GENERATIVE ADVERSARIAL NETWORK INCLUDING REFERRING IMAGE SEGMENTATION FOR TEXT-GUIDED IMAGE MANIPULATION

Yuto Watanabe[†], *Ren Togo*[††], *Keisuke Maeda*[†††], *Takahiro Ogawa*[††††], *Miki Haseyama*[††††]

[†]School of Engineering, Hokkaido University, Japan
[††]Education and Research Center for Mathematical and Data Science, Hokkaido University, Japan
[†††]Office of Institutional Research, Hokkaido University, Japan
[††††]Faculty of Information Science and Technology, Hokkaido University, Japan
E-mail:{y_watanabe, togo, maeda, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

ABSTRACT

This paper proposes a novel generative adversarial network to improve the performance of image manipulation using natural language descriptions that contain desired attributes. Text-guided image manipulation aims to semantically manipulate an image aligned with the text description while preserving text-irrelevant regions. To achieve this, we newly introduce referring image segmentation into the generative adversarial network for image manipulation. The referring image segmentation aims to generate a segmentation mask that extracts the text-relevant region. By utilizing the feature map of the segmentation mask in the network, the proposed method explicitly distinguishes the text-relevant and irrelevant regions and has the following two contributions. First, our model can pay attention only to the text-relevant region and manipulate the region aligned with the text description. Second, our model can achieve an appropriate balance between the generation of accurate attributes in the text-relevant region and the reconstruction in the text-irrelevant regions. Experimental results show that the proposed method can significantly improve the performance of image manipulation.

Index Terms— Text-guided image manipulation, generative adversarial network, referring image segmentation.

1. INTRODUCTION

Taking images has become an important role in people's lives because of the rapid spread of mobile devices such as smartphones. With this trend, there are growing demands for applications that enable users to manipulate images for making them look better or reflecting the user's preference. However, existing applications require complex operations to manipulate images. To easily manipulate images, there are several techniques such as image inpainting [1–3], image colorization [4–6], style transfer [7–10] and domain or attribute transformation [11–13]. These methods focus on specific tasks such as converting a style of an image to a certain artist's style or colorizing a grayscale image, and it is still challenging to realize image manipulation reflecting the user's demands.

By using natural language descriptions, several methods focus on more user-friendly image manipulation called text-guided image manipulation [14–17]. These methods adopt generative adversarial networks [18] to manipulate the text-relevant region aligned with the text description given by a user and preserve the other regions. The methods pay attention to the text-relevant region by adopting attention mechanisms and semantically combine the region with the text description. However, these methods are still insufficient for generating accurate manipulated images. As shown in Fig. 1, red words in the text description specify the object and the manipulation, and we can see that it is difficult to manipulate a specific object when there are multiple same types of objects within the image. Moreover, the manipulation of attributes described in the text description is insufficiently conducted, and the attributes of the input image such as "white" still remain in manipulated images. Specifically, the previous methods have the following two problems. (1) Previous attention mechanisms have limitations in such a complex scene that contains multiple objects in the image, and (2) the generative adversarial network for image manipulation is extremely unstable. The first problem stems from the limitation of the previous attention mechanism in the network to explicitly learn the words such as "top" that distinguish the objects in the image. The second problem is that the task of text-guided image manipulation expects the network to completely different behavior that generates new attributes in the text-relevant region while preserving the other regions.

In this paper, we propose a novel text-guided image manipulation method and solve the above two problems. We newly introduce referring image segmentation [19] into the previous generative adversarial network for image manipulation [15]. The referring image segmentation aims to generate a segmentation mask to extract the text-relevant region in the input image. Our key idea is utilizing the segmentation mask to help the attention to the text-relevant region and stabilize the network. Specifically, we semantically associate the segmentation mask with the text description and allow the network to focus on generating new attributes. Moreover, by utilizing the segmentation mask, we distinguish the features calculated from the input image into the text-relevant and irrelevant regions. We use the distinguished features to mainly reconstruct text-irrelevant regions at the end of the model while keeping the balance of the network. We experimentally confirm the effectiveness of the proposed method on Caltech-UCSD Birds (CUB) [20]. Our two contributions can be summarized as follows.

