ACMA-GAN: Adaptive Cross-Modal Attention for Text-to-Image Generation

Longlong Zhou, Xiao-Jun Wu(✉), and Tianyang Xu

The School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, People's Republic of China
{wu_xiaojun,tianyang.xu}@jiangnan.edu.cn

Abstract. Automatically generating realistic and natural high resolution images from text descriptions is a complicated problem in the cross-modal research field. Recently, multi-stage conditional generative adversarial networks based on word attention are the mainstream of Text-to-Image generation. A close examination of these methods reveals two fundamental issues. Firstly, the granularity difference between the words and local image features makes the words cannot accurately express the local image features. Second, the discriminators cannot extract enough image information, which will result in poor discrimination effect. In this paper, we address these issues by proposing an adaptive cross-modal attention generative adversarial network (ACMA-GAN). Specifically, we design (1) an adaptive word attention module, which can reform the granularity of words and mine the context information of words; (2) a feature alignment module, which uses the pre-trained CNN model to improve the feature extraction ability of discriminator. Extensive experiments on CUB-200 and MS-COCO datasets demonstrate that our method is superior to the existing methods.

Keywords: cGAN · adaptive word attention · feature alignment

1 Introduction

In recent years, adversarial discriminant techniques are widely used in various fields [1,2,7,8,17–19], especially Text-to-Image generation (T2I) [4,6,9,10,14,16,20–23]. T2I is based on conditional generation adversarial network (cGANs) [7,8], which is a variant of generative adversarial networks (GANs) [2]. At present, T2I methods are mostly based on multi-stage network structure. Each stage contains a generator and a discriminator to control the generation of images of a specific size.

It is undeniable that AttnGAN [16] has an inestimable place in the field of Text-to-Image generation, it is the basis for the vast majority of three-stage generative models. AttnGAN [16] introduces word attention at both 64 and 128

This work is supported in part by the National Natural Science Foundation of China (Grant No. 62020106012, 62106089).

H. Lu et al. (Eds.): ICIG 2023, LNCS 14358, pp. 111–123, 2023.
https://doi.org/10.1007/978-3-031-46314-3_9

resolutions, which is essential to improve the details of the generated images. A significant problem also arises: whether the word features are at an identical granularity to the local image features at different resolutions. It is well known that local image features are pixel-level, whereas the granularity of word features is significantly higher than pixel-level local image features, and usually a word corresponds to a set of local feature vectors. Therefore, it is important to effectively reduce the granularity difference between words and local Image features before using word attention. DM-GAN [23] uses the image features after global pooling on the intermediate image features as modulation information, then the word features and image features are weighted and summed through a gating mechanism to obtain a new word representation, which will be used in word attention. We believe that the design of DM-GAN [23] still has some shortcomings. Firstly, the global image feature vector and sentence vector should have similar semantics and granularity. DM-GAN [23] does not explore the possibility of using sentence vector as modulation information. Secondly, whether the weighted sum of word feature vector and global feature vector will affect the context information between words. To alleviate these shortcomings, we designed an Adaptive Word Attention Module (AWAM). In AWAM, we concatenate the sentence feature vector and global image feature vector to the word feature vectors, and then use a self-attention module to obtain the new words. The new words constructed by our method can be regarded as a weighted sum of words, sentence and image features, and we not only consider the granularity relationship between word and local image features, but also take into account the context information of words.

As we all know, the discriminant ability of discriminator is based on its ability to extract features, if the discriminator can not extract enough effective information, it will be difficult to make a good discrimination. However, the discriminators in AttnGAN [16] and DM-GAN [23] are composed of few convolution and activation operations, so the feature extraction ability is poor. In order to improve the feature extraction ability of discriminators, we designed a Feature Alignment Module (FAM). In FAM, We use the excellent image classification network Inception-v3 model [13] as our feature extraction template, and then align the image features extracted by discriminator and pre-trained Inception-v3 model [13] at local and global levels. Specifically, we design a feature alignment loss to gradually reduce the distance of discriminator and Inception-v3 model [13] during training (Fig. 1).

