MANIGAN: TEXT-GUIDED IMAGE MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a novel generative adversarial network for visual attributes manipulation (ManiGAN), which is able to semantically modify the visual attributes of given images using natural language descriptions. The key to our method is to design a novel co-attention module to combine text and image information rather than simply concatenating two features along the channel direction. Also, a detail correction module is proposed to rectify mismatched attributes of the synthetic image, and to reconstruct text-unrelated contents. Finally, we propose a new metric for evaluating manipulation results, in terms of both the generation of text-related attributes and the reconstruction of text-unrelated contents. Extensive experiments on benchmark datasets demonstrate the advantages of our proposed method, regarding the effectiveness of image manipulation and the capability of generating high-quality results.

1 INTRODUCTION

Image manipulation refers to the task of changing various aspects of given images from low-level colour or texture (Zhang et al., 2016; Gatys et al., 2016) to high-level semantics (Zhu et al., 2016), and has numerous potential applications in video games, image editing, and computer-aided design. Recently, with the development of deep learning and generative models, automatic image manipulation becomes possible, including image inpainting (Iizuka et al., 2016; Pathak et al., 2016), image colourisation (Zhang et al., 2016), style transfer (Gatys et al., 2016; Johnson et al., 2016), and domain or attribute translation (Lample et al., 2017; Isola et al., 2017).

However, all the above works mainly focus on specific tasks, and only few studies (Dong et al., 2017; Nam et al., 2018) concentrate on more general and user-friendly image manipulation by using natural language descriptions. Also, as shown in Fig.1, current state-of-the-art methods can only generate low-quality images and fail to effectively manipulate given images on more complicated datasets, such as COCO (Lin et al., 2014). The less effective performance is mainly because (1) simply concatenating text and image cross-domain features along the channel direction, the model may fail to precisely correlate words and corresponding visual attributes, and thus cannot modified specific attributes required in the text, and (2) conditioned only on a global sentence vector, current state-of-the-art methods lack important fine-grained information at the word-level, which prevents an effective manipulation using natural language descriptions.

In this paper, we aim to manipulate given images using natural language descriptions. In particular, we focus on modifying visual attributes (e.g., category, texture, colour, and background) of input images by providing texts that describe desired attributes. To achieve this, we propose a novel generative adversarial network for visual attributes manipulation (ManiGAN), which allows to effectively manipulate given images using natural language descriptions and to produce high-quality results.

The contribution of our proposed method is fourfold: (1) instead of simply concatenating hidden features generated from a natural language description and image features encoded from the input image along the channel direction, we propose a novel co-attention module where both features can collaborate to reconstruct the input image and also keep the synthetic result semantically aligned with the given text description, (2) a detail correction module (DCM) is introduced to rectify mismatched attributes, and to reconstruct text-unrelated contents existing in the input image, (3) a new metric is proposed, which can appropriately reflect the generation of text-related visual attributes and the reconstruction of text-unrelated contents involved in the image manipulation, and (4) extensive experiments on the CUB (Wah et al., 2011) and COCO (Lin et al., 2014) datasets are performed

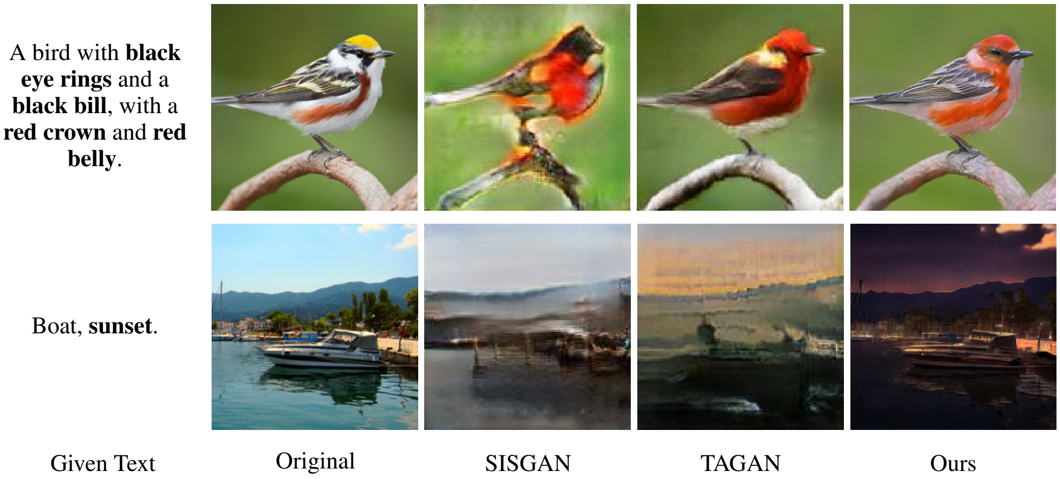

A bird with **black eye rings** and a **black bill**, with a **red crown** and **red belly**.
Boat, **sunset**.
Given Text
Original
SISGAN
TAGAN
Ours

Figure 1: Examples of image manipulation using natural language descriptions. Current state-of-the-art methods only generate low-quality images, and fail to do manipulation on COCO. In contrast, our method allows the input images to be manipulated accurately corresponding to the given text descriptions while preserving text-unrelated contents.

to demonstrate the superiority of our model, which outperforms existing state-of-the-art methods both qualitatively and quantitatively.

2 Related Work

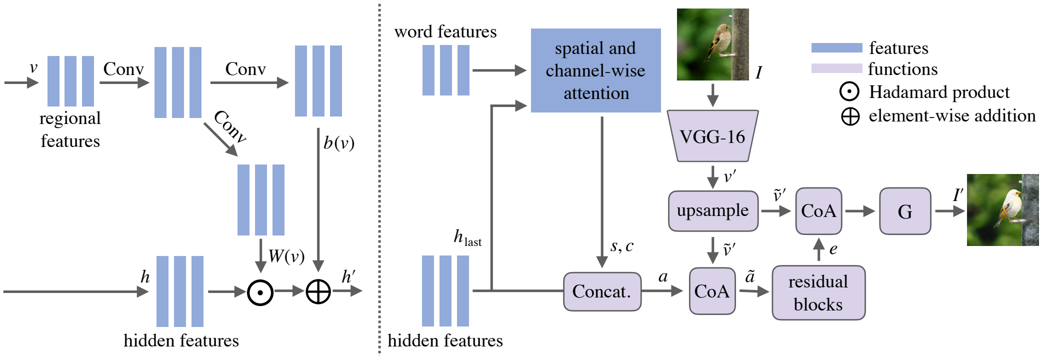
There are few studies focusing on image manipulation using natural language descriptions. Dong et al. (2017) proposed a GAN-based encoder-decoder architecture to disentangle the semantics of both input images and text descriptions. Nam et al. (2018) implemented a similar architecture, but introduced a text-adaptive discriminator that can provide specific word-level training feedback to the generator. However, both methods are limited in performance due to a less effective text-image concatenation method and a coarse sentence condition.

Our work is also related to conditional image manipulation. Brock et al. (2016) introduced a VAE-GAN hybridisation model to modify natural images by exploring the latent features. Isola et al. (2017) and Zhu et al. (2017) introduced paired and unpaired image-to-image translation methods based on conditional adversarial networks, respectively. However, all these methods focus mainly on image-to-image same-domain translation instead of image manipulation using cross-domain text descriptions.