- We semantically associate a feature map of the segmentation mask with the text description to pay attention only to the text-relevant region and manipulate the region aligned with the given text description.
- By utilizing the features of the input image with distinguishing the text-relevant and irrelevant regions, our model achieves a balance between the generation of accurate attributes and the reconstruction of the other remaining regions.

This work was partly supported by JSPS KAKENHI Grant Numbers JP20K19857 and JP20K19856.

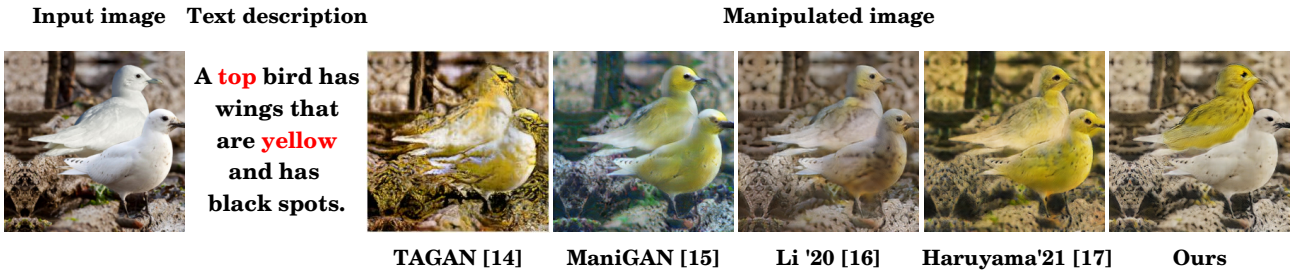

Fig. 1. Examples of the manipulated images from the input image and the text description. The results of the state-of-the-art methods still have limitations to manipulate only the text-relevant region aligned with the text description.

We show the results of text-guided image manipulation by the proposed method in Figs. 1 and 3.

2. OUR IMAGE MANIPULATION

Our goal is to generate a manipulated image I' that has new attributes based on a given text description T and preserves text-irrelevant regions in an input image I. We adopt ManiGAN [15] architecture as the base framework. ManiGAN is the generative adversarial network for image manipulation consisting of a text-image affine combination module (ACM) and a detail correction module (DCM). We newly introduce guided ACM (GACM) and guided DCM (GDCM) by extending ACM and DCM to segmentation-guided modules. As shown in Fig. 2, the proposed method has two units consisting of a main unit with the multi-stage and a GDCM unit. The generators G_{1-3} of each stage generate the images progressively in three different scales, 64×64, 128×128 and 256×256 pixels by upsampling the output features from the previous stage.

2.1. Training Strategy

To achieve accurate image manipulation, we newly utilize a segmentation mask M obtained by the referring image segmentation in our generative adversarial network. In our network, we transform the segmentation mask M to a feature map by copying and resizing M with a feature transformation module (FTM). The role of the feature map is to manipulate only the text-relevant region and help the reconstruction of the text-irrelevant regions at the end of the model. Specifically, by using GACM, the proposed method associates the feature map with the text description T to pay attention to the text-relevant region and conducts its manipulation. Thus, the module expects the generators G_{1-3} to generate the images having the new attributes only in the text-relevant region. Moreover, the feature map enables the feature extraction from the input image I with distinguishing the text-relevant and irrelevant regions. We use the distinguished features to mainly reconstruct the text-irrelevant regions in GDCM. Finally, the generator G_{GDCM} generates the manipulated image I'. The two modules and objective functions which contribute to the accurate image manipulation are described below.

