# Learning to Cartoonize Using White-box Cartoon Representations Supplementary materials

Xinrui Wang[1,2,3] Jinze Yu[2]
[1]ByteDance, [2]The University of Tokyo, [3]Style2Paints Research

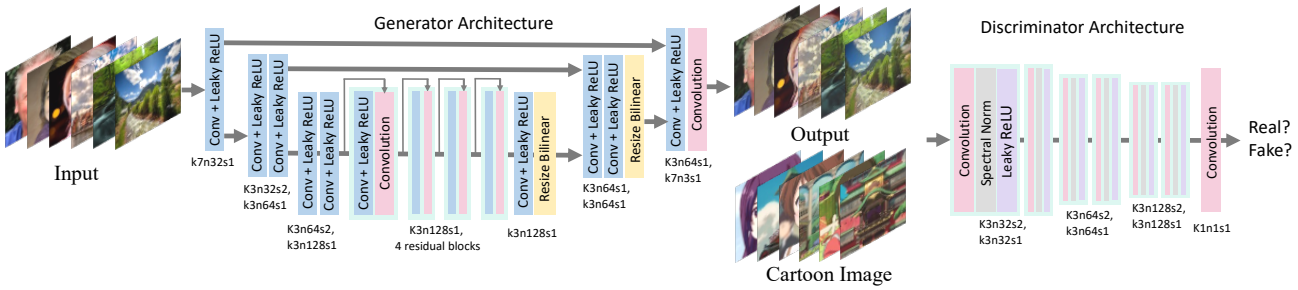


Figure 1: The architecture of generator network and discriminator network.

In this supplementary material, we show more experimental results, including the architecture of generator network and discriminator network, results of our method in different use cases, comparison between our results and cartoons in the same scenes, and examples used in the user study. The inference code with pre-trained model and some test images are also submitted in the supplementary material for reproducibility.

## 1. Network Architecture

We show the architecture of generator network and discriminator network in the above figure. The generator network is a fully-convolutional U-Net-like [6] network. We use convolution layers with stride2 for down-sample and bilinear interpolation layers for upsample to avoid checkerboard artifacts. The network consists of only three kind of layers: convolution, Leaky ReLU (LReLU) [4] and bilinear-resize layers. This enables it to be easily embedded in edge devices such as mobile phones. PatchGAN [2] is adapted in the discriminator network, where the last layer is a convolution layer. Each pixel in the output feature map correspond to a patch in the input image, with the patch size equals to the perceptive field, and is used to judge whether the patch belongs to cartoon images or generated images. The PatchGAN enhances the discriminative ability on details and accelerates training. Spectral normalization [5] is placed after every convolution layer (except the last one) to enforce Lipschitz constrain on the network and stabilize training.

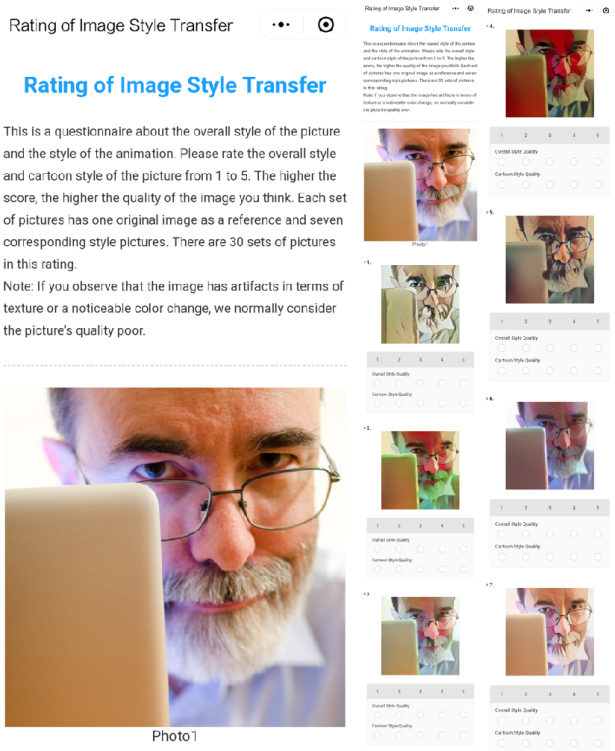


Figure 2: Smartphone application interface of our user study.

## 2. Results in Different Use Cases

In Section 4.1 of the main paper, we apply our method on different scenes and show the cartoonized results. Due to the limitation of space, Only 16 pairs of examples with small resolution are presented. Here we collect images from more use cases with higher resolution, and show the results generated by our method.

