard to fix those classes but ignore the major part. Based on the number of instances of each class, the contribution of generating good riders is significantly less than the contribution of generating good cars. This experiment justifies our first suggestion that only few classes are highlighted in the attention map.

6. Attention Map during Training

We present some intermediate results with its attention map in Figure 6, Figure 7 and Figure 8. The white area in the attention map indicates that region is important. Please note that the attention map focuses on the target object in the early training stage while it will focus on some small regions later. This scenario is consistent with the attention behavior in AGGAN [5].

7. More Translation Results

More translation results are provided in Figure 9, Figure 10, Figure 11 and Figure 12, 13.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] 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.

[3] Minyoung Huh, Shao-Hua Sun, and Ning Zhang. Feedback adversarial learning: Spatial feedback for improving generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1476–1485, 2019.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[5] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 3693–3703, 2018.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[7] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017.

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

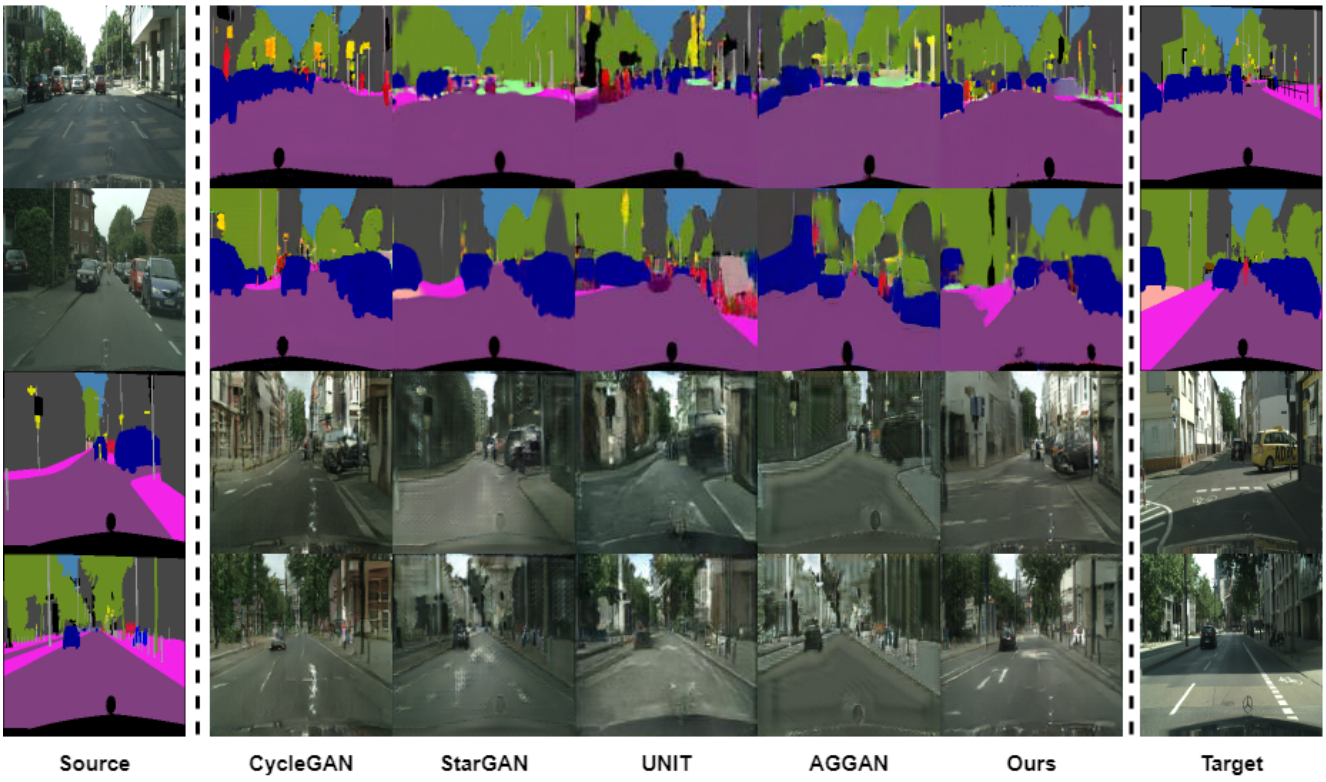

Figure 1. Different unsupervised translation methods for labels↔photos mapping, trained on *Cityscape* images.