Recently, text-to-image generation has drawn much attention due to the success of GANs in generating photo-realistic images. Reed et al. (2016) first proposed to use conditional GANs to generate plausible images from given text descriptions. Zhang et al. (2017) stacked multiple GANs to generate high-resolution images from coarse- to fine-scale. Xu et al. (2018) implemented a spatial attention mechanism to explore the fine-grained information at the word-level. However, all aforementioned methods mainly focus on generating new photo-realistic images from texts, and not on manipulating specific visual attributes of given images using natural language descriptions.

3 Generative Adversarial Networks for Image Manipulation

Let I denote an input image required to be modified, and S' denote a text description given by a user. We aim to semantically manipulate the input image I using the given text S', and also keep the visual attributes of the modified image I' semantically aligned with S' while preserving text-unrelated contents existing in I. To achieve this, we first adopt the ControlGAN (Li et al., 2019), as our basic framework, as it can effectively control text-to-image generation, and manipulate visual attributes of synthetic images. Then, we propose two novel components: (1) co-attention module,



(a) Co-Attention Module (CoA)

(b) Generator used in the Detail Correction Module

Figure 2: The architecture of the co-attention module and the generator used in the detail correction module. In (b), CoA denotes the co-attention module.

and (2) detail correction module to achieve effective image manipulation. We elaborate our model as follow, and the full architecture diagram is shown in Appendix A.

3.1 Co-Attention Module

As shown in Fig. 2 (a), our co-attention module takes two inputs: (1) the hidden features $h \in \mathbb{R}^{C\times H\times D}$, where C is the number of channels, H and D are the height and width of the feature map, respectively, and (2) the regional image features $v \in \mathbb{R}^{256\times 17\times 17}$ of the input image I encoded by the GoogleNet (Szegedy et al., 2015). The activation value $h' \in \mathbb{R}^{C\times H\times D}$ is given by $h' = h \odot W(v) + b(v)$, where $W(v)$ and $b(v)$ are the learned weights and biases dependent on the regional features v, and $\odot$ denotes Hadamard element-wise product. We use W and b to represent the functions that convert the regional features v to scaling and bias values. Then, the activation value h' serves as the input for the next stage. We also apply the co-attention module before implementing an image generation network to produce synthetic images; please see Appendix A for more details.

This linear combination form has been widely used in normalisation techniques (Park et al., 2019; Dumoulin et al., 2016; Huang & Belongie, 2017; De Vries et al., 2017), but, different from them, (1) our co-attention module is only applied at specific positions instead of all normalisation layers, which requires less computational resources, and (2) our co-attention module is designed to incorporate text and image cross-domain information, where W helps the model to focus on text-related visual attributes, while b provides input image information to help to reconstruct text-unrelated contents. Also, we experimentally find that implementing our co-attention module at all normalisation layers fails to produce reasonable images, which indicates that the normalisation techniques may not be suitable for the tasks requiring different domain information. Following Park et al. (2019), the functions W and b are implemented by a simple two-layer convolutional network, see Fig. 2 (a).

What has been learned by the co-attention module? To better understand what has been learned by our co-attention module, we conduct an ablation study shown in Fig. 3 to evaluate the effectiveness of W and b. As we can see, without W, some visual attributes cannot be perfectly generated (e.g., white belly in row 1 and the red head in row 2), and without b, the text-unrelated contents (e.g., background) are hard to preserve, which verify our assumption that W behaves as an attention function to help the model focus on text-related visual attributes, and b helps to complete missing text-unrelated details existing in the input image. Also, the visualisation of the channel feature maps of $W(v)$ shown in the last three columns of Fig. 3 validates the attention mechanism of W.

3.2 Detail Correction Module

The main purpose of our model is to incorporate input images and then generate modified images aligned with given text descriptions. Then, it may inevitably produce some new visual attributes or mismatched contents that are not required in the given texts. To fix this issue, we propose a

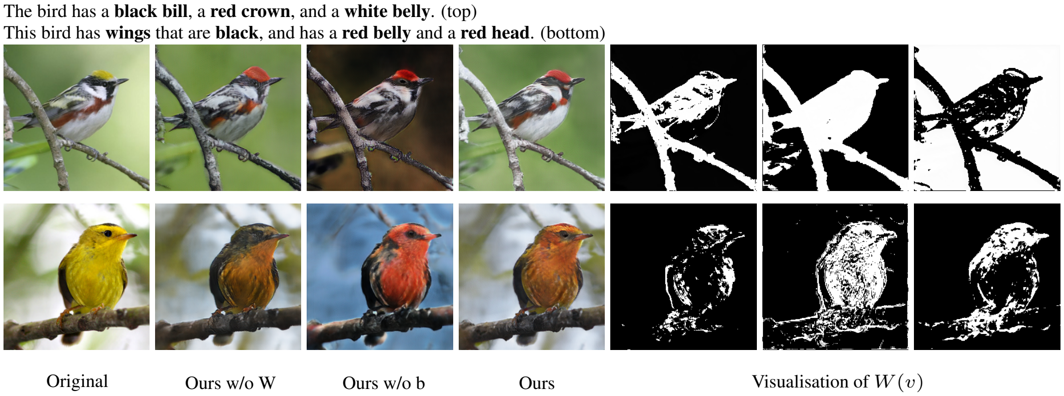

The bird has a **black bill**, a **red crown**, and a **white belly**. (top)
This bird has **wings** that are **black**, and has a **red belly** and a **red head**. (bottom)
Original
Ours w/o W
Ours w/o b
Ours
Visualisation of $W(v)$

Figure 3: Ablation studies of the learned W and b. The texts on the top are the given descriptions containing desired visual attributes, and the last three columns are the channel feature maps of $W(v)$.

detail correction module (DCM) to rectify inappropriate attributes, and to reconstruct text-unrelated contents existing in the input images.

The DCM consists of a generator and a discriminator, and is trained alternatively by minimising both objective functions. The generator, shown in Fig. 2 (b), takes three inputs: (1) the last hidden features $h_{\text{last}} \in \mathbb{R}^{C' \times H' \times D'}$ from the main module (we call our model without the DCM as main module), (2) the word features, and (3) visual features $v' \in \mathbb{R}^{128 \times 128 \times 128}$ that are extracted from the input image I by the VGG-16 (Simonyan & Zisserman, 2014) pretrained on ImageNet (Russakovsky et al., 2015). We have also applied GoogleNet (Szegedy et al., 2015) and ResNet (He et al., 2016) for feature extraction, but both do not perform well. Please refer to Appendix D for a detailed description of the detail correction module.

3.3 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 discriminator loss $\mathcal{L}_D$.

Generator objective. The loss function for the generator follows those used in ControlGAN (Li et al., 2019), but we introduce a regularisation term $\mathcal{L}_{\text{reg}} = 1 - \frac{1}{C_I H_I W_I} ||I' - I||$ to prevent the network achieving identity mapping, which can penalise large perturbations when the generated image becomes the same as the input image.