In summary, the key contributions of our paper are as follows:

- We propose a novel model ACMA-GAN with adaptive cross-modal attention to generate realistic images.
- We design an adaptive word attention module (AWAM) to reform the granularity of words. The new words constructed by AWAM preserve the context information between words.
- A feature alignment loss is proposed to improve the feature extraction ability of discriminator. To be specific, We use a pre-trained CNN model to force the discriminator to extract more similar features.

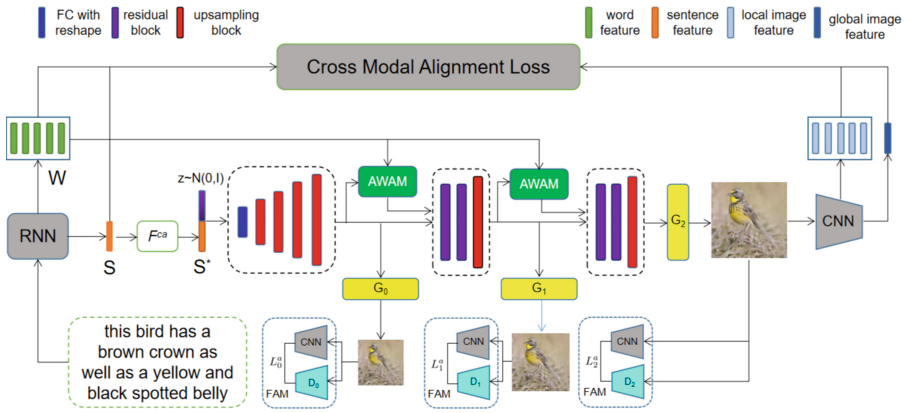

Fig. 1. The structure of our ACMA-GAN.

- Extensive experiments confirm that our ACMA-GAN outperforms most advanced methods.

2 Methodology

In this section, we first illustrate the overall structure of our ACMA-GAN, then analyze the Adaptive Word Attention Module (AWAM) and Feature Alignment Module (FAM), finally we analyze the loss function of generator and discriminator in detail.

2.1 Model Overview

Our ACMA-GAN contains a text encoder, an image encoder, a three-stage generator, and three discriminators. The text encoder and the image encoder are the pre-trained Bi-LSTM [12] and Inception-v3 [13] models, respectively.

The generator takes random noise, sentence vector and word vectors as input. Sentence vector is concatenated with noise after a conditioning augmentation module F^{ca} as the input of the first stage. The output of the previous stage and AWAM are the inputs for the second and third stages. We use the pre-trained RNN and CNN models to calculate a cross-modal alignment loss to optimize the training of generator.

The discriminator contains a feature alignment loss L^a in addition to the adversarial loss, and we use the Inception-v3 model [13] to optimize the feature extraction capability of our discriminator by narrowing the distance of local features and global features in two CNN models.

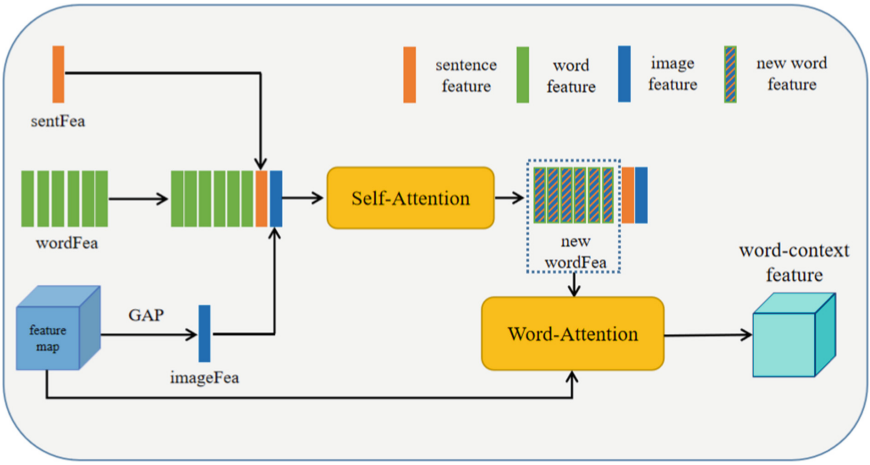

Fig. 2. The pipeline of Adaptive Word Attention Module (AWAM).