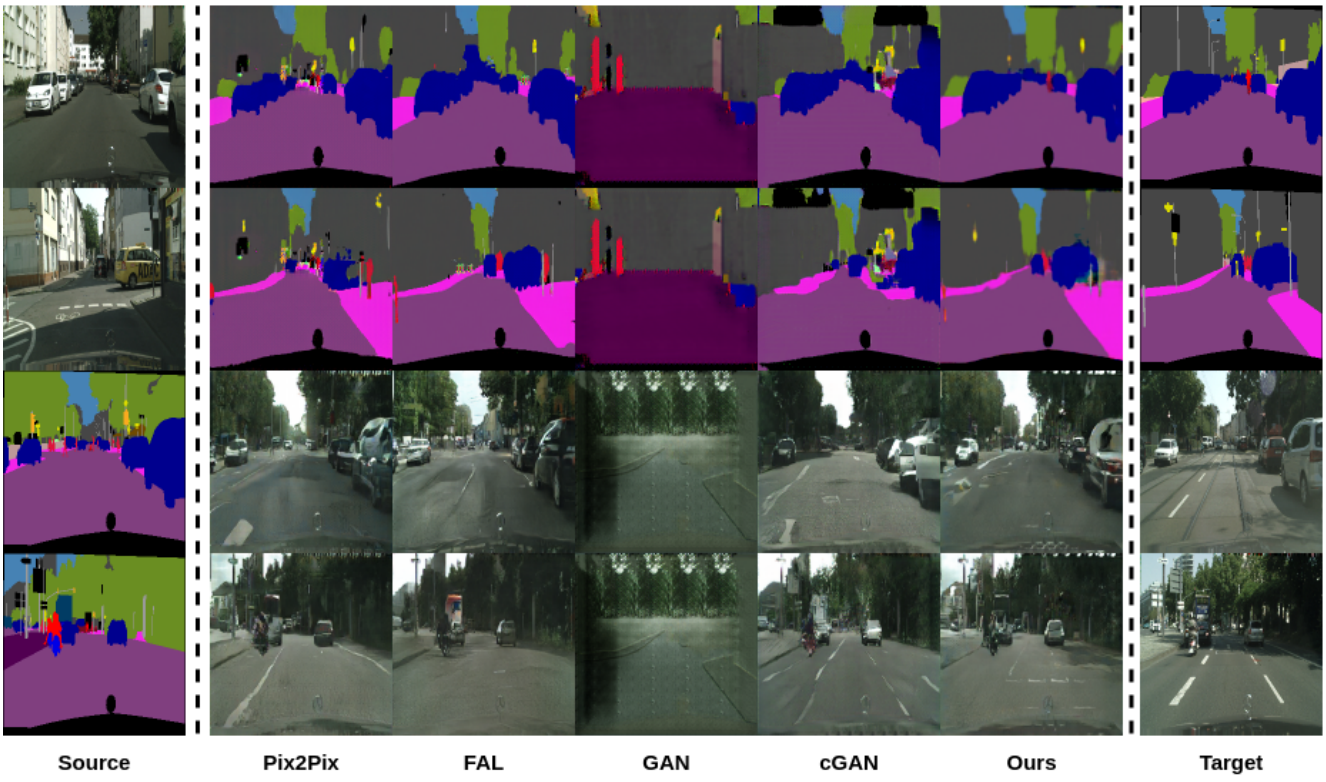

Figure 2. Different supervised translation methods for labels↔photos mapping, trained on *Cityscape* images.

