ManiGAN: Text-Guided Image Manipulation

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

A. Architecture

We adopt the ControlGAN [3] as the basic framework and replace batch normalisation with instance normalisation [6] everywhere in the generator network except in the first stage. Basically, the affine combination module (ACM) can be inserted anywhere in the generator, but we experimentally find that it is best to incorporate the module before upsampling blocks and image generation networks; see Fig. 2.

A.1. Residual Block

Each residual block contains two convolutional layers, two instance normalisation (IN) [6], and one GLU [1] nonlinear function. The architecture of the residual block used in the detail correction module is shown in Fig. 1.

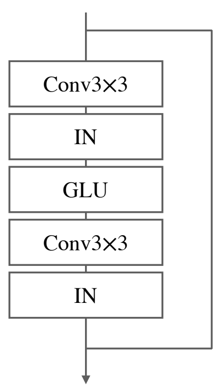

Figure 1. The architecture of the residual block.

B. Objective Functions

We train the main module and detail correction module separately, and the generator and discriminator in both modules are trained alternatively by minimising both the generator loss $\mathcal{L}_G$ and the discriminator loss $\mathcal{L}_D$.

Generator objective. The loss function for the generator follows those used in ControlGAN [3], but we introduce a regularisation term:

$$\mathcal{L}_{\text{reg}} = 1 - \frac{1}{CHW}||I' - I||, \tag{1}$$

to prevent the network achieving identity mapping, which can penalise large perturbations when the generated image becomes the same as the input image.

$$\mathcal{L}_G = \underbrace{-\frac{1}{2}E_{I' \sim PG}\left[\log(D(I'))\right]}_{\text{unconditional adversarial loss}} \underbrace{-\frac{1}{2}E_{I' \sim PG}\left[\log(D(I', S))\right]}_{\text{conditional adversarial loss}} + \mathcal{L}_{\text{ControlGAN}} + \lambda_1 \mathcal{L}_{\text{reg}}, \tag{2}$$

$$\mathcal{L}_{\text{ControlGAN}} = \lambda_2 \mathcal{L}_{\text{DAMSM}} + \lambda_3(1 - \mathcal{L}_{\text{corre}}(I', S)) + \lambda_4 \mathcal{L}_{\text{rec}}(I', I), \tag{3}$$

where I is the real image sampled from the true image distribution P_{data}, S is the corresponding matched text that correctly describes the I, I' is the generated image sampled from the model distribution PG. The unconditional adversarial loss makes the synthetic image I' indistinguishable from the real image I, the conditional adversarial loss aligns the generated image I' with the given text description S, $\mathcal{L}_{\text{DAMSM}}$ [8] measures the text-image similarity at the word-level to provide fine-grained feedback for image generation, $\mathcal{L}_{\text{corre}}$ [3] determines whether word-related visual attributes exist in the image, and $\mathcal{L}_{\text{rec}}$ [3] reduces randomness involved in the generation process. λ_1, λ_2, λ_3, and λ_4 are hyperparameters controlling the importance of additional losses. Note that we do not use $\mathcal{L}_{\text{rec}}$ when we train the detail correction module.

Discriminator objective. The loss function for the discriminator follows those used in ControlGAN [3], and the function used to train the discriminator in the detail correction module is the same as the one used in the last stage of the main module.

$$\begin{aligned}\mathcal{L}_D = & \underbrace{-\frac{1}{2}E_{I \sim P_{\text{data}}}\left[\log(D(I))\right] - \frac{1}{2}E_{I' \sim PG}\left[\log(1 - D(I'))\right]}_{\text{unconditional adversarial loss}} \\ & \underbrace{-\frac{1}{2}E_{I \sim P_{\text{data}}}\left[\log(D(I, S))\right] - \frac{1}{2}E_{I' \sim PG}\left[\log(1 - D(I', S))\right]}_{\text{conditional adversarial loss}} \\ & + \lambda_3((1 - \mathcal{L}_{\text{corre}}(I, S)) + \mathcal{L}_{\text{corre}}(I, S'_{})),\end{aligned} \tag{4}$$

where S' is a given text description randomly sampled from the dataset. The unconditional adversarial loss determines whether the given image is real, and the conditional adversarial loss reflects the semantic similarity between images and texts.