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

$$\mathcal{L}_{\text{ControlGAN}} = \lambda_2 \mathcal{L}_{\text{DAMSM}} + \lambda_3 (1 - \mathcal{L}_{\text{corre}}(I', S)) + \lambda_4 \mathcal{L}_{\text{rec}}(I', I), \tag{2}$$

where 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, $\mathcal{L}_{\text{DAMSM}}$ (Xu et al., 2018) measures the text-image similarity at the word-level to provide fine-grained feedback for image generation, $\mathcal{L}_{\text{corre}}$ (Li et al., 2019) determines whether words-related visual attributes exist in the image, and $\mathcal{L}_{\text{rec}}$ (Li et al., 2019) 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 (Li et al., 2019), 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.

$$\begin{aligned}\mathcal{L}_D = & \underbrace{-\frac{1}{2}E_{I\sim P_{\text{data}}}\left[\log(D(I))\right] - \frac{1}{2}E_{I'\sim PG}\left[\log(1-D(I'))\right]}_{\text{unconditional adversarial loss}} \\ & \underbrace{-\frac{1}{2}E_{I\sim P_{\text{data}}}\left[\log(D(I,S))\right] - \frac{1}{2}E_{I'\sim PG}\left[\log(1-D(I',S))\right]}_{\text{conditional adversarial loss}} \\ & + \lambda_3((1-\mathcal{L}_{\text{corre}}(I,S)) + \mathcal{L}_{\text{corre}}(I,S')),\end{aligned} \tag{3}$$

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.

Analysis. To prevent the model picking the input image as the solution, i.e., the model becomes an identity mapping network, we first introduce a regularisation term $\mathcal{L}_{\text{reg}}$ to penalise large perturbations when the generated image becomes the same as the input image, and then we stop the training early when the model reaches a stage achieving the best trade-off between the generation of new visual attributes aligned with given text descriptions and the reconstruction of text-unrelated contents existing in the input images. As for when to stop training, it is based on our proposed measurement metric, called manipulative precision (see Fig. 4), which is discussed in Sec. 4.

4 Experiments

To evaluate our model, extensive quantitative and qualitative experiments are carried out. Two state-of-the-art approaches on image manipulation using natural language descriptions, SISGAN (Dong et al., 2017) and TAGAN (Nam et al., 2018), are compared on the CUB birds (Wah et al., 2011) and more complicated COCO (Lin et al., 2014) datasets. Results for these two baselines are reproduced based on the code released by the authors. Please refer to Appendix A, B, and C for a detailed description of our network structures, the datasets, and training configurations.

Quantitative results. As mentioned above, our model can generate high-quality images compared with state-of-the-art methods. To demonstrate this, we adopt the inceptions score (IS) (Salimans et al., 2016) as the quantitative evaluation measure. In our experiments, we evaluate the IS on a large number of manipulated samples generated from mismatched pairs, i.e., randomly chosen input images manipulated by randomly selected text descriptions.

However, as the IS cannot reflect the quality of the content preservation, the L_1 pixel difference (diff) is calculated between the input images and corresponding modified images. Moreover, using the pixel difference alone may falsely report a good reconstruction due to over-training that the model becomes an identity mapping network. To address this issue, we propose a new measurement metric, called manipulative precision (MP), incorporating both the text-image similarity (sim) (Li et al., 2019) and the pixel difference, where the text-image similarity is calculated by performing the cosine similarity on the text features and corresponding image features encoded from the modified images. This is based on the intuition that if the manipulated images are generated from an identity mapping network, then the text-image similarity should be low, as the synthetic images cannot perfectly keep a semantic consistence with given text descriptions. Thus, the measurement metric is defined as $\text{MP} = (1 - \text{diff}) \times \text{sim}$.

As shown in Table 1, our method has the highest MP values on both the CUB and COCO datasets compared with the state-of-the-art approaches, which demonstrates that our method can better generate text-related visual attributes, and also reconstruct text-unrelated contents existing in the input images. The model without main module (i.e., only having the DCM) gets the highest IS, the lowest L_1 pixel difference, and low text-image similarity. This is because the model has become a identity mapping network and loses the capability of image manipulation.

Qualitative results. Figs. 5 and 6 show the visual comparison between our ManiGAN, SISGAN (Dong et al., 2017), and TAGAN (Nam et al., 2018) on the CUB and COCO datasets, respectively.

Table 1: Quantitative comparison: inception score (IS), text-image similarity (sim), L_1 pixel difference (diff), and manipulative precision (MP) of state-of-the-art approaches and ManiGAN on the CUB and COCO datasets. “w/o CoA” denotes without co-attention module. “w/ Concat.” denotes using concatenation method to combine hidden and image features. “w/o main” denotes without main module. “w/o DCM” denotes without detail correction module. For IS, similarity, and MP, higher is better; for pixel difference, lower is better.

	CUB				COCO			
Method	IS	sim	diff	MP	IS	sim	diff	MP
SISGAN	2.24	.045	.508	.022	3.44	.077	.442	.042
TAGAN	3.32	.048	.267	.035	3.28	.089	.545	.040
Ours w/o CoA	4.01	**.138**	.491	.070	5.26	.121	.537	.056
Ours w/ Concat.	3.81	.135	.512	.065	13.48	.085	.532	.039
Ours w/o main	**8.48**	.084	**.235**	.064	**17.59**	.080	**.169**	.066
Ours w/o DCM	3.84	.123	.447	.068	6.99	**.138**	.517	.066
Ours	8.47	.101	.281	**.072**	14.96	.087	.216	**.068**

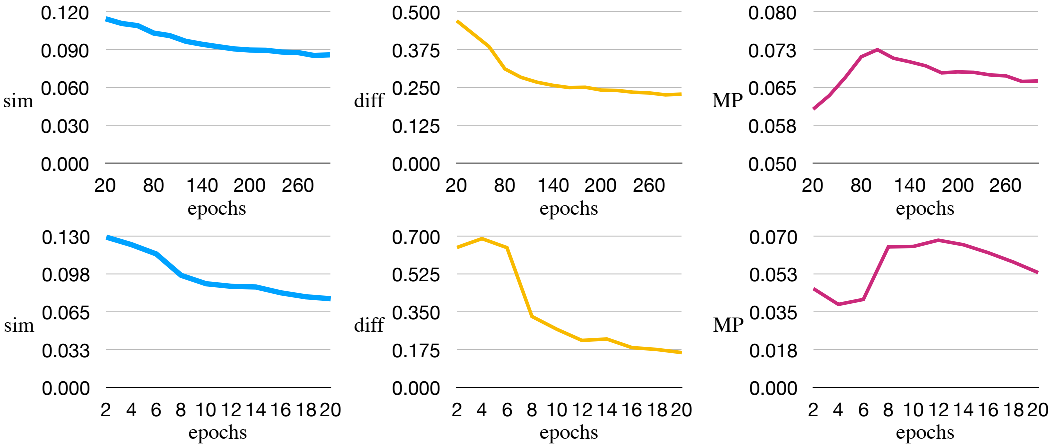


Figure 4: Text-image similarity, L_1 pixel difference, and manipulative precision values at different epochs on the CUB (top) and COCO (bottom) datasets. We suggest to stop training the DCM module when the model gets the highest MP values shown in the last column.