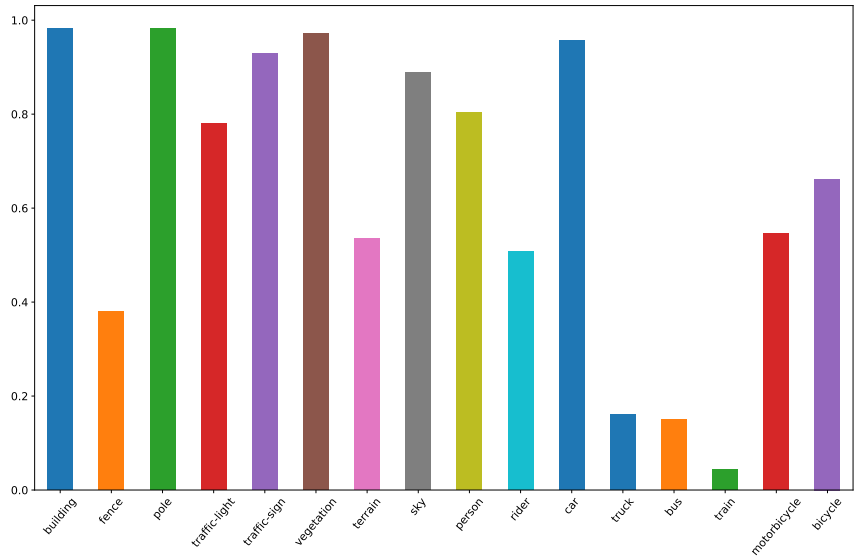

Figure 3. The *statistic frequency* for all 16 classes appeared in the *Cityscape* translation.

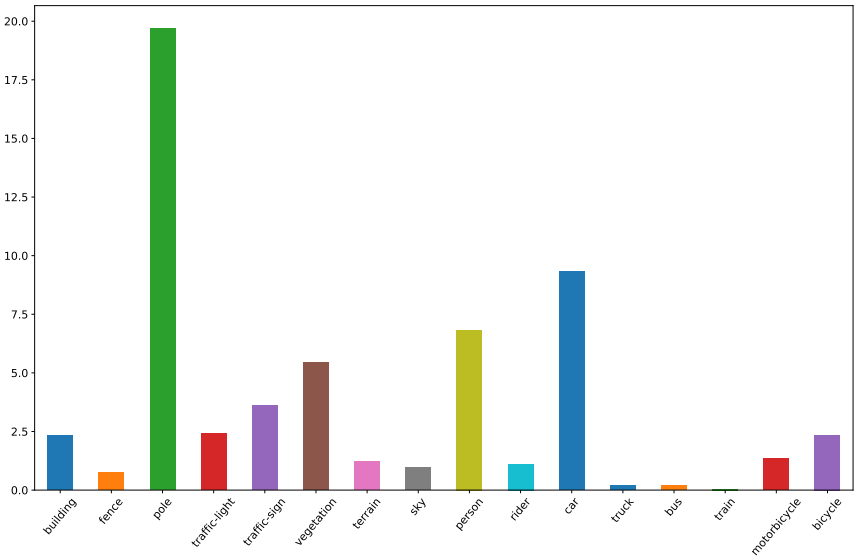

Figure 4. The *average frequency* for all 16 classes appeared in the *Cityscape* translation.