2.2 Adaptive Word Attention Module

The previous methods AttnGAN [16] and DM-GAN [23] fail to reduce the granularity difference between words and local image features, and DM-GAN [23] neglects the context of words when it rewrite the word embeddings. To address these problem, we propose an Adaptive Word Attention Module (AWAM). As shown in Fig. 2, we use global average pooling on the image features to obtain an image feature vector with the same dimension as the sentence and word vectors, then we concatenate the image vector, sentence vector and word vectors together, finally input them into a self-attention Module for rewriting word embeddings. The self-attention is as follows:

$$\begin{aligned} o_j &= v(\sum_{i=1}^{N} \beta_{j,i} h(x_i)), \\ h(x) &= W_h x, \\ v(x) &= W_v x. \end{aligned} \tag{1}$$

where x_i represents the content vector before self-attention and o_j represents the content vector after self-attention, $\beta_{j,i}$ represents the attention score between the jth and ith content vectors, and W_h and W_v are the weights of the fully connected layer. $\beta_{j,i}$ is calculated as follows:

$$\begin{aligned} \beta_{j,i} &= \frac{exp(s_{i,j})}{\sum_{i=1}^{N} exp(s_{i,j})}, \\ s_{i,j} &= f(x_i)^T g(x_j), \\ f(x) &= W_f x, \\ g(x) &= W_g x. \end{aligned} \tag{2}$$

where x_i and x_j represent the ith and jth content vectors, and W_f and W_g are the weights of the fully connected layer.

The new words after self-attention not only retain the context information between words, but also reduce the granularity difference between words and local image features. The new word features and image feature are input into the word attention module to calculate the word-content feature. Finally, the original image feature and the word-content feature are used as the input of the next stage. The word-content feature are computed as follows:

$$\begin{aligned} c_j &= \sum_{i=0}^{T-1} \beta_{j,i} e_i, \\ \beta_{j,i} &= \frac{exp(s_{j,i})}{\sum_{k=0}^{T-1} exp(s_{j,k})}, \\ s_{j,i} &= v_j \cdot e_i. \end{aligned} \tag{3}$$

where c_j represents the word-content feature vector at the jth pixel position, $\beta_{j,i}$ represents the attention score between the local feature at the jth pixel position and the ith word, v_j represents the jth sub-region of the image, and e_i represents the ith word.

In our adaptive word attention module, the new words constructed from the words, sentence and image information can better play the ability of word attention. The sentence feature vector and image feature vector constructed by the self-attention module are discarded in the word attention module, because their granularity is higher than the local features of the image, using them will not be conducive to the construction of word-content feature.

2.3 Feature Alignment Module

In order to improve the feature extraction ability of the discriminator and promote the better performance of the discriminative network, we design a Feature Alignment Module (FAM). In FAM, we design a feature alignment loss between the discriminator and the pretrained Inception-v3 model [13] at both local and global levels. Before the feature alignment, we use convolution operation to adjust the dimension of local and global features in discriminator, aiming to project the features of Inception-v3 model [13] and discriminator into a common space. The feature alignment loss in FAM is calculated as follows:

$$
\begin{aligned}
L_{D_i}^{FA} = -[&\sum_{i=1}^{N} log\frac{exp(\gamma R(x_i^G, y_i^G))}{\sum_{j=1}^{N} exp(\gamma R(x_i^G, y_j^G))} \\
&+ \sum_{i=1}^{N} log\frac{exp(\gamma R(x_i^G, y_i^G))}{\sum_{j=1}^{N} exp(\gamma R(x_j^G, y_i^G))} \\
&+ \sum_{i=1}^{N} log\frac{exp(\gamma R(x_i^R, y_i^R))}{\sum_{j=1}^{N} exp(\gamma R(x_i^R, y_j^R))} \\
&+ \sum_{i=1}^{N} log\frac{exp(\gamma R(x_i^R, y_i^R))}{\sum_{j=1}^{N} exp(\gamma R(x_j^R, y_i)^R)}]
\end{aligned}
\tag{4}
$$

where x_i^G and x_i^R represent the global and local image features extracted from discriminator, y_i^G and y_i^R represent the global and local image features extracted from pretrained Inception-v3 model, $R(\cdot)$represents the matching function, and γ is a smoothing factor.

2.4 Objective Function

Adversarial Loss. As with our baseline AttnGAN [16], we use the cross-entropy loss as our adversarial loss. The adversarial loss of discriminator is defined as:

$$
\begin{aligned}
L_{D_i}^{adv} = -\frac{1}{2}[&\mathbb{E}_{x\sim p_{data}} log D_i(x) + \mathbb{E}_{\hat{x}\sim p_{G_i}} log(1 - D_i(\hat{x})) \\
&+ \mathbb{E}_{x\sim p_{data}} log D_i(x, s) + \mathbb{E}_{\hat{x}\sim p_{G_i}} log(1 - D_i(\hat{x}, s))]
\end{aligned}
\tag{5}
$$

where x and $\hat{x}$ represent the real and generated images and s represents the text condition. The adversarial loss of generator is defined as:

$$
L_{G_i}^{adv} = -\frac{1}{2}[\mathbb{E}_{\hat{x}\sim p_{G_i}} log D_i(\hat{x}) + \mathbb{E}_{\hat{x}\sim p_{G_i}} log D_i(\hat{x}, s)] \tag{6}
$$

Cross-Modal Alignment Loss. The cross-modal alignment loss used on generator has the same functional template as the feature alignment loss on discriminator, which can motivates the generator to generate semantically consistent images. It is defined as:

$$
\begin{aligned}
L_{CMA} = -[&\sum_{i=1}^{N} log\frac{exp(\gamma R(x_i, y_i))}{\sum_{j=1}^{N} exp(\gamma R(x_i, y_j))} \\
&+ \sum_{i=1}^{N} log\frac{exp(\gamma R(x_i, y_i))}{\sum_{j=1}^{N} exp(\gamma R(x_j, y_i))}]
\end{aligned}
\tag{7}
$$

where (x_i, y_i) is the image-text pair, $R(\cdot)$ represents the matching function between image and text, γ is a smoothing factor. The cross-modal alignment loss contains two components, x_i represents the global image feature when y_i represents the sentence feature, and x_i represents the local image feature when y_i represents the word feature.

Generator Loss. Based on the adversarial loss $L_{G_i}^{adv}$, we add the cross-modal alignment loss L_{CMA}, thus the whole generator loss is defined as:

$$L_G = \sum_{i=0}^{N} L_{G_i}^{adv} + \lambda_1 L_{CMA} \tag{8}$$

Discriminator Loss. Based on the adversarial loss $L_{D_i}^{adv}$, we add the feature alignment loss $L_{D_i}^{FA}$, thus the whole discriminator loss is defined as:

$$L_D = \sum_{i=0}^{N} (L_{D_i}^{adv} + \lambda_2 L_{D_i}^{FA}) \tag{9}$$

3 Experiments

In this section, we will demonstrate the feasibility and effectiveness of our proposed innovations through comprehensive and rigorous experiments. Firstly, we introduce the datasets and metrics. Then, we quantitatively and qualitatively analyzed the superiority of our method. Finally, we verify the generalization performance of our method through ablation experiments.