The contents of the collected images include male celebrities (shown in Figure 3), female celebrities (shown in Figure 4), animals (shown in Figure 5), plants (shown in Figure 6), food (shown in Figure 7), natural sceneries (shown in Figure 8 and Figure 9), city views (shown in Figure 10 and Figure 11), and other objects (shown in Figure 12). Overall, the presented results demonstrate that our method can generate high-quality cartoonized images, and can be applied on diverse use cases and real-world scenes. All the images of person are from CelebA dataset [3], while the rest of images are from the DIV2K dataset [1]

## 3. Comparison with Cartoon Images

To further illustrate the diverse use cases in the real-world and the cartoonization quality of our method, we present the comparison between the results of our method and cartoon images in the same scene. We collect several cartoon images from Shinkai Makoto's films and their counterpart real-world photos taken in the same scenes. Our method is then applied on the collected real-world photos, and results of our method and cartoon images are shown in Figure 13 and Figure 14.

## References


[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 2

[2] 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. 1

[3] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2

[4] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of International Conference on Machine Learning*, volume 30, 2013. 1

[5] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 1

[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. 1

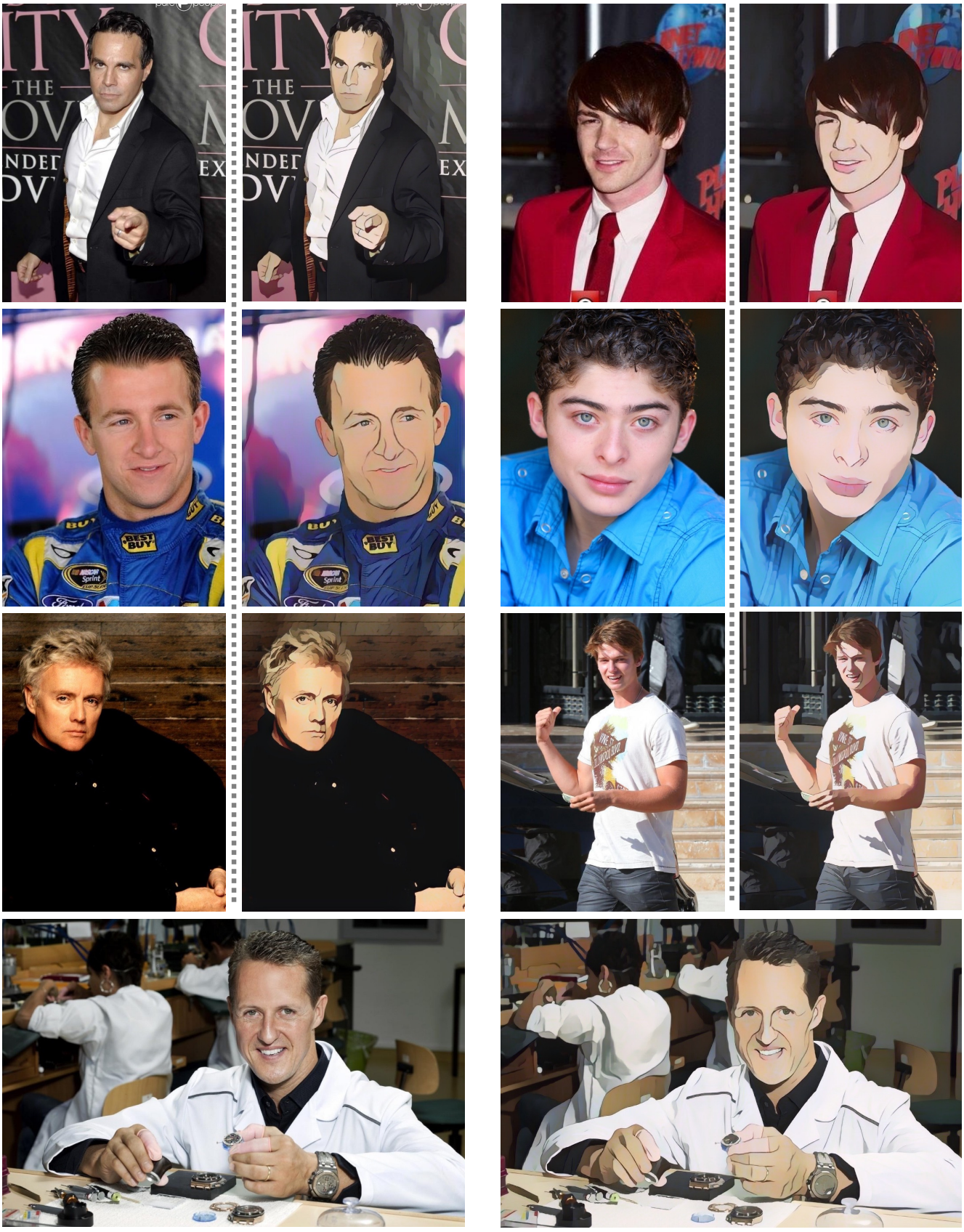


Figure 3: Cartoonized male Celebrities.