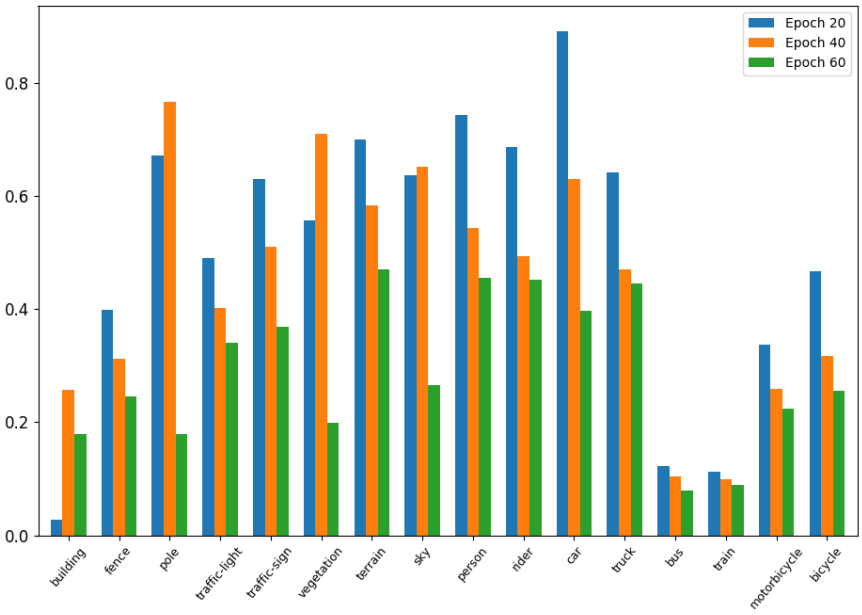

Figure 5. The average per-class attention intensity during training, in epochs 20, 40, 60.

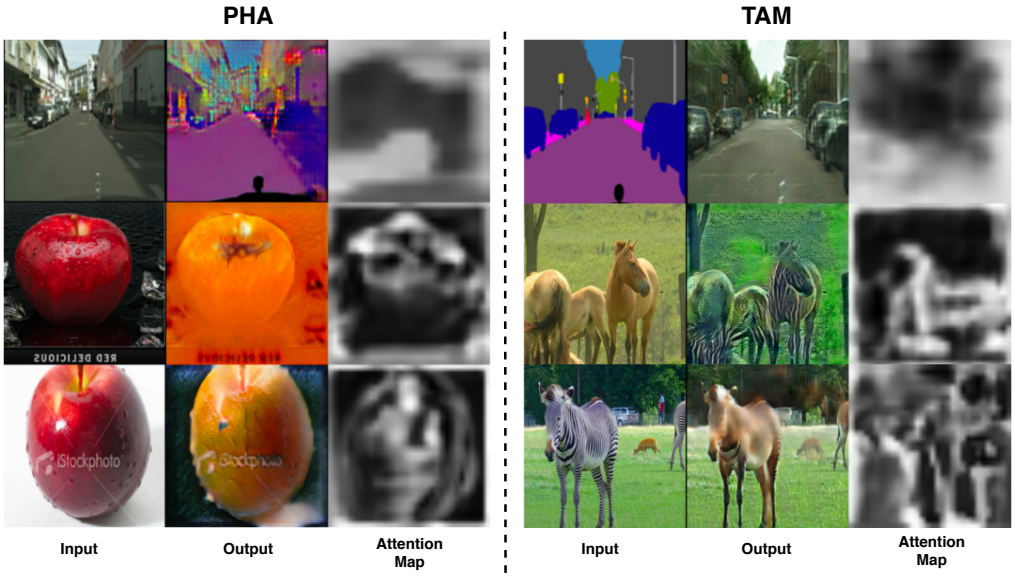

Figure 6. Inputs, outputs and corresponding attention maps at training epoch 10. Left: attention map generated by PHA; Right: attention map generated by TAM.

