

# Gated cross word-visual attention-driven generative adversarial networks for text-to-image synthesis

Borun Lai<sup>1</sup>, Lihong Ma<sup>1</sup>, and Jing Tian<sup>2</sup>(✉)

<sup>1</sup> School of Electronics Information Engineering, South China University of Technology, Guangzhou, China

[eebrlai@mail.scut.edu.cn](mailto:eebrlai@mail.scut.edu.cn), [eelhma@scut.edu.cn](mailto:eelhma@scut.edu.cn)

<sup>2</sup> Institute of Systems Science, National University of Singapore, Singapore, 119615  
[tianjing@nus.edu.sg](mailto:tianjing@nus.edu.sg)

**Abstract.** The main objective of text-to-image (Txt2Img) synthesis is to generate realistic images from text descriptions. We propose to insert a gated cross word-visual attention unit (GCAU) into the conventional multiple-stage generative adversarial network Txt2Img framework. Our GCAU consists of two key components. First, a cross word-visual attention mechanism is proposed to draw fine-grained details at different subregions of the image by focusing on the relevant words (via the visual-to-word attention), and select important words by paying attention to the relevant synthesized subregions of the image (via the word-to-visual attention). Second, a gated refinement mechanism is proposed to dynamically select important word information for refining the generated image. Extensive experiments are conducted to demonstrate the superior image generation performance of the proposed approach on CUB and MS-COCO benchmark datasets.

## 1 Introduction

The objective of text-to-image (Txt2Img) is to generate a realistic image from a given text description that is consistent with the text semantics. Deep learning techniques, particularly, Generative Adversarial Networks (GANs), have become an effective generative approach in Txt2Img synthesis [1, 2, 11]. It has many significant applications, such as image enhancement [10], text-image matching [8].

The GAN-based approaches encode the text description as a global sentence vector and then apply it as a conditional constraint to generate an image that matches the text description. They can be classified into two categories: (i) one-stage methods, and (ii) multiple-stage methods. The one-stage methods generate images by adding up-sampling layers in a single generator. However, this may cause the generated image to be inconsistent with the input text description [15]. Thus, a matching-aware zero-centered gradient penalty method is proposed in [19] to make the generated image better match the text description. The multiple-stage methods generate an initial low-resolution image by the first

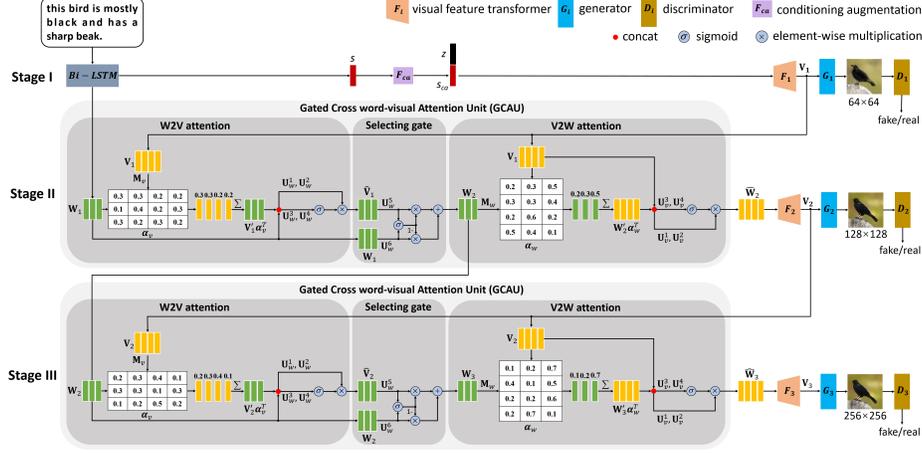
generator and then refine it by subsequent generators to create high-resolution progressively, where the global sentence feature is used as a conditional constraint to the discriminator at each stage to ensure that the generated image matches the text description [24, 25].