3.1 Datasets and Metrics

To demonstrate the capability of our proposed ACMA-GAN, we conduct experiments on CUB [15] and COCO [5] datasets. The CUB dataset contains 200 bird categories with 11,788 images, where 150 categories with 8,855 images are used for training while the remaining 50 categories with 2,933 images for testing. There are ten text descriptions for each image in CUB dataset. The COCO dataset includes a training set with 80k images and a test set with 40k images. Each image in the COCO dataset has five text descriptions. We quantify the performance of ACMA-GAN in terms of Inception Score (IS) [11], Fréchet Inception Distance (FID) [3], and R-precision [16]. In the testing phase, 30 000 images are randomly generated.

3.2 Quantitative Results

In this subsection we will analyze the effect of our innovation points in terms of three metrics: Inception Score [11], FID [3] and R-Precision [16]. As shown in Table 1, ACMA-GAN obtains the second highest Inception Score and FID on CUB dataset, slightly behind TIME [6], but obtains the best R-precision. ACMA-GAN obtains the state-of-the-art results on all metrics for COCO dataset.

Compared with our baseline AttnGAN [16] on CUB and COCO datasets, ACMA-GAN improves 11.47% and 24.26% on Inception Score, improves 38.62% and 28.97% on FID, and improves 11.53% and 9.97% on R-Precision.

Table 1. Comparing the results of our ACMA-GAN with other advanced methods. AttnGAN [16] is our baseline model. The bold is best.

Methods	Inception Score↑		FID↓		R-Precision↑	
	CUB	COCO	CUB	COCO	CUB	COCO
StackGAN [21]	3.70	8.45	51.89	74.05	N/A	N/A
StackGAN++ [22]	4.04	8.30	15.30	81.59	N/A	N/A
AttnGAN [16]	4.36	25.89	23.98	35.49	67.82	83.53
ControlGAN [4]	4.58	24.06	N/A	N/A	69.33	82.43
SEGAN [14]	4.67	27.86	18.17	32.28	N/A	N/A
DM-GAN [23]	4.75	30.49	16.09	32.64	72.31	88.56
TIME [6]	**4.91**	30.85	**14.30**	31.14	71.57	89.57
ACMA-GAN	4.86	**32.17**	14.72	**25.21**	**75.64**	**91.86**

The quantitative results show that our ACMA-GAN model has significant advantages over other state-of-the-art methods in generating high-quality images and improving image diversity, as well as in maintaining semantic consistency, especially for COCO dataset of complex scenes (Fig. 3).

3.3 Qualitative Results

By comparing the images generated by ACMA-GAN with those generated by AttnGAN [16] and DM-GAN [23], qualitative results will indicate the validity of the generated images from a visual perspective.

As shown in Fig. 4. The text in column 2 gives "This bird is brown and white", however the abdomen of the bird generated by AttnGAN [16] is hardly noticeable as brown, and the abdomen of the bird generated by DM-GAN [23] is completely white, whereas the bird generated by our method is perfectly consistent with the semantic of the given text and has good morphology. The text in column 4 contains a description of "a large crowd", which is not reflected in the images generated by AttnGAN [16] and DM-GAN [23], while The images generated by our method reflect the concept of "a large crowd".

In summary, compared with the mainstream AttnGAN [16] and DM-GAN [23] our method can better understand the semantics of text descriptions and then synthesize images with consistent content, and the images generated by our method contain more well-posed objects and have higher image diversity.

3.4 Ablation Study

In this section, we first quantitatively analyze the use of added information in AWAM, then quantitatively analyze the proposed Adaptive Word Attention Module (AWAM) and Feature Alignment Module (FAM), and qualitatively analyze the differences of word attention between ACMA-GAN and AttnGAN [16]. Finally, we qualitatively analyze the generalization performance of ACMA-GAN.

