Supplementary Material

Bowen Li[1], **Xiaojuan Qi**[2], **Philip H. S. Torr**[1], **Thomas Lukasiewicz**[1]
[1]University of Oxford, [2]University of Hong Kong
{bowen.li, thomas.lukasiewicz}@cs.ox.ac.uk
xjqi@eee.hku.hk, philip.torr@eng.ox.ac.uk

1 Architecture

As shown in Fig. 3, the generator consists of a text encoder, image encoders, and a series of upsampling and residual blocks, where the text encoder is a pre-trained bidirectional RNN [2, 9], and the image encoders are pre-trained Inception-v3 [6] and VGG-16 [5] networks.

1.1 Residual Block

As shown in Fig. 1, each residual block consists of two 3×3 convolution layers, two instance normalisations (INs) [7], and one GLU [1] non-linear activation function.

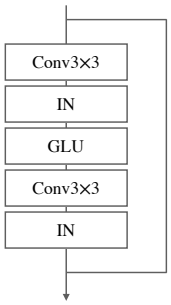


Figure 1: Architecture of the residual block.

1.2 Upsampling Block

As shown in Fig. 2, each upsampling block consists of one upsample function with nearest mode, one instance normalisation (IN), one 3×3 convolution layer, and one GLU non-linear activation function.

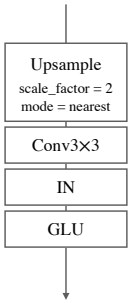
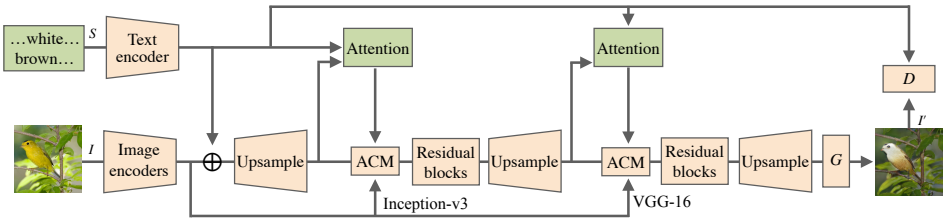


Figure 2: Architecture of the upsampling block.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.



architecture of our proposed lightweight network

Figure 3: Architecture of our model.

1.3 Trend of Manipulation Results

Following [3], we use paired data $(I, S) \rightarrow I$ to train our model, where S is the text description matching the image I. Therefore, there is a trade-off between the reconstruction of the original contents existing in the input images and the generation of new attributes aligned with the given text descriptions. To verify this trade-off, we investigate the change of the manipulation results when the training epoch increases. As shown in Figs. 4 and 5, we can easily observe that the visual attributes of the input images are modified smoothly, matching the given text descriptions, e.g., blue head, black eyerings, and red belly in Fig. 4, and green grass background in Fig. 5. However, when the epoch increases further, new modified attributes are gradually replaced by the original contents in the input image, and finally the synthetic images become almost the same as the input images.

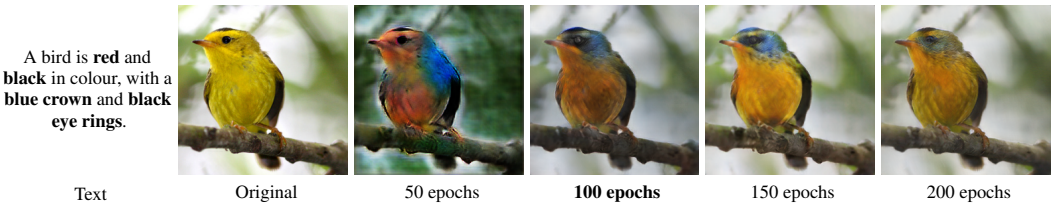


Figure 4: Trend of the manipulation results over epoch increases on the CUB dataset.

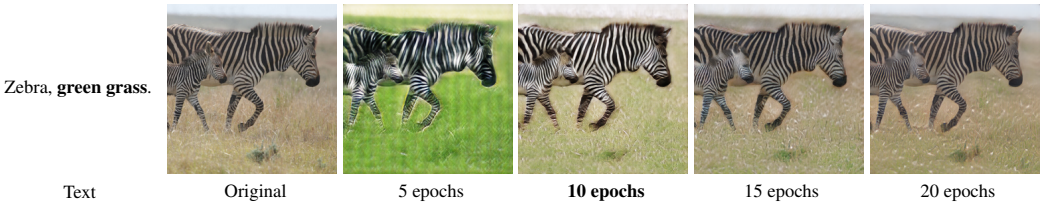


Figure 5: Trend of the manipulation results over epoch increases on the COCO dataset.