Considering that fine-grained details are critical in the generated image, the attention mechanism has been exploited for Txt2Img generation. AttnGAN [22] drew image details by computing the attention distribution of all word feature vectors on each visual feature vector. However, the unchanged text representation is used at each stage of image refinement. Moreover, if the attention weights are wrongly estimated at the beginning, some important word information may be ignored. An attention regularization loss was proposed in SEGAN [17] to highlight important words. A threshold is set so that the attention weight of important words (above the threshold) can be gradually increased and the attention weight of irrelevant words (below the threshold) can be gradually decreased. The limit is that it is not easy to determine the appropriate value range of the threshold. A dynamic memory mechanism was proposed in DM-GAN [27] to refine image details dynamically. The memory writing gate would select important word information according to the global image information and word information, and save them in memory slots. The memory is addressed and read according to the correlation between each subregion of the image and the memory, thus gradually completing the refinement of the image. Its limitation is that it considers the contribution of all subregions of the image to each word as equal. KT-GAN [18] focused on adjusting attention distribution by using hyperparameters to extract important word information. However, KT-GAN is a time-consuming method, and it requires accurately-estimated attention weight of each word.

The fundamental challenge in Txt2Img synthesis is how to exploit the information from the input sentence, which guides details generation in the image. Our approach yields the following two contributions.

- Firstly, each word in the input sentence provides different information depicting the image content. The image information should be taken into account to determine the importance of every word, and the word information should also be considered to determine the importance of every subregion of the image. For that, we propose a cross word-visual attention mechanism. It draws details at different image subregions by focusing on the relevant words via visual-to-word (V2W) attention, and select important words by focusing on the relevant image subregions via word-to-visual (W2V) attention.
- Secondly, if the same word representation is utilized at multiple phases of image refinement, the procedure may become ineffective. For that, we propose a gated refinement mechanism to dynamically select the important word information from the updated word representation based on the updated image representation at multiple image refinement stages.

We propose to include these two contributions into a multiple-stage GAN-based Txt2Img synthesis framework by combining them to construct a gated cross word-visual attention unit.



**Fig. 1.** The framework of our proposed approach. The proposed gated cross-word visual attention unit, which contains a W2V attention, a V2W attention, and a selecting gate, is used for the Stage II and Stage III.

The remainder of this paper is organized as follows. The proposed Txt2Img synthesis approach is presented in Sect. 2 by developing a gated cross word-visual attention method. It is evaluated in extensive experiments in Sect. 3. Finally, Section 4 concludes this paper.

## 2 Proposed Txt2Img synthesis approach

We leverage a conventional multiple-stage GAN-based Txt2Img framework, where a low-resolution initial image is firstly generated and then refined via several stages to obtain the final high-resolution synthesized image. Let  $\mathbf{V}_i$  and  $\mathbf{W}_i$  be visual features and word features, respectively.  $F_{ca}$  represents the Conditioning Augmentation [24] that converts the sentence vector to the conditioning vector,  $z \sim N(0, 1)$  is a random noise vector,  $F_i$  represents the visual feature transformer at the  $i$ -th stage,  $G_i$  represents the generator at the  $i$ -th stage,  $D_i$  represents the discriminator at the  $i$ -th stage.

As shown in Fig. 1, we propose to insert a gated cross word-visual attention unit (GCAU) at each stage (except the first stage) of this Txt2Img framework. Our GCAU contains a W2V attention, a V2W attention, and a selecting gate. These three components are described in detail as follows.

### 2.1 Cross word-visual attention

Denote the word feature matrix  $\mathbf{W}_i \in \mathbb{R}^{D_w \times N_w}$ , the visual feature matrix  $\mathbf{V}_i \in \mathbb{R}^{D_v \times N_v}$ , where  $D_w$  and  $D_v$  are dimensions of a word feature vector and a visual feature vector,  $N_w$  and  $N_v$  are numbers of word feature vectors and visual feature vectors.