2.2. Guided Text-Image Affine Combination Module (GACM)

To associate image and text features, the proposed method newly introduces GACM. As shown in Fig. 2, GACM takes two inputs, the feature map $\boldsymbol{M}_{\mathrm{GACM}} \in \mathbb{R}^{256\times17\times17}$ by FTM, and the hidden features $\boldsymbol{H} \in \mathbb{R}^{32\times H_i\times W_i}$ encoded from the text description T by a pre-trained LSTM [21] or output from the previous GACM. H_i and W_i are the input size of i th GACM. Note that H_1 and W_1 are 64 pixels, H_2 and W_2 are 128 pixels, and H_3 and W_3 are 256 pixels, respectively. The feature map $\boldsymbol{M}_{\mathrm{GACM}}$ is processed with the two convolutional layers to obtain a weight matrix $\boldsymbol{W}(\boldsymbol{M}_{\mathrm{GACM}})$ and a bias $\boldsymbol{B}(\boldsymbol{M}_{\mathrm{GACM}})$ that have the same size as $\boldsymbol{H}$. $\boldsymbol{W}(\boldsymbol{M}_{\mathrm{GACM}})$ can encode the text-relevant region, and $\boldsymbol{B}(\boldsymbol{M}_{\mathrm{GACM}})$ can encode the text-irrelevant regions in the input image I. We associate the cross-modality of the segmentation mask M and the text description T and calculate the hidden features $\boldsymbol{H}' \in \mathbb{R}^{32\times H_i\times W_i}$ as

$$\boldsymbol{H}' = \boldsymbol{H} \odot \boldsymbol{W}(\boldsymbol{M}_{\mathrm{GACM}}) + \boldsymbol{B}(\boldsymbol{M}_{\mathrm{GACM}}), \tag{1}$$

where $\odot$ is the Hadamard product. The proposed method adopts the multi-stage architecture consisting of GACMs, and the scale of the hidden features $\boldsymbol{H}'$ is progressively larger. In our network, the features in the text-relevant region are consistently passed on to the next GACM. Eventually, the proposed method realizes the model that can pay attention only to the text-relevant region and manipulate the corresponding region aligned with the text description T.

2.3. Guided Detail Correction Module (GDCM)

To modify mismatched attributes and reconstruct the text-irrelevant regions in the image output from G_3, we newly introduce GDCM. Our key idea is utilizing the feature map $\boldsymbol{M}_{\mathrm{GDCM}} \in \mathbb{R}^{32\times256\times256}$ obtained by inverting the output of FTM in GDCM and newly concatenating the input image I and the feature map $\boldsymbol{M}_{\mathrm{GDCM}}$. Specifically, we obtain the visual features $\boldsymbol{V} \in \mathbb{R}^{32\times256\times256}$ of the input image I by the pre-trained VGG16 network [22]. To distinguish the text-relevant and irrelevant regions in the visual features $\boldsymbol{V}$, we concatenate the visual features $\boldsymbol{V}$ with the feature map $\boldsymbol{M}_{\mathrm{GDCM}}$ and generate the features $\tilde{\boldsymbol{V}} \in \mathbb{R}^{32\times256\times256}$. By using the features $\tilde{\boldsymbol{V}}$ in GDCM, our model can achieve the reconstruction of the text-irrelevant regions while preserving the accurate attributes of the text-relevant region.

We utilize the features $\tilde{\boldsymbol{V}}$ and the word features encoded from the pre-trained LSTM [23] to refine the hidden features $\boldsymbol{H}_{\mathrm{last}} \in \mathbb{R}^{32\times256\times256}$ that are output from the last stage in the main unit. Specifically, we adopt an attention mechanism introduced in [23] and concatenate its results with $\boldsymbol{H}_{\mathrm{last}}$ to generate intermediate features $\boldsymbol{A} \in \mathbb{R}^{32\times256\times256}$. Then GACM associates $\boldsymbol{A}$ with $\tilde{\boldsymbol{V}}$ in the same manner as Eq. (1), where $\boldsymbol{A}$ and $\tilde{\boldsymbol{V}}$ correspond to $\boldsymbol{H}$ and $\boldsymbol{M}_{\mathrm{GACM}}$, and generates the features $\boldsymbol{A}' \in \mathbb{R}^{32\times256\times256}$. By obtaining the features $\boldsymbol{A}'$, our model can modify the missing information of the text-relevant region in the features $\boldsymbol{A}$ and reconstruct the text-irrelevant regions. Finally, we refine $\boldsymbol{A}'$ with a residual block to obtain the manipulated image I' by the generator G_{GDCM}.