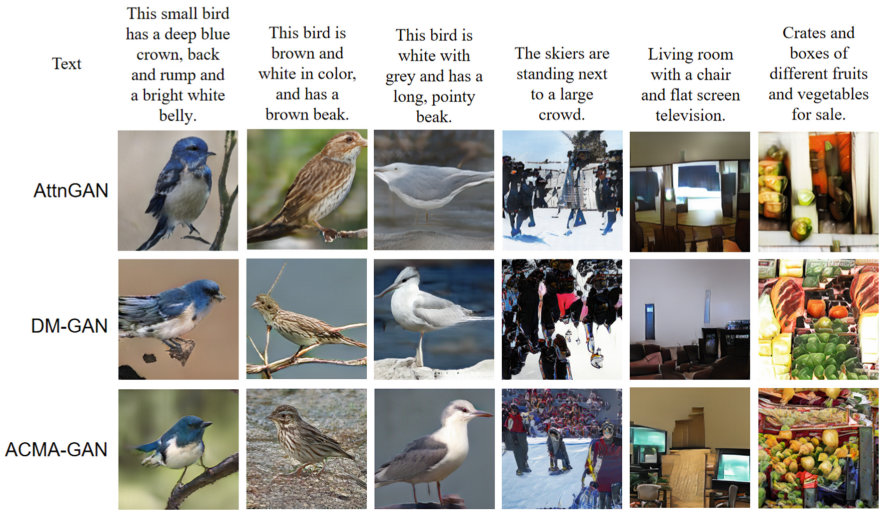

Fig. 3. Examples (CUB: 1st–3rd columns), COCO: (4th–6th columns) are generated by AttnGAN [16], DM-GAN [23] and our proposed ACMA-GAN.

Table 2. The results of using different added information in AWAM.

ID	Components		IS↑	FID↓	R-Precision↑
	SentFea	ImageFea			
0	-	-	4.57 ± 0.05	21.41	70.66 ± 0.69
1	✓	-	4.69 ± 0.05	16.85	73.12 ± 0.73
2	-	✓	4.76 ± 0.06	16.56	75.16 ± 0.76
3	✓	✓	4.81 ± 0.04	15.78	75.64 ± 0.54

Added Information. We fine-tune the AttnGAN [16] model and obtain a better baseline model. We analyzed in the introduction that DM-GAN [23] only uses the image features after the global average pooling as the modulation information to optimize the semantics of words, which can obtain better word attention results, but whether the sentence information also has this ability has not been explored. So we explored the effect of using sentence features and image features to obtain new words in the adaptive word attention module. As shown in Tables 2, when we used sentence features, both Inception Score, FID and R-Precision were improved, which indicates that the way to optimize the semantics of words with sentence features. When we use both sentence features and image features as modulation information to optimize the word semantics, our model can obtain the best results, which indicates that it is necessary to automatically optimize the word semantics according to the current image features and sentence features before using word attention, and then we can get the best attention results.

Table 3. The performance of AWAM and FAM on CUB dataset.

Architecture	Inception Score	FID	R-Precision
baseline	4.57	21.41	70.66
+AWAM	4.81	15.78	74.78
+FAM	4.83	16.96	74.20
+AWAM+FAM	4.86	14.72	75.64

Effectiveness of AWAM and FAM. As shown in Table 3, when we replace the word attention module of our baseline model with AWAM, the Inception Score, FID, and R-Precision are improved by 5.25%, 26.30%, and 5.83%, when we add the FAM to our baseline, the Inception Score, FID and R-Precision are improved by 5.69%, 20.78% and 5.01%, these results indicate that both the AWAM and FAM are effective. The best results are obtained by using both the AWAM and FAM on our baseline model, which indicates the proposed AWAM and FAM are mutually beneficial and not in conflict.