**W2V Attention** Firstly, it transforms visual features from a visual semantic space to a word semantic space by a  $1 \times 1$  convolution operator  $\mathbf{M}_v(\cdot)$  to obtain a mapped visual feature matrix  $\mathbf{V}'_i \in \mathbb{R}^{D_w \times N_v}$  as

$$\mathbf{V}'_i = \mathbf{M}_v(\mathbf{V}_i). \quad (1)$$

Then, it calculates a similarity matrix  $\mathbf{W}'_i \mathbf{V}'_i$  between the mapped visual feature matrix  $\mathbf{V}'_i$  and the word feature matrix  $\mathbf{W}_i$ . By calculating the attention distribution  $\alpha_v \in \mathbb{R}^{N_w \times N_v}$  on each mapped visual feature vector, the normalized attention distribution is obtained as

$$\alpha_v = \text{softmax}(\mathbf{W}'_i \mathbf{V}'_i). \quad (2)$$

Next, according to the attention distribution  $\alpha_v$ , each mapped visual feature vector is weighted and summed up to obtain the visual-context feature matrix as  $\mathbf{V}'_i \alpha_v^T$ .

**V2W Attention** It follows a similar procedure as W2V as follows. Firstly, it applies a  $1 \times 1$  convolution operator  $\mathbf{M}_w(\cdot)$  to obtain a mapped word feature matrix  $\mathbf{W}'_i \in \mathbb{R}^{D_v \times N_w}$  as

$$\mathbf{W}'_i = \mathbf{M}_w(\mathbf{W}_i) \quad (3)$$

Then, it calculates the attention distribution  $\alpha_w \in \mathbb{R}^{N_v \times N_w}$  on each mapped word feature vector to obtain the normalized attention distribution as

$$\alpha_w = \text{softmax}(\mathbf{V}'_i \mathbf{W}'_i). \quad (4)$$

Next, each mapped word feature vector is weighted and summed up to obtain the word-context feature matrix as  $\mathbf{W}'_i \alpha_w^T$ .

Finally, following the idea of the Attention on Attention (AoA) method [4], we further concatenate the visual-context feature matrix and word feature matrix, then apply two separate linear transformations conditioned on the concatenated result. Then we add another attention using element-wise multiplication to eventually obtain the W2V attentional information  $\hat{\mathbf{V}}_i$  as

$$\hat{\mathbf{V}}_i = (\mathbf{U}_w^1 \mathbf{V}'_i \alpha_v^T + \mathbf{U}_w^2 \mathbf{W}_i + b_w^1) \otimes \sigma(\mathbf{U}_w^3 \mathbf{V}'_i \alpha_v^T + \mathbf{U}_w^4 \mathbf{W}_i + b_w^2), \quad (5)$$

where  $\sigma(\cdot)$  is the sigmoid activation function,  $\otimes$  denotes the element-wise multiplication,  $\mathbf{U}_w^1, \mathbf{U}_w^2, \mathbf{U}_w^3, \mathbf{U}_w^4 \in \mathbb{R}^{D_w \times D_w}$ ,  $b_w^1, b_w^2 \in \mathbb{R}^{D_w}$ . It highlights visual subregions that each word should pay attention to, it measures the word importance that will be sent to the selecting gate in the gated refinement mechanism for important word selection. The AoA method [4] is also applied to obtain the V2W attentional information.

## 2.2 Gated refinement

We propose a selecting gate to dynamically select the important word feature at different image refinement stages. It adopts a structure of a memory writing

gate [27], but we modify it in two ways. Firstly, we use the word information refined from the previous stage as the input, instead of the fixed initial word information in [27]. Secondly, we adaptively combine features from different visual subregions according to the W2V attentional information, instead of treating them equally in [27].