It can be seen that both state-of-the-art methods are only able to produce low-quality results and cannot effectively manipulate input images on the COCO dataset. However, our method is capable to perform an accurate manipulation and keep a highly semantic consistence between synthetic images and given text descriptions, while preserving text-unrelated contents. For example, shown in the last column of Fig. 6, SISGAN and TAGAN both fail to achieve an effective manipulation, while our model modifies the *green grass* to *dry grass* and also maps the *cow* into a *sheep*.

Note that as birds can have many detailed descriptions (e.g., colour for different parts), we use a long sentence to manipulate them, while the text descriptions for COCO are more abstract and focus mainly on categories, thus we use words to do manipulation for simplicity, which has the same effect as using long detailed text descriptions.

The effectiveness of the co-attention module. To verify the effectiveness of the co-attention module, we use the concatenation method to replace all co-attention modules, which concatenates hidden features h and regional features v along the channel direction, shown in Figs. 7 and 8 (d). As we can see that our full model can synthesise an object having exactly the same shape, pose, and position as the one existing in the input image, and also generate new visual attributes aligned with the given text description on the synthetic image. In contrast, as shown in the last two columns of Figs. 7 and 8 (d), with concatenation method, the model cannot reconstruct birds on the CUB bird dataset, and fails to do manipulation on the COCO dataset.

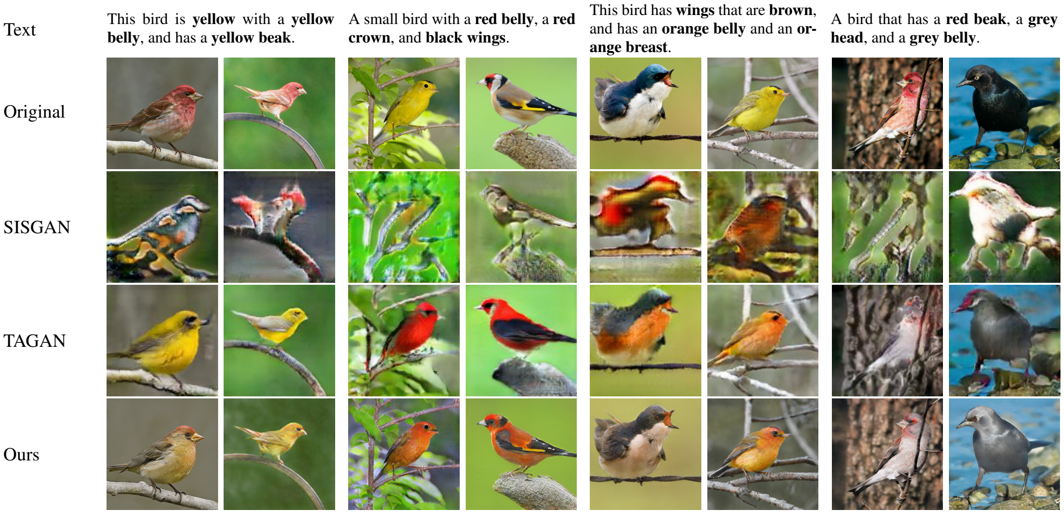

Text
This bird is yellow with a yellow belly, and has a yellow beak.
A small bird with a red belly, a red crown, and black wings.
This bird has wings that are brown, and has an orange belly and an orange breast.
A bird that has a red beak, a grey head, and a grey belly.
Original
SISGAN
TAGAN
Ours

Figure 5: Qualitative comparison of three methods on the CUB birds dataset.

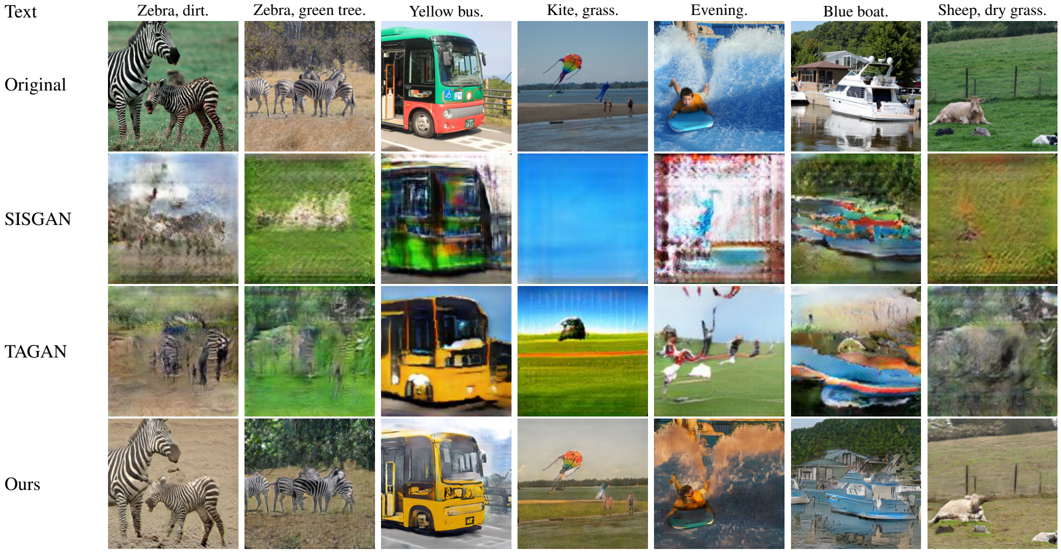


Figure 6: Qualitative comparison of three methods on the COCO dataset.

Also, to further validate the effectiveness of the co-attention module, we conduct an ablation study shown in Fig. 8 (c). It can be seen that our model without co-attention module that we just concatenate text and image features before feeding into the main module, which is used in Dong et al. (2017) and Nam et al. (2018), fails to produce reasonable images on both datasets. In contrast, our full model can better generate text-required attributes and also reconstruct text-unrelated contents shown in the last column. Table 1 also verifies the effectiveness of our co-attention module, as the values of IS and MP increase significantly when we implement the co-attention module.

The effectiveness of the detail correction module and main module. As shown in Fig. 8 (f), our model without detail correction module may miss some visual attributes (e.g., the bird missing the tail at row 2, the zebra missing the mouth at row 3), or generate new contents (e.g., new background at row 1, different appearance of bus at row 4), which indicates that the detail correction module can correct inappropriate attributes and reconstruct the text-unrelated contents. Fig. 8 (e) shows that without the main module, our model fails to do image manipulation on both datasets, which just achieves an identity mapping. This is mainly because the model cannot precisely correlate words with corresponding visual attributes, which mostly has been done in the main module.