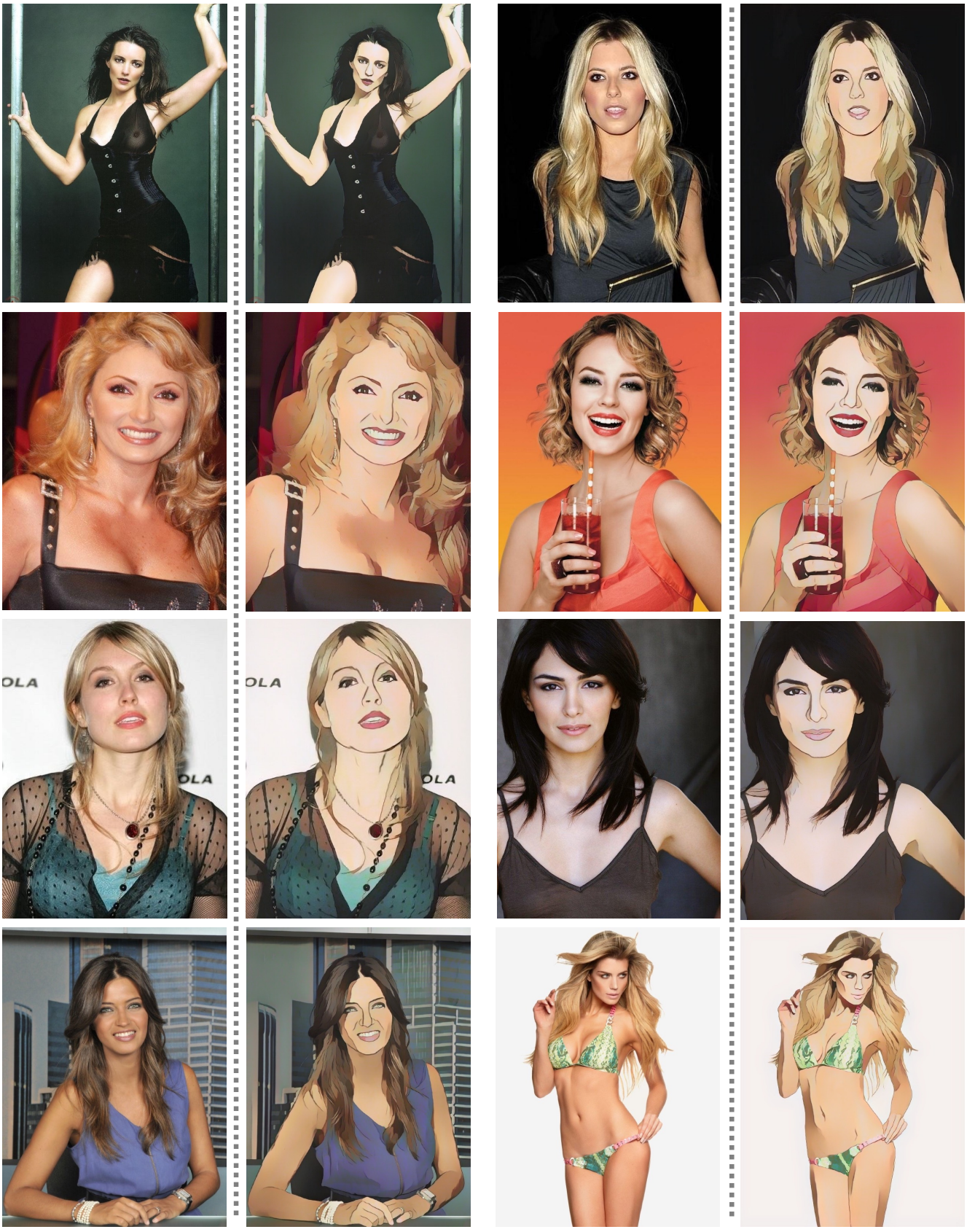


Figure 4: Cartoonized female Celebrities.