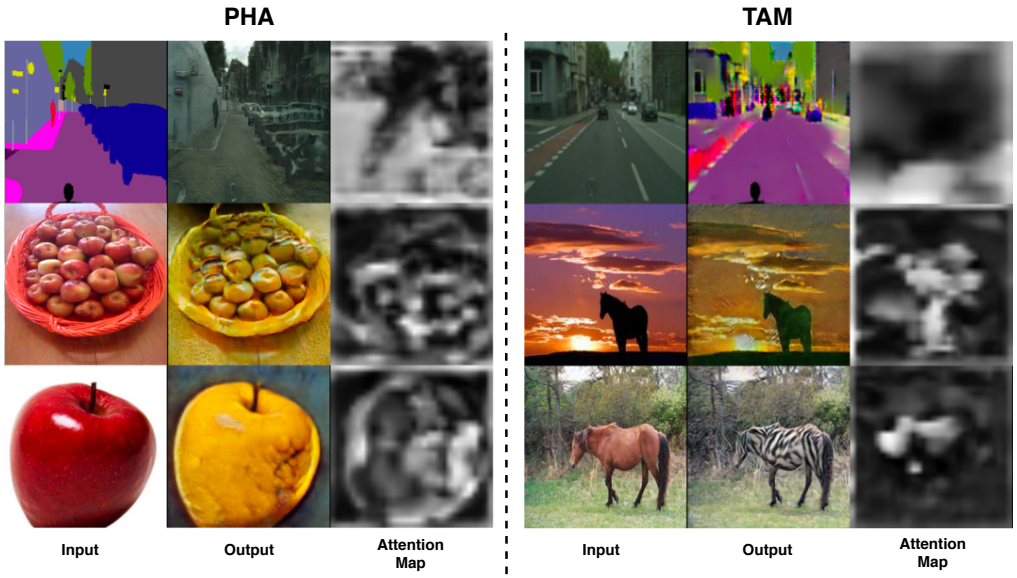

Figure 7. Inputs, outputs and corresponding attention maps at training epoch 50. Left: attention map generated by PHA; Right: attention map generated by TAM.

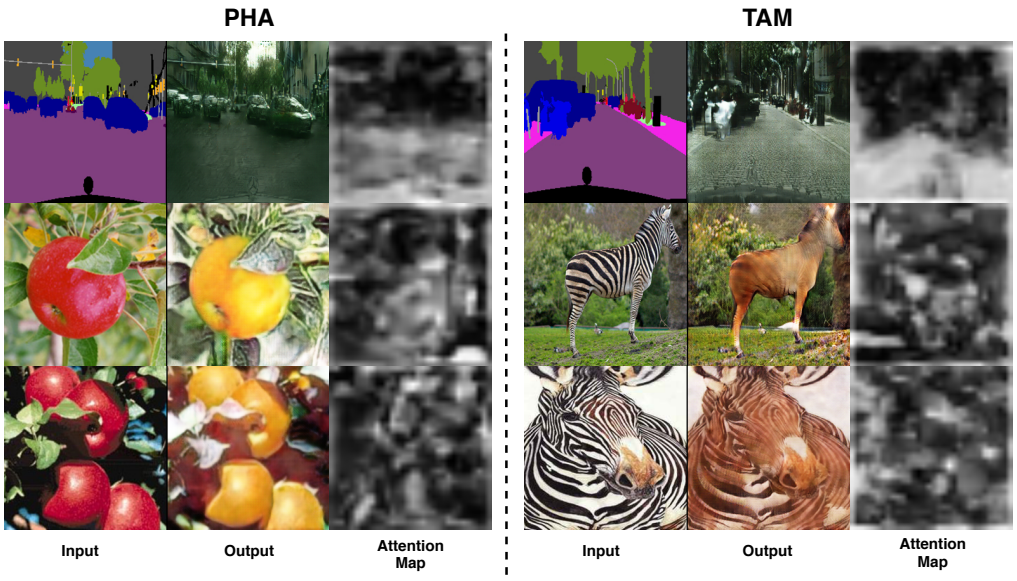

Figure 8. Inputs, outputs and corresponding attention maps at training epoch 100. Left: attention map generated by PHA; Right: attention map generated by TAM.

Figure 9. Image-to-Image translation results generated by different approaches on object translation and scenery translation. Every two rows from top: *apple↔orange*, *zebra↔horse*, *night↔day* and *winter↔summer*. More result is available in the supplementary