Attention Results. Qualitative analysis of word attention is shown in Fig. 4, where we show the top five words with the highest attention score and mark them with different colors in the text, and highlight the areas attentioned by each word in the image. The given text contains three descriptions: "long black legs," "brown feathers" and "black beak," but AttnGAN [16] mistakenly makes the brown feathers blue, but the generated image by our method is completely consistent with these three descriptions. The attention areas in AttnGAN [16] are not match the semantics of the words, which disturbs the generation of the image. In our method, "leg" and "beak" are accurately matched to the corresponding area, and the word "have" is the public description of "leg", "feather" and "beak", which should be related to the three words in the context of text. In our attention results, the word "have" noticed two areas "leg" and "beak", although not perfectly noticed all the relevant regions, it is also enough to show that our adaptive word attention can accurately focus on the word-related areas while maintaining the word context semantics.

Generalisation Ability. As shown in Fig. 5, when we change the partial description of the bird in the text, our ACMA-GAN can generates the images of the bird with the corresponding semantics. These results indicate that our ACMA-GAN is sensitive to the input text, and has good generalisation performance.

Fig. 4. Attention results for AttnGAN [16] and our ACMA-GAN at 128 resolution.

Fig. 5. Experiment of text sensitivity.

4 Conclusion

In this paper, we propose a novel Text-to-Image generation model ACMA-GAN. We design an Adaptive Word Attention Module (AWAM), which uses both image features and sentence features to modify the word embeddings. The updated words can better play the role of cross-modal attention. In addition, a feature alignment loss is designed to use the pre-trained image classification model to encourage the discriminator to extract more image features, so as to improve the feature extraction ability of the discriminator. Extensive experiments on two common datasets confirm that ACMA-GAN significantly outperforms other state-of-the-art methods.

References

1. Frolov, S., Hinz, T., Raue, F., Hees, J., Dengel, A.: Adversarial text-to-image synthesis: a review. Neural Netw. **144**, 187–209 (2021)
2. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
3. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
4. Lee, M., Seok, J.: Controllable generative adversarial network. IEEE Access **7**, 28158–28169 (2019)
5. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
6. Liu, B., Song, K., Zhu, Y., de Melo, G., Elgammal, A.: Time: text and image mutual-translation adversarial networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2082–2090 (2021)
7. Mirza, M., Osindero, S.: Conditional generative adversarial nets. Computer Science, pp. 2672–2680 (2014)
8. Miyato, T., Koyama, M.: cGANs with projection discriminator. arXiv preprint arXiv:1802.05637 (2018)
9. Qiao, T., Zhang, J., Xu, D., Tao, D.: MirrorGAN: learning text-to-image generation by redescription. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1505–1514 (2019)
10. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning, pp. 1060–1069. PMLR (2016)
11. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. Adv. Neural. Inf. Process. Syst. **29**, 2234–2242 (2016)
12. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Sig. Process. **45**(11), 2673–2681 (1997)
13. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
14. Tan, H., Liu, X., Li, X., Zhang, Y., Yin, B.: Semantics-enhanced adversarial nets for text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10501–10510 (2019)
15. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset (2011)
16. Xu, T., et al.: 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)
17. Xu, T., Feng, Z., Wu, X.J., Kittler, J.: Adaptive channel selection for robust visual object tracking with discriminative correlation filters. Int. J. Comput. Vis. **129**, 1359–1375 (2021)
18. Xu, T., Feng, Z., Wu, X.J., Kittler, J.: Toward robust visual object tracking with independent target-agnostic detection and effective Siamese cross-task interaction. IEEE Trans. Image Process. **32**, 1541–1554 (2023)

19. Xu, T., Zhu, X.F., Wu, X.J.: Learning spatio-temporal discriminative model for affine subspace based visual object tracking. Vis. Intell. **1**(1), 4 (2023)
20. Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2327–2336 (2019)
21. Zhang, H., et al.: 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)
22. Zhang, H., et al.: StackGAN++: realistic image synthesis with stacked generative adversarial networks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1947–1962 (2018)
23. Zhu, M., Pan, P., Chen, W., Yang, Y.: DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5802–5810 (2019)