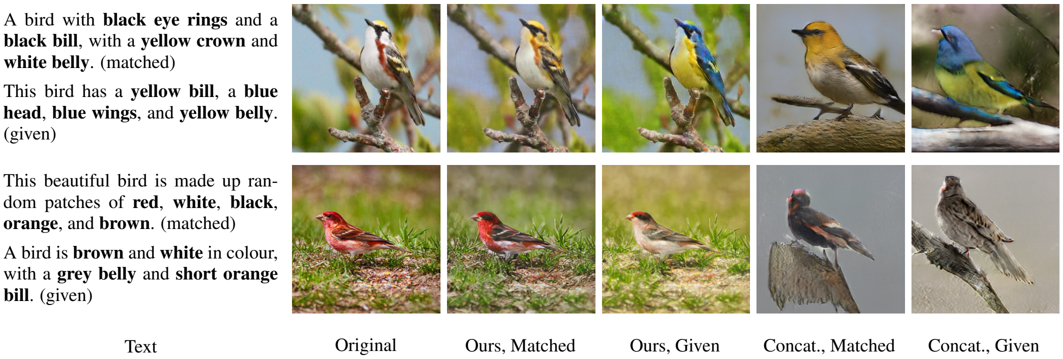

A bird with black eye rings and a black bill, with a yellow crown and white belly. (matched)
This bird has a yellow bill, a blue head, blue wings, and yellow belly. (given)
This beautiful bird is made up random patches of red, white, black, orange, and brown. (matched)
A bird is brown and white in colour, with a grey belly and short orange bill. (given)
Text
Original
Ours, Matched
Ours, Given
Concat., Matched
Concat., Given

Figure 7: Analysis of the co-attention module. “Matched” represents the texts matching original images. “Given” represents the texts provided by users. “Concat.” denotes that instead of using co-attention, hidden features are concatenated with image features along the channel direction.

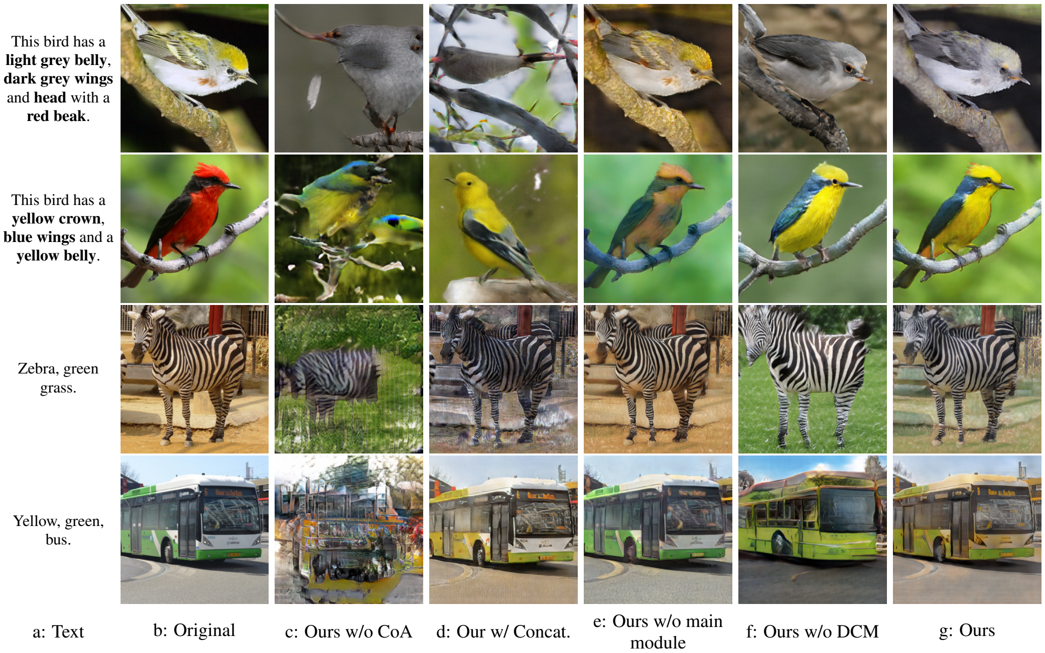

This bird has a light grey belly, dark grey wings and head with a red beak.
This bird has a yellow crown, blue wings and a yellow belly.
Zebra, green grass.
Yellow, green, bus.
a: Text
b: Original
c: Ours w/o CoA
d: Our w/ Concat.
e: Ours w/o main module
f: Ours w/o DCM
g: Ours

Figure 8: Ablation studies. a: given text describing the desired visual attributes; b: input image; c: removing the co-attention module and only concatenating image features and text features before feeding into the main module; d: using concatenation method to replace all co-attention modules; e: removing the main module and just training the DCM only; f: removing the DCM and just training the main module only; g: our full model.

5 Conclusion

We have proposed a novel generative adversarial network for visual attributes manipulation, called ManiGAN, which can semantically manipulate the input images using natural language descriptions. Two novel components are proposed in our model: (1) the co-attention module enables cooperation between hidden features and image features where both features can collaborate to reconstruct the input image and also keep the synthetic result semantically aligned with the given text description, and (2) the detail correction module can rectify mismatched visual attributes of the synthetic result, and also reconstruct text-unrelated contents existing in the input image. Extensive experimental results demonstrate the superiority of our proposed method, in terms of both the effectiveness of image manipulation and the capability of generating high-quality results.

REFERENCES

Andrew Brock, Theodore Lim, James M. Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016.

Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C. Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pp. 6594–6604, 2017.

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*, pp. 5706–5714, 2017.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, 2016.

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*, pp. 770–778, 2016.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (TOG)*, 35(4):110, 2016.

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*, pp. 1125–1134, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pp. 694–711. Springer, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pp. 5967–5976, 2017.

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

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*, pp. 740–755. Springer, 2014.

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*, pp. 42–51, 2018.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

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

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

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*, pp. 1316–1324, 2018.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5907–5915, 2017.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision*, pp. 649–666. Springer, 2016.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of the European Conference on Computer Vision*, pp. 597–613. Springer, 2016.

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*, pp. 2223–2232, 2017.

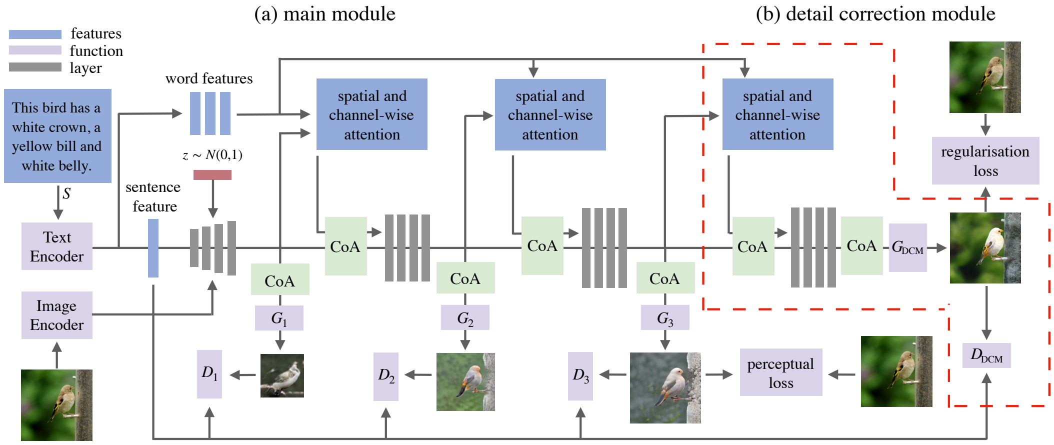

(a) main module
(b) detail correction module
features
function
layer
This bird has a white crown, a yellow bill and white belly.
S
Text Encoder
Image Encoder
word features
z ~ N(0,1)
sentence feature
spatial and channel-wise attention
spatial and channel-wise attention
spatial and channel-wise attention
CoA
CoA
CoA
CoA
CoA
CoA
CoA
CoA
G1
G2
G3
G_DCM
D1
D2
D3
D_DCM
perceptual loss
regularisation loss

Figure 9: The architecture of the ManiGAN. The red dashed box indicates detail correction module, the CoA denotes the co-attention module.