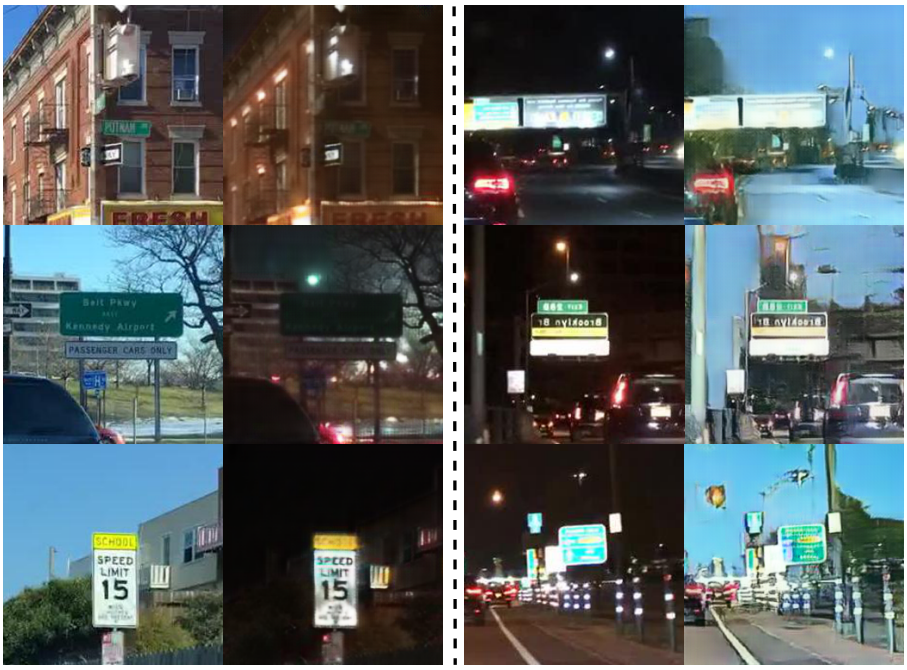

Figure 10. Additional translation results on *day2night* dataset. From left to right: real daytime images, fake night images, real night images, fake daytime images.

Figure 11. Additional translation results on *apple2orange* dataset. From left to right: real apple images, fake orange images, real orange images, fake apple images.

Figure 12. Additional translation results on *horse2zebra* dataset. From left to right: real horse images, fake zebra images, real zebra images, fake horse images.

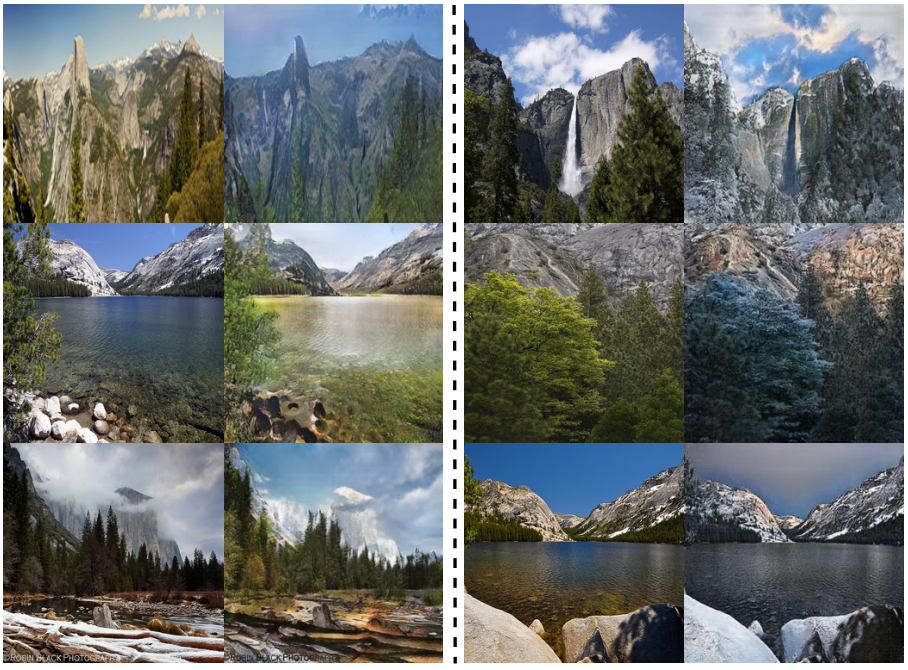

Figure 13. Additional translation results on *winter2summer* dataset. From left to right: real winter images, fake summer images, real summer images, fake winter images.