2 Additional Results

Fig. 6 shows various colour manipulations on the same images. In Figs. 7 and 8, we show additional comparison results between our method and ManiGAN [3] on the CUB [8] and COCO [4] datasets.

Figure 6: Various colour manipulations on the same images.

Figure 7: Additional comparison results between ManiGAN [3] and Ours on the CUB bird dataset.

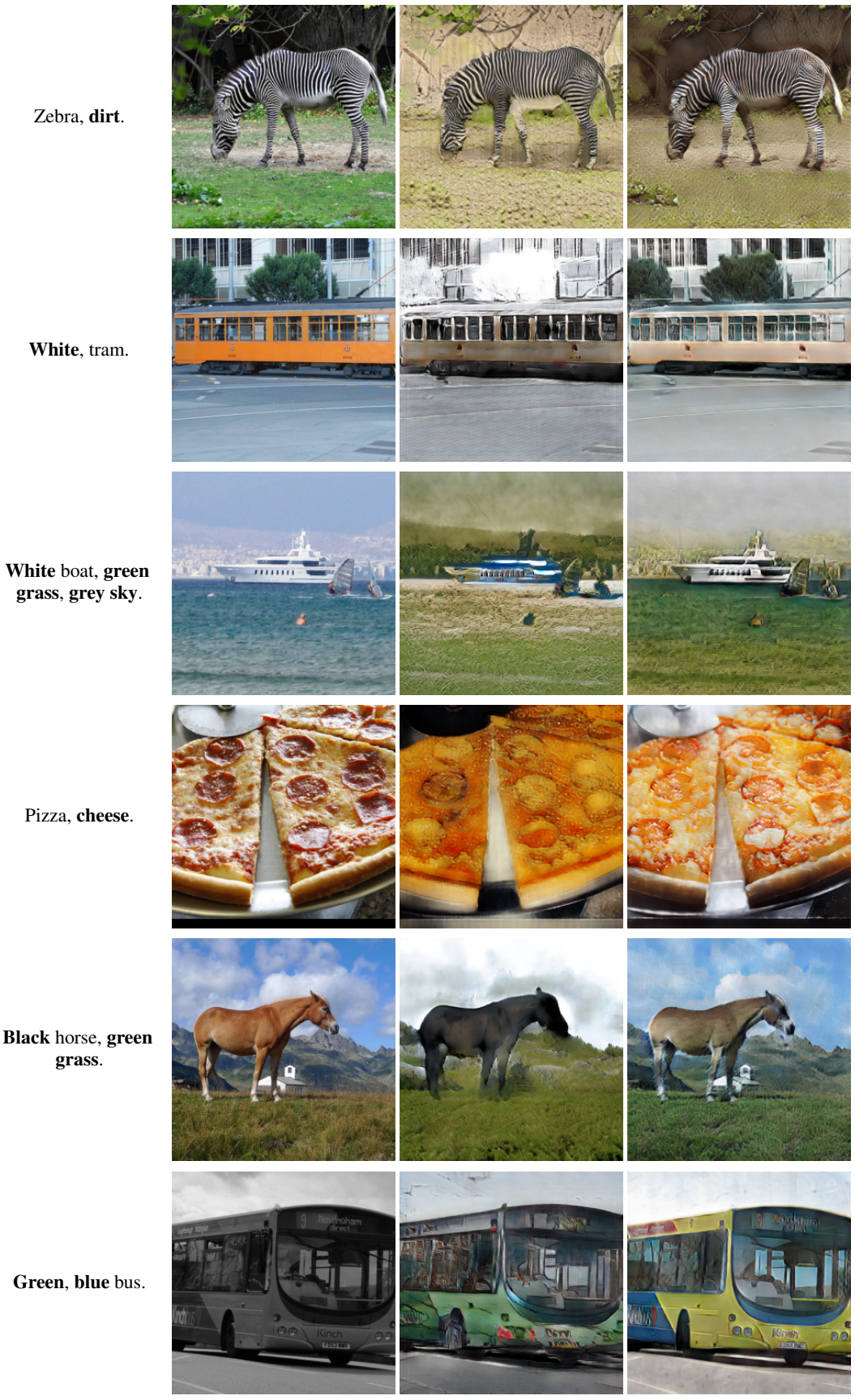


Figure 8: Additional comparison results between ManiGAN [3] and Ours on the COCO dataset.

References

[1] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 933–941, 2017.

[2] B. Li, X. Qi, T. Lukasiewicz, and P. Torr. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems*, pages 2063–2073, 2019.

[3] B. Li, X. Qi, T. Lukasiewicz, and P. Torr. ManiGAN: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[7] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[8] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 dataset. 2011.

[9] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.