A Architecture Details

We adopt the ControlGAN (Li et al., 2019) as the basic framework and replace batch normalisation with instance normalisation (Ulyanov et al., 2016) everywhere in the generator network except in the first stage. Basically, the co-attention module 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. 9.

B Datasets

Our method is evaluated on the CUB birds (Wah et al., 2011) and the MS COCO (Lin et al., 2014) datasets. The CUB dataset contains 8,855 training images and 2,933 test images, and each image has 10 corresponding text descriptions. As for the COCO dataset, it contains 82,783 training images and 40,504 validation images, and each image has 5 corresponding text descriptions. We preprocess this two datasets based on the methods introduced in Zhang et al. (2017).

C Training Details

In our setting, we train the detail correction module (DCM) separately from the main module. Once the main module has converged, we train the DCM subsequently and set the main module as the eval mode. There are three stages in the main module, and each stage contains a generator and a discriminator. We train three stages at the same time, and three different-scale images 64×64, 128×128, 256×256 are generated progressively.

The main module is trained for 600 epochs on the CUB dataset and 120 epochs on the COCO dataset using the Adam optimiser (Kingma & Ba, 2014) with the learning rate 0.0002, and $\beta_1 = 0.5$, $\beta_2 = 0.999$. We do not use any learning rate decay, but for visualising generator output at any given point during the training, we use an exponential running average for the weights of the generator with decay 0.999.

As for the DCM, there is a trade-off between generation of text-related attributes and the reconstruction of text-unrelated contents. Based on the manipulative precision (MP) values (see Fig. 4), we find that training 100 epochs for the CUB, and 12 epochs for the COCO to achieve an appropriate balance between generation and reconstruction. The other training setting are the same as in the main module. The hyperparameters λ_1, λ_2, λ_3, and λ_4 are set to 1, 5, 0.5, and 1 for the CUB dataset, and 15, 5, 0.5, and 1 for COCO, respectively.

D Architecture of the Detail Correction Module

First, the visual features v' are converted into the same size as the hidden features h_{last} via a convolutional layer F, denoted $\tilde{v}' = Fv'$, where $\tilde{v}' \in \mathbb{R}^{128 \times H' \times D'}$. Then, we adopt the spatial attention and channel-wise attention introduced in (Li et al., 2019) to generate spatial attentive word-context features $s \in \mathbb{R}^{C' \times H' \times D'}$ and channel-wise attentive word-context features $c \in \mathbb{R}^{C' \times H' \times D'}$, and concatenate these two features with the hidden features h_{last} along the channel direction to generate new hidden features $a \in \mathbb{R}^{(3*C') \times H' \times D'}$. Next, to incorporate the visual features $\tilde{v}'$, we adopt the co-attention module here, donated $\tilde{a} = a \odot W'(\tilde{v}') + b'(\tilde{v}')$, where W' and b' are learned weights and bias dependent on visual features $\tilde{v}'$. Then, the transformed features $\tilde{a}$ are fed into a series of residual blocks followed by a convolutional layer to generate hidden features e. Before feeding e into a network to generate the output image, we apply the co-attention module on the e again to further strengthen the visual information; see Fig. 2 (b).

E Trend of Manipulation Results

We also track the trend of manipulation results over epoch increases, as shown in Fig. 10. The image is smoothly modified to achieve the best balance between the generation of new visual attributes (e.g., dirt background) and the reconstruction of text-unrelated contents (e.g., the appearance of zebras). However, when the epoch goes larger, the generated visual attributes (e.g., dirt background) aligned with the given text description are erased, and the synthetic image becomes more and more similar to the input image. This verifies the existence of the trade-off between the generation of new visual attributes required in the given text description and the reconstruction of contents existing in the input image.

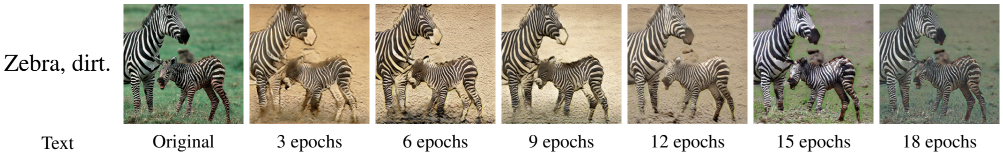


Figure 10: Trend of the manipulation results over epoch increases on the COCO dataset.

F Additional Results

We show additional comparison results between our ManiGAN, SISGAN (Dong et al., 2017), and TAGAN (Nam et al., 2018) on the CUB (Wah et al., 2011) and COCO (Lin et al., 2014) datasets.