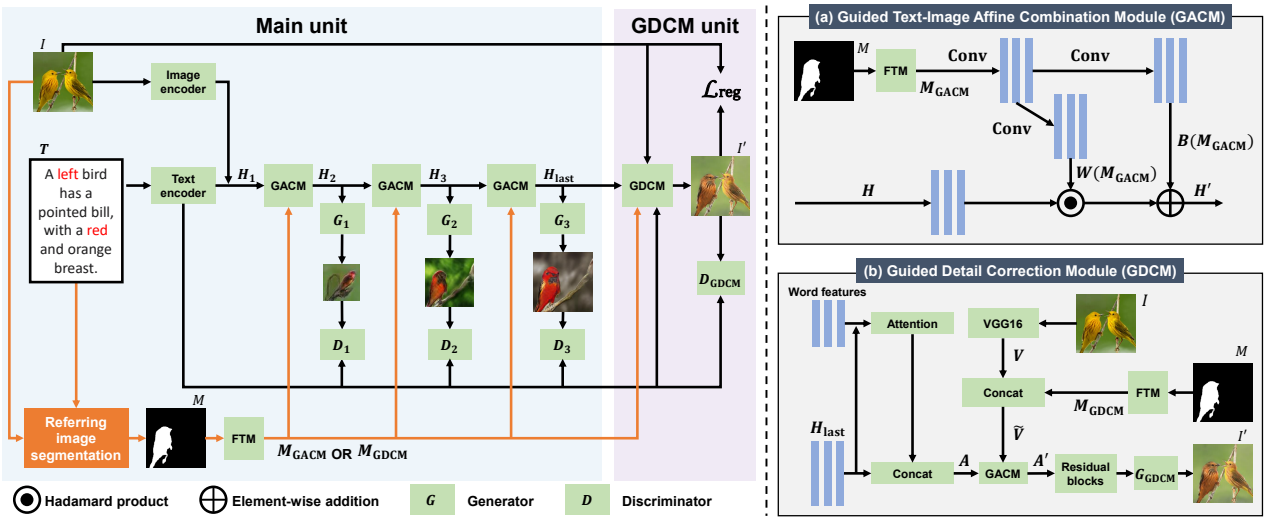

Fig. 2. The multi-stage architecture of our generative adversarial network for text-guided image manipulation. We newly utilize the feature map of the segmentation mask M obtained by the referring image segmentation in GACM and GDCM.

2.4. Objective Functions

To balance the accurate generation of the new attributes with the reconstruction, we train the generators $G_{1-3,\mathrm{GDCM}}$ and the discriminators $D_{1-3,\mathrm{GDCM}}$ of the generative adversarial network in the main unit and the GDCM unit by alternatively minimizing the generator loss $\mathcal{L}_G$ and the discriminator loss $\mathcal{L}_D$.

Generator objective. Based on [15, 24], the generator loss $\mathcal{L}_G$ consists of an adversarial loss $\mathcal{L}_{\mathrm{Gad}}$, a perceptual loss $\mathcal{L}_{\mathrm{per}}$, a text-image correlation loss $\mathcal{L}_{\mathrm{cor}}$, a text-image matching loss $\mathcal{L}_{\mathrm{DAMSM}}$ [23] and a regularization term $\mathcal{L}_{\mathrm{reg}}$. To ensure diversity that prevents the network from learning an identity mapping, we adopt the regularization term in the training of the GDCM unit. The regularization term and the generator loss $\mathcal{L}_G$ are defined as follows:

$$\mathcal{L}_{\mathrm{reg}} = -\frac{1}{C_I H_I W_I}||I' - I||^2_{\mathrm{F}}, \tag{2}$$

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{Gad}} + \mathcal{L}_{\mathrm{per}}(I', I) + \{1 - \mathcal{L}_{\mathrm{cor}}(I', T)\} + \mathcal{L}_{\mathrm{DAMSM}}(I', T) + \mathcal{L}_{\mathrm{reg}}(I', I), \tag{3}$$

where C_I, H_I and W_I are the number of color channels, the height and width of the input image I, respectively. $||\cdot||^2_{\mathrm{F}}$ is the Frobenius norm. The details of $\mathcal{L}_{\mathrm{Gad}}$, $\mathcal{L}_{\mathrm{per}}$, $\mathcal{L}_{\mathrm{cor}}$ and $\mathcal{L}_{\mathrm{DAMSM}}$ can be seen in [23, 24].