Figure 5: Cartoonized animals.


Figure 6: Cartoonized plants.

Real-world photo Cartoonized result

Figure 7: Cartoonized food.


Figure 8: Cartoonized sceneries.


Figure 9: Cartoonized sceneries.

Real-world photo Cartoonized result

Figure 10: Cartoonized city scenes.


Figure 11: Cartoonized city scenes.


Figure 12: Cartoonized different objects.


Figure 13: Comparison between cartoon images and results of our methods in the same scene

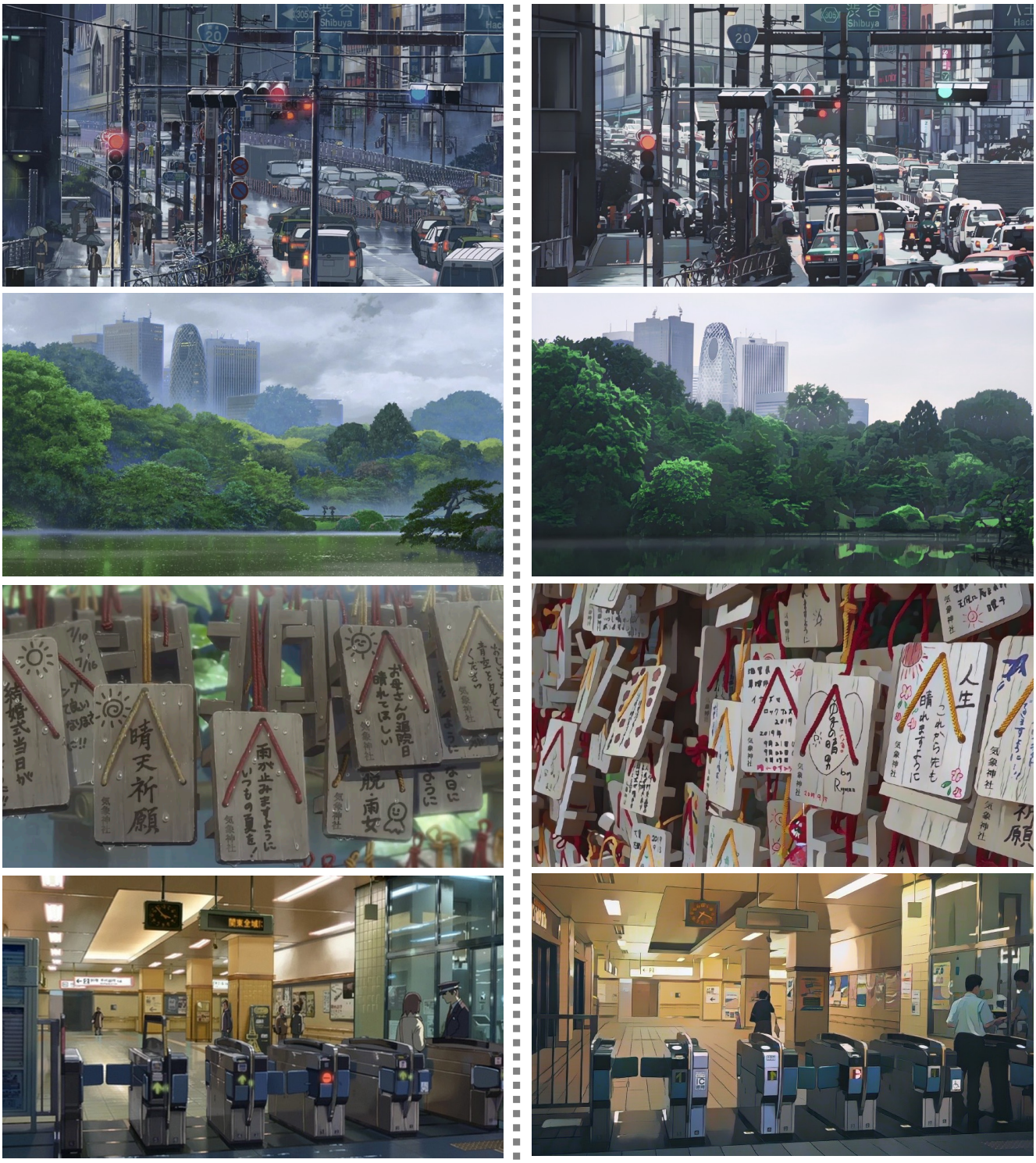


Figure 14: Comparison between cartoon images and results of our methods in the same scene.