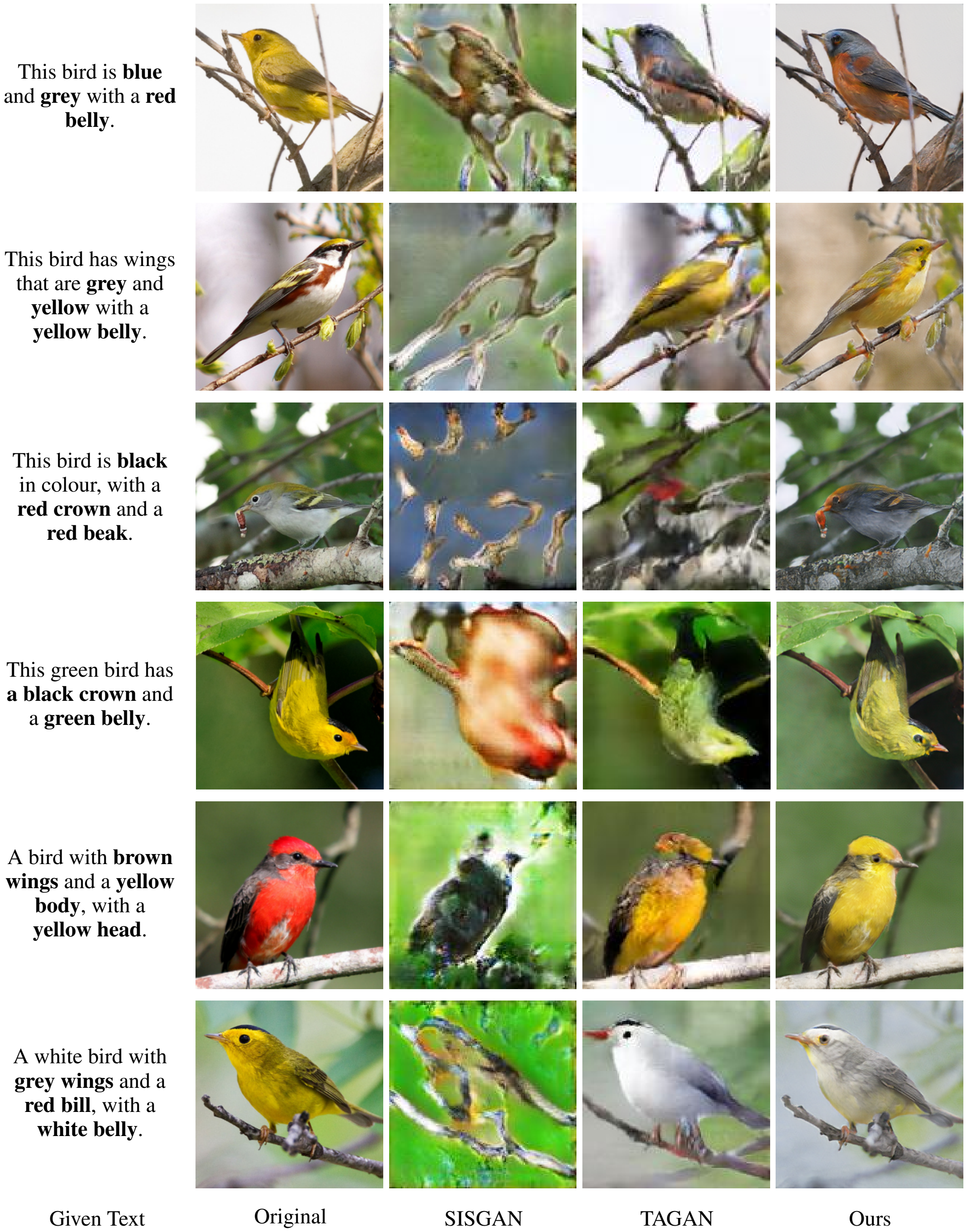

This bird is **blue** and **grey** with a **red belly**.
This bird has wings that are **grey** and **yellow** with a **yellow belly**.
This bird is **black** in colour, with a **red crown** and a **red beak**.
This green bird has **a black crown** and a **green belly**.
A bird with **brown wings** and a **yellow body**, with a **yellow head**.
A white bird with **grey wings** and a **red bill**, with a **white belly**.
Given Text
Original
SISGAN
TAGAN
Ours

Figure 11: Additional results between ManiGAN, SISGAN, and TAGAN on the CUB bird dataset.

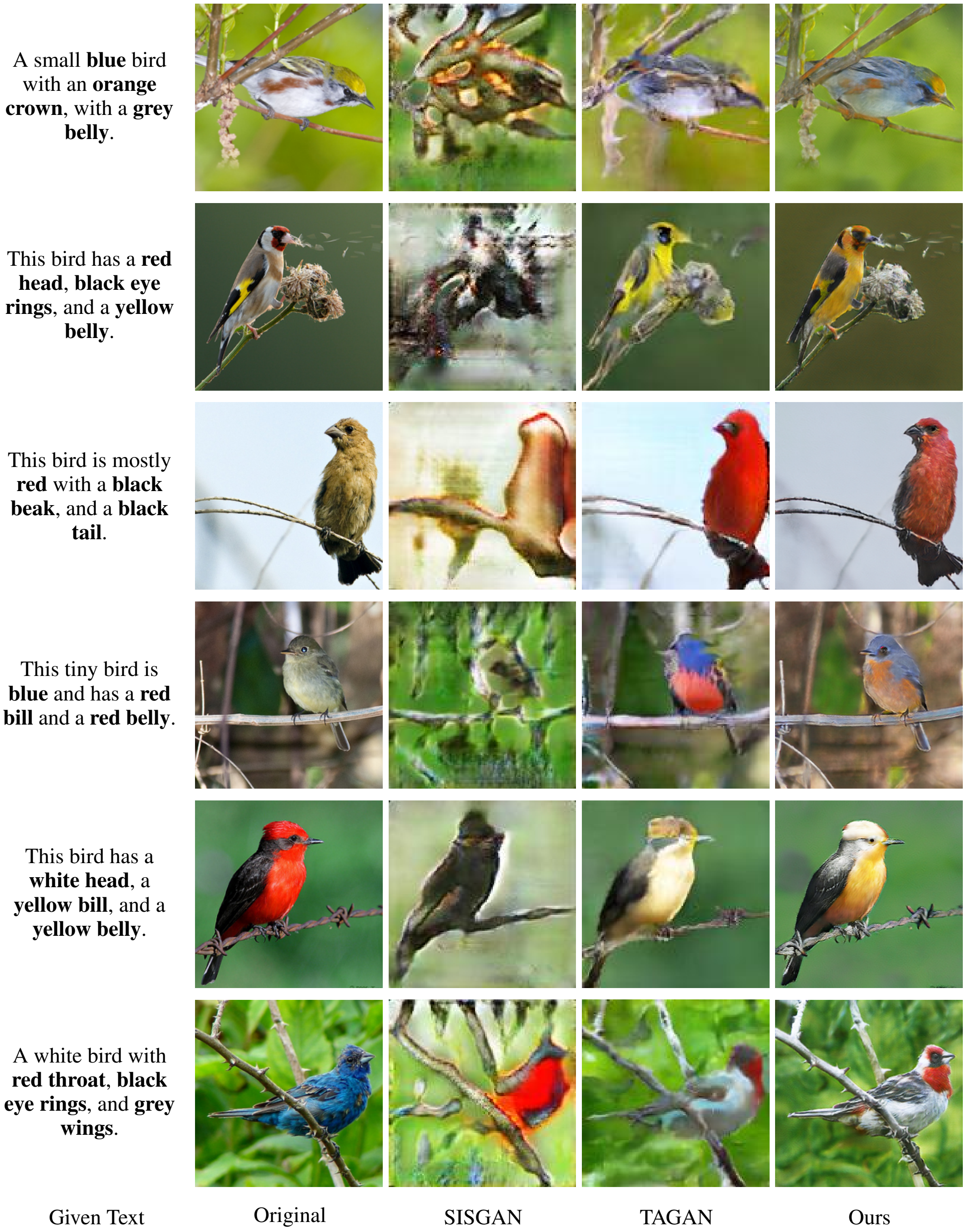

A small **blue** bird with an **orange crown**, with a **grey belly**.
This bird has a **red head**, **black eye rings**, and a **yellow belly**.
This bird is mostly **red** with a **black beak**, and a **black tail**.
This tiny bird is **blue** and has a **red bill** and a **red belly**.
This bird has a **white head**, a **yellow bill**, and a **yellow belly**.
A white bird with **red throat**, **black eye rings**, and **grey wings**.
Given Text
Original
SISGAN
TAGAN
Ours

Figure 12: Additional results between ManiGAN, SISGAN, and TAGAN on the CUB bird dataset.