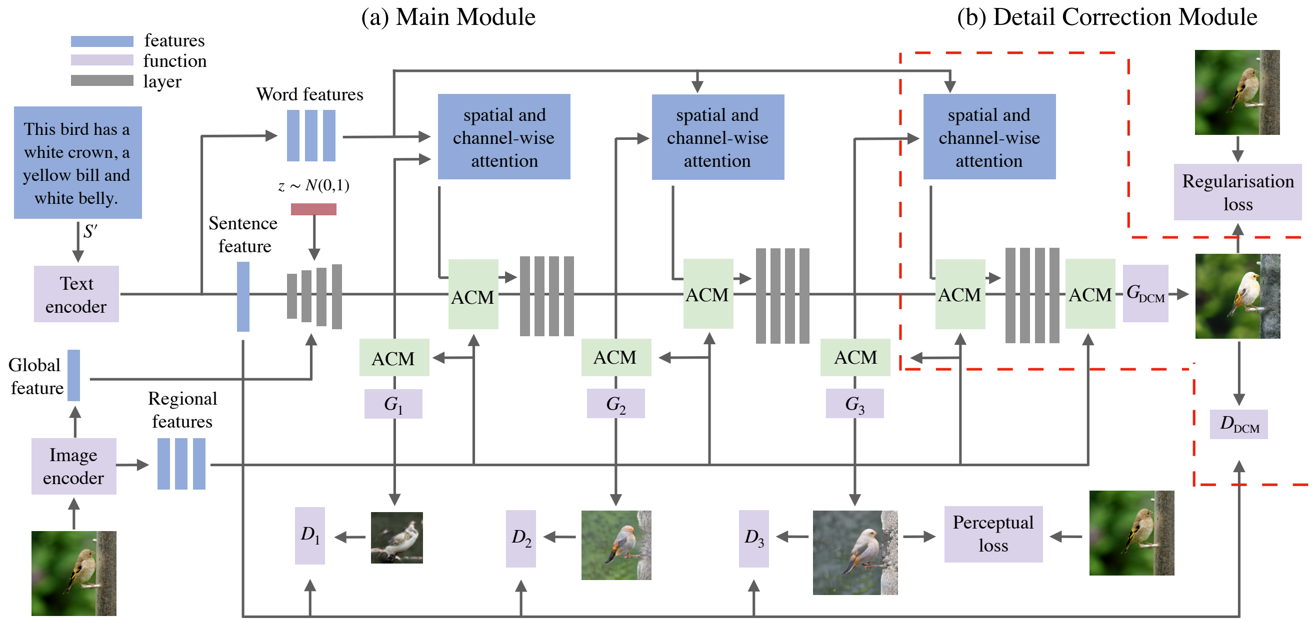

Figure 2. The architecture of ManiGAN. ACM denotes the text-image affine combination module. Red dashed box indicates the architecture of the detail correction module.

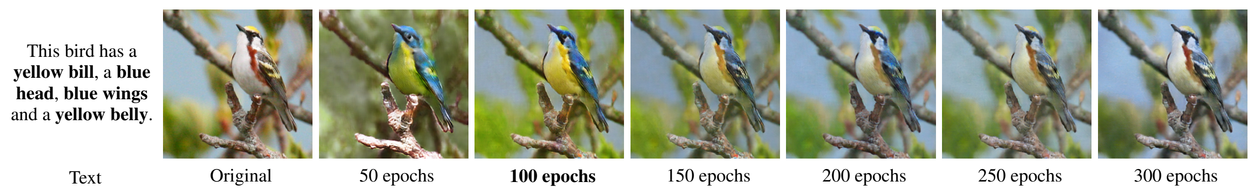

Figure 3. Trend of the manipulation results over epoch increases on the CUB dataset.

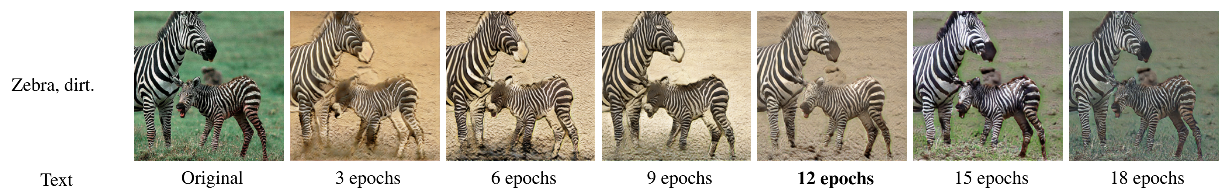

Figure 4. Trend of the manipulation results over epoch increases on the COCO dataset.