Discriminator objective. Based on [24], the discriminator loss $\mathcal{L}_D$ consists of an adversarial loss $\mathcal{L}_{\mathrm{Dad}}$ and the text-image correlation loss $\mathcal{L}_{\mathrm{cor}}$. The discriminator loss $\mathcal{L}_D$ is defined as follows:

$$\mathcal{L}_D = \mathcal{L}_{\mathrm{Dad}} + \{1 - \mathcal{L}_{\mathrm{cor}}(I, T)\} + \mathcal{L}_{\mathrm{cor}}(I, T'), \tag{4}$$

where T' is a mismatched text description that is randomly chosen from the dataset. The detail of $\mathcal{L}_{\mathrm{Dad}}$ is shown in [24].

3. EXPERIMENTS

The image manipulation of our method was evaluated on Caltech-UCSD Birds (CUB) [20]. CUB consists of 11,788 images each with 10 text descriptions, and there are 8,855 training images and 2,933 test images, respectively. We used the state-of-the-art text-guided image manipulation methods, TAGAN [14], ManiGAN [15], Li'20 [16] and Haruyama'21 [17] as the comparative methods. To confirm the effectiveness of our method, we performed quantitative and qualitative comparisons.

Implementation. We used the referring image segmentation model [19] pre-trained on RefCOCO [25] to generate the segmentation mask to extract the text-relevant region. We followed [15] and trained the GDCM unit separately from the main unit as shown in Fig. 2. Specifically, we trained the GDCM unit after having converged the training of the main unit and set the main unit as the evaluation mode in the training of the GDCM unit. We trained the main unit for 600 epochs and the GDCM unit for 100 epochs on CUB using Adam [26].

Evaluation Metrics. To evaluate the quality of the manipulated images by our method, Inception Score (IS) [27] and Fréchet Inception Distance (FID) [28] were applied. However, the accuracy of the image manipulation should be ideally evaluated whether the manipulated image is aligned with the text description. Since the evaluations of the image quality by IS and FID are insufficient for our task, we conducted the subject experiments. In the subject experiments, we used two metrics, Accuracy and Realism following [16]: (1) Accuracy: whether the text-relevant region of the manipulated image is aligned with the given text description, and whether the text-irrelevant regions are preserved, and (2) Realism: whether the manipulated image looks realistic.

3.1. Quantitative Comparison

In the subject experiments, we created the new dataset to contain two birds based on CUB and extended the text descriptions to describe one bird in each image as shown in Fig. 3. From the dataset, we randomly selected 30 images with randomly chosen the corresponding text descriptions. Then we asked subjects to compare five results after looking at the input image and the given text description based on the two metrics, Accuracy and Realism, and give scores from one to five, respectively. Finally, we collected 540 results from 18 subjects.

Fig. 3. Qualitative comparison of the manipulated images by our method and the four state-of-the-art methods [14–17]. Note that red words in the text description indicate the words that specify the object and their corresponding manipulation.

Table 1. The results of IS, FID, Accuracy and Realism used in the subject experiments of our method and the four state-of-the-art methods. (↑ means that higher is better, and ↓ means that lower is better.)

Method	IS(↑)	FID(↓)	Accuracy(↑)	Realism(↑)
TAGAN [14]	3.64	57.20	2.43	2.82
ManiGAN [15]	4.58	11.30	2.64	3.01
Li'20 [16]	4.64	**9.10**	2.68	3.67
Haruyama'21 [17]	4.54	9.47	2.64	3.01
Ours	**5.86**	9.33	**4.16**	**3.83**

Table 1 shows that our method achieves the highest IS and the second lowest FID, and both Accuracy and Realism of our method have the highest scores compared to the state-of-the-art methods. This indicates that the naturalness and the aesthetics of the manipulated images by our method are almost equal to or better than those by the state-of-the-art methods from the results of IS and FID. Moreover, by the results of Accuracy and Realism in the subject experiments, it can be seen that manipulated images by our method align with the text description and preserve the text-irrelevant regions successfully. We applied Welch's t-test to our method and Li'20 [16] and confirmed that our method had statistically significant differences of 1% (p-values < 0.01) in Accuracy and Realism.