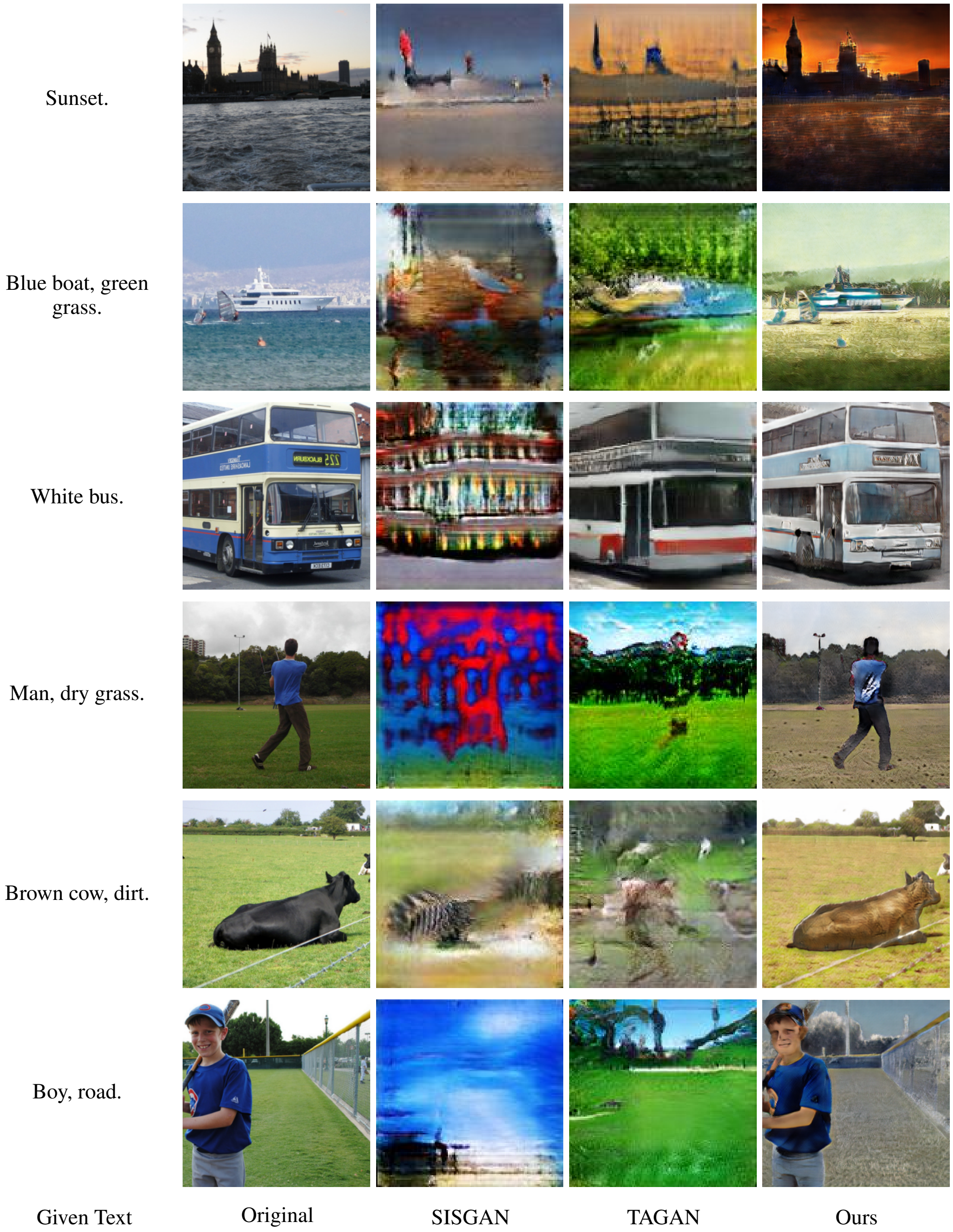


Figure 13: Additional results between ManiGAN, SISGAN, and TAGAN on the COCO dataset.

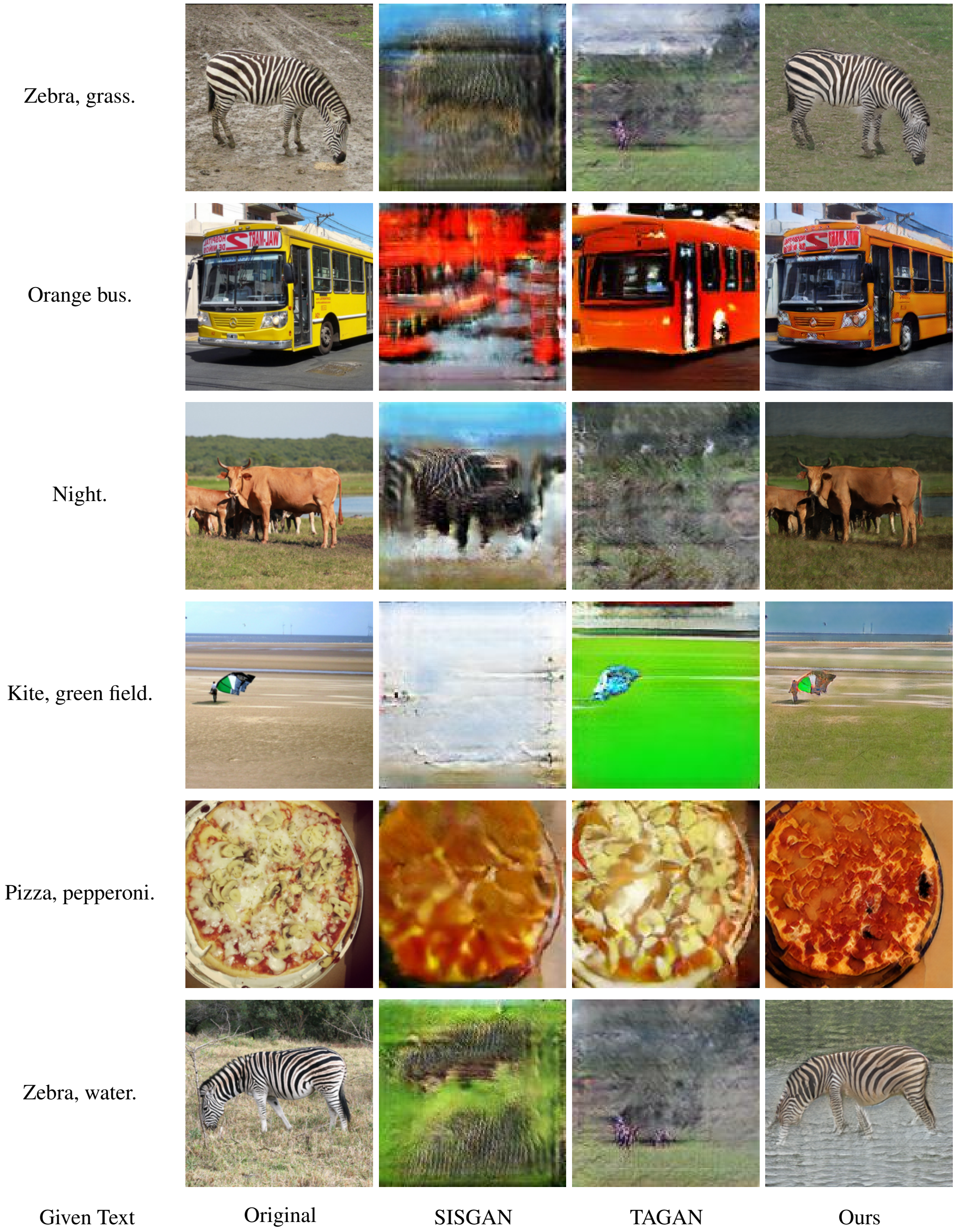


Figure 14: Additional results between ManiGAN, SISGAN, and TAGAN on the COCO dataset.