C. Trend of Manipulation Results

We track the trend of manipulation results over epoch increases, as shown in Figs. 3 and 4. The original images are smoothly modified to achieve the best balance between the generation of new visual attributes (e.g., blue head, blue wings and yellow belly in Fig. 3, dirt background in Fig. 4) and the reconstruction of text-irrelevant contents of the original images (e.g., the shape of the bird and the background in Fig. 3, the appearance of zebras in Fig. 4). However, when the epoch goes larger, the generated visual attributes (e.g., blue head, blue wings, and yellow belly of the bird, dirt background of the zebras) aligned with the given text descriptions are gradually erased, and the synthetic images become more and more similar to the original images. This verifies the existence of the trade-off between the generation of new visual attributes required in the given text descriptions and the reconstruction of text-irrelevant contents existing in the original images.

D. Additional Comparison Results

In Figs. 5, 6, 7, and 8, we show additional comparison results between our ManiGAN, SISGAN [2], and TAGAN [5] on the CUB [7] and COCO [4] datasets. Please watch the accompanying video for detailed comparison.

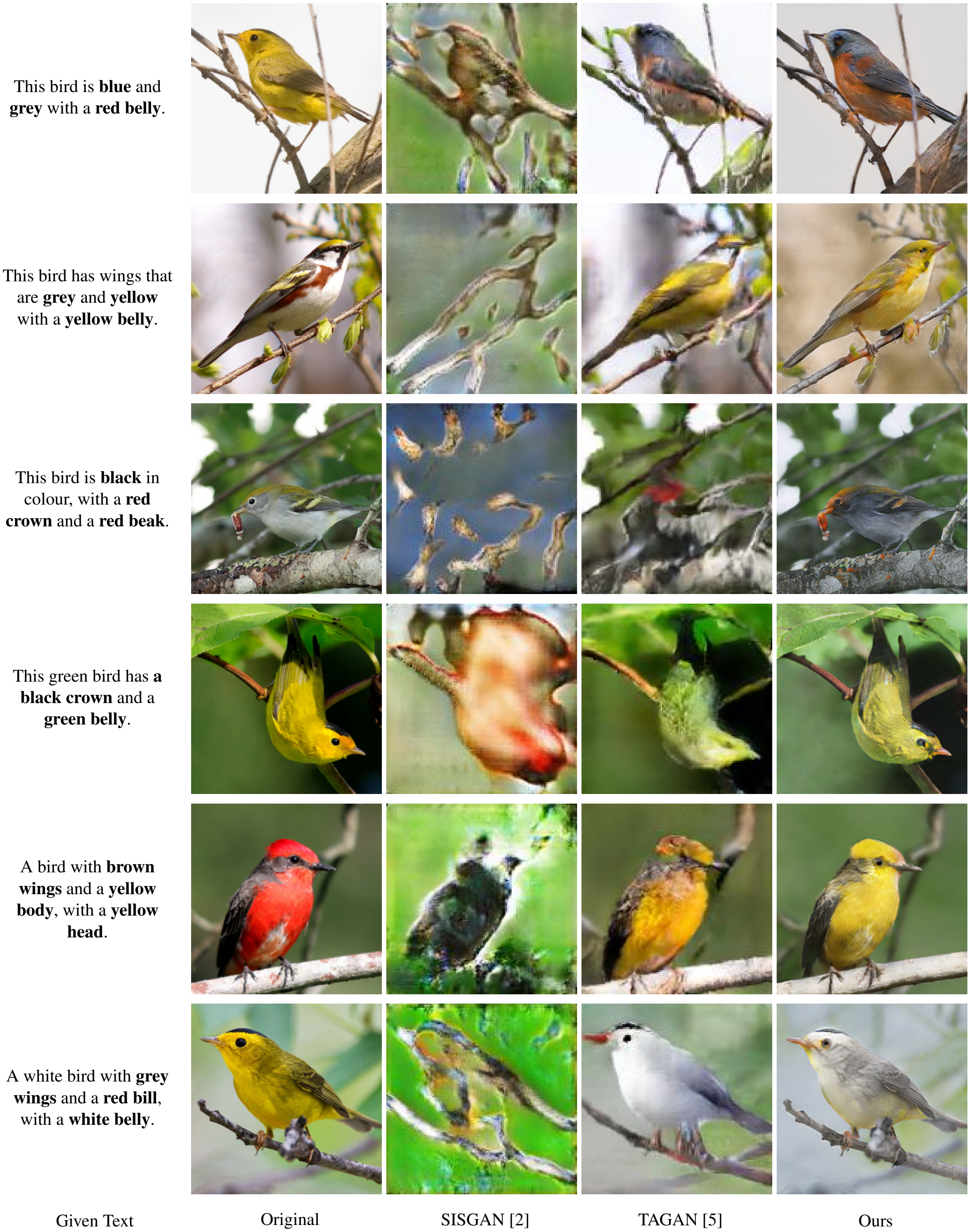

Figure 5. Additional comparison results between ManiGAN, SISGAN, and TAGAN on the CUB bird dataset.