3.2. Qualitative Comparison

We show the qualitative comparison between the manipulated images of our method and the four state-of-the-art methods [14–17] on CUB and the new dataset in Fig. 3. As shown in the samples (A)-(D) in Fig. 3, the comparison between our method and TAGAN [14] indicates that our method can generate more realistic manipulated images with higher resolution because of the multi-stage architecture. Furthermore, our method can exactly generate new attributes and suppress the manipulation of the text-irrelevant regions more successfully than all the state-of-the-art methods [14–17]. For example, in samples (A) and (B) on CUB, the text-relevant regions of the manipulated images by our method contain the accurate attributes described in the text descriptions. On the other hand, the manipulated images by the comparative methods still contain the attributes of the input images. Next, in samples (C) and (D) with complex scenes on the new dataset, our method can successfully manipulate only the text-relevant region and reconstruct the text-irrelevant regions. On the other hand, the comparative methods manipulated all the birds described in the text description and have limitations in reconstructing the text-irrelevant regions. These experimental results proved the superiority of our text-guided image manipulation compared to the four state-of-the-art methods [14–17].

4. CONCLUSION

We have proposed a new generative adversarial network to conduct text-guided image manipulation. By introducing the referring image segmentation that generates the segmentation mask to extract the text-relevant region, the model can explicitly distinguish the text-relevant and irrelevant regions. As a result, our method can exactly manipulate only the text-relevant region aligned with the text description without the manipulation of the other regions. Experimental results show the advantages of our method compared to the state-of-the-art methods both quantitatively and qualitatively.

5. REFERENCES

[1] D. Pathak, P. Krahenbuhl, J. Donahue *et al.*, "Context encoders: Feature learning by inpainting," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[2] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[3] Y. Jo and J. Park, "SC-FEGAN: Face editing generative adversarial network with user's sketch and color," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1745–1753.

[4] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2015, pp. 415–423.

[5] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. European Conference on Computer Vision*, 2016, pp. 649–666.

[6] S. Iizuka, E. Simo-Serra, and H. 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*, vol. 35, no. 4, pp. 1–11, 2016.

[7] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[8] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[9] J.-Y. Zhu, T. Park, P. Isola *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2017, pp. 2223–2232.

[10] A. Madaan, A. Setlur, T. Parekh *et al.*, "Politeness transfer: A tag and generate approach," *arXiv preprint arXiv:2004.14257*, 2020.

[11] P. Isola, J.-Y. Zhu, T. Zhou *et al.*, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[12] G. Lample, N. Zeghidour, N. Usunier *et al.*, "Fader networks: Manipulating images by sliding attributes," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5967–5976, 2017.

[13] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang *et al.*, "CrDoCo: Pixel-level domain transfer with cross-domain consistency," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1791–1800.

[14] S. Nam, Y. Kim, and S. J. Kim, "Text-adaptive generative adversarial networks: manipulating images with natural language," *arXiv preprint arXiv:1810.11919*, 2018.

[15] B. Li, X. Qi, T. Lukasiewicz *et al.*, "ManiGAN: Text-guided image manipulation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7880–7889.

[16] B. Li, X. Qi, P. H. Torr *et al.*, "Lightweight generative adversarial networks for text-guided image manipulation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 020–22 031, 2020.

[17] T. Haruyama, R. Togo, K. Maeda *et al.*, "Segmentation-aware text-guided image manipulation," in *Proc. IEEE International Conference on Image Processing*, 2021, pp. 2433–2437.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.

[19] S. Huang, T. Hui, S. Liu *et al.*, "Referring image segmentation via cross-modal progressive comprehension," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 488–10 497.

[20] C. Wah, S. Branson, P. Welinder *et al.*, "The Caltech-UCSD Birds-200-2011 dataset," *California Institute of Technology*, 2011.

[21] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

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

[23] T. Xu, P. Zhang, Q. Huang *et al.*, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1316–1324.

[24] B. Li, X. Qi, T. Lukasiewicz *et al.*, "Controllable text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 32, pp. 2065–2075, 2019.

[25] L. Yu, P. Poirson, S. Yang *et al.*, "Modeling context in referring expressions," in *Proc. European Conference on Computer Vision*, 2016, pp. 69–85.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] T. Salimans, I. Goodfellow, W. Zaremba *et al.*, "Improved techniques for training GANs," *arXiv preprint arXiv:1606.03498*, 2016.

[28] M. Heusel, H. Ramsauer, T. Unterthiner *et al.*, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6626–6637, 2017.