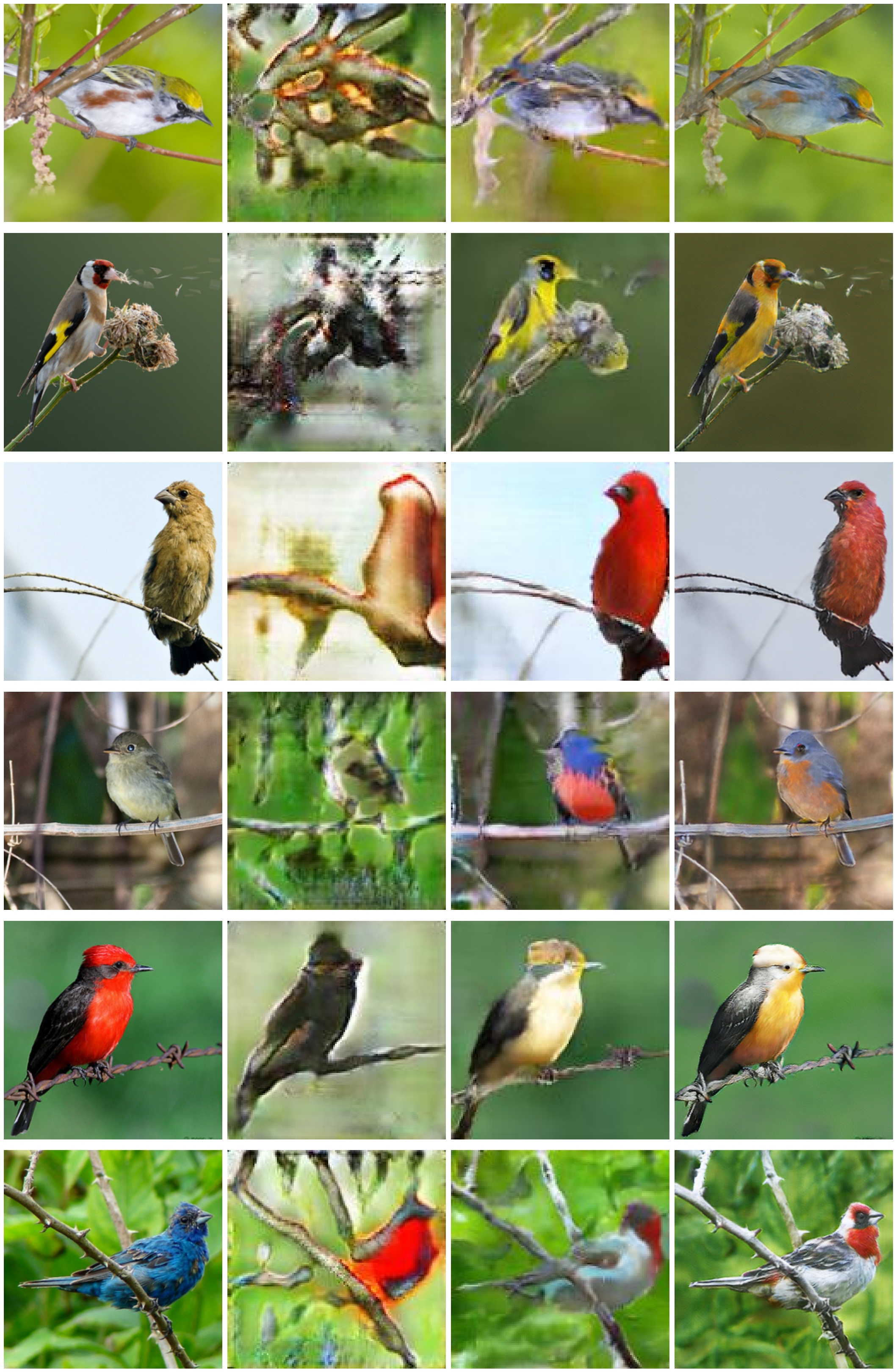

Figure 6. Additional comparison results between ManiGAN, SISGAN, and TAGAN on the CUB bird dataset.

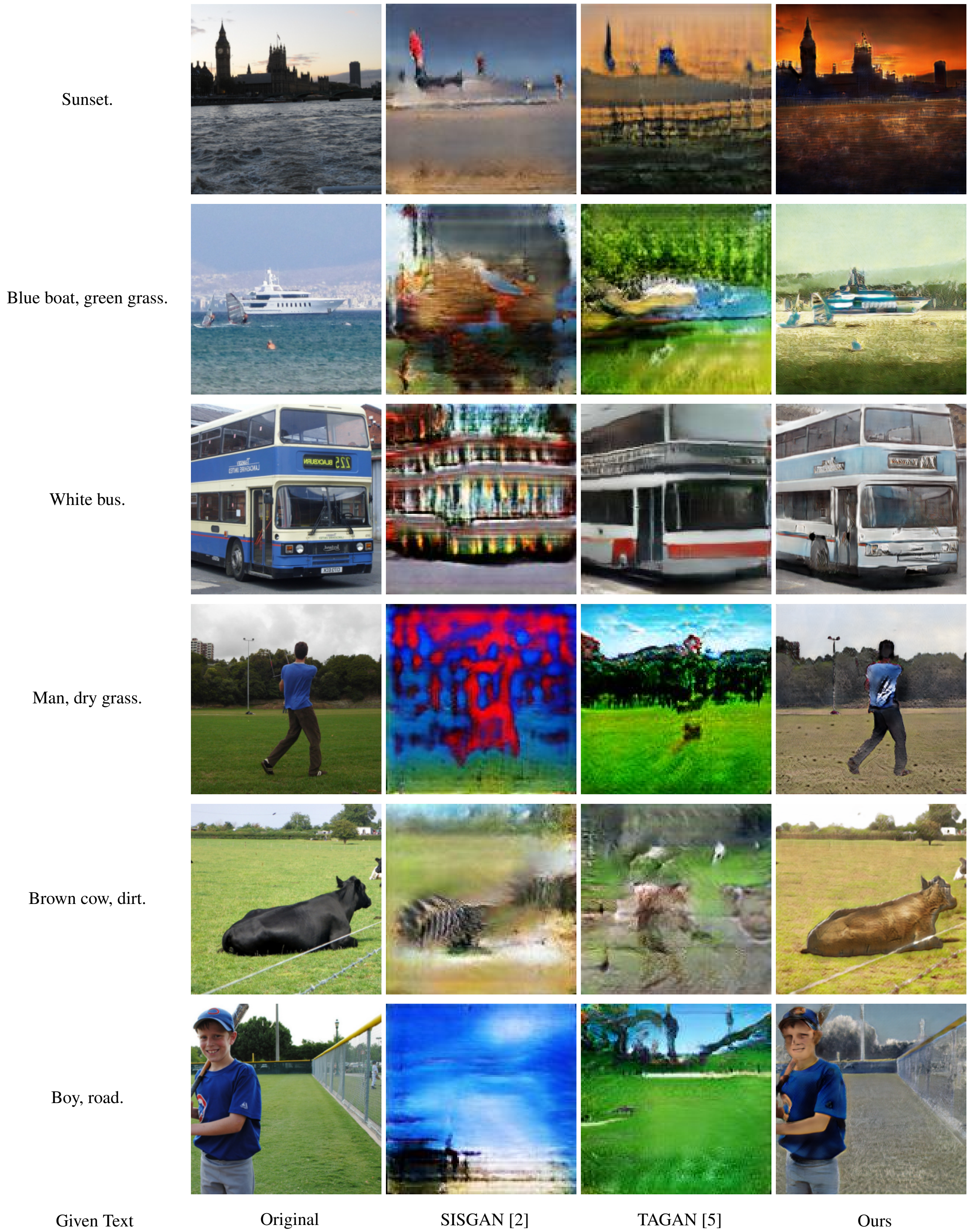

Figure 7. Additional comparison results between ManiGAN, SISGAN, and TAGAN on the COCO dataset.

Figure 8. Additional comparison results between ManiGAN, SISGAN, and TAGAN on the COCO dataset.

References

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

[2] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017.

[3] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Controllable text-to-image generation. *arXiv preprint arXiv:1909.07083*, 2019.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[5] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. In *Advances in Neural Information Processing Systems*, pages 42–51, 2018.

[6] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[7] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. 2011.

[8] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong 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.