Our selecting gate is defined as follows. It inputs the previous word information  $\mathbf{W}_{i-1}$  and the W2V attentional information  $\hat{\mathbf{V}}_{i-1}$ , which are firstly transformed by linear transformations  $\mathbf{U}_w^5, \mathbf{U}_w^6 \in \mathbb{R}^{1 \times D_w}$ , and normalized by the sigmoid activation function  $\sigma(\cdot)$  as

$$g(\mathbf{W}_{i-1}, \hat{\mathbf{V}}_{i-1}) = \sigma(\mathbf{U}_w^5 \mathbf{W}_{i-1} + \mathbf{U}_w^6 \hat{\mathbf{V}}_{i-1}). \quad (6)$$

Then, it removes the past word information to be forgotten and obtains the attentional information to be memorized as

$$\mathbf{W}_i = g(\mathbf{W}_{i-1}, \hat{\mathbf{V}}_{i-1}) \hat{\mathbf{V}}_{i-1} + (1 - g(\mathbf{W}_{i-1}, \hat{\mathbf{V}}_{i-1})) \mathbf{W}_{i-1}. \quad (7)$$

### 2.3 Objective function

The objective function is defined by combining all stages of image refinement as

$$L = \sum_i L_{G_i} + \lambda_1 L_{CA} + \lambda_2 \sum_i L_{DAMSM}(\mathbf{W}_i), \quad (8)$$

where  $L_{G_i}$  is an adversarial loss [22],  $\lambda_1$  and  $\lambda_2$  are the corresponding weights of a conditioning augmentation loss  $L_{CA}$  [27] and a loss  $L_{DAMSM}(\mathbf{W}_i)$ , which is modified from the DAMSM loss [22].  $L_{G_i}$  encourages the generated image to be realistic and match the given text description,  $L_{CA}$  avoids overfitting in model training, and  $L_{DAMSM}(\mathbf{W}_i)$  encourages each subregion of the image to match each word in the given text description as much as possible.

It is important to note that we use the refined word information  $\mathbf{W}_i$  in the objective function. We compute the DAMSM loss between the generated image and the refined word information at each stage of image refinement. This is different from the initial word information  $\mathbf{W}$  that is used in [22]. This modified loss enables the generated image to match the text description by ensuring the semantic consistency between the visual subregions and the selected word information at each stage. It also enables our W2V attention to accurately highlight visual subregions more relevant to the word.

### 2.4 Summary of our method

At the  $i$ -th stage of image refinement, our GCAU takes the previous word feature  $\mathbf{W}_{i-1}$  and visual feature  $\mathbf{V}_{i-1}$  as the inputs, and output the current word feature  $\mathbf{W}_i$  and visual feature  $\mathbf{V}_i$  as follows.

Step 1. Apply the W2V attention to calculate the current W2V attentional information  $\hat{\mathbf{V}}_{i-1}$  from  $\mathbf{W}_{i-1}$  and  $\mathbf{V}_{i-1}$ .

Step 2. Apply the selecting gate to select current important word information  $\mathbf{W}_i$  from  $\hat{\mathbf{V}}_{i-1}$  and  $\mathbf{W}_{i-1}$ .

Step 3. Apply the V2W attention to calculate the V2W attentional information  $\hat{\mathbf{W}}_i$  from  $\mathbf{W}_i$  and  $\mathbf{V}_{i-1}$ . Then we concatenate  $\hat{\mathbf{W}}_i$  with itself and input to the visual feature transformer to obtain the updated visual feature  $\mathbf{V}_i$ .

Step 4. Generate the refined image  $\mathbf{I}_i$  from  $\mathbf{V}_i$  via  $\mathbf{G}_i$ .

### 3 Experimental Results

Extensive experiments in this section are carried out to compare our proposed approach with other previous state-of-the-art Txt2Img synthesis approaches to verify the performance of our approach. In addition, we conduct an ablation study to verify the performance of each component.

#### 3.1 Datasets

We use two public benchmark datasets, including CUB [20] and MS-COCO [9] datasets. The CUB [20] dataset is a single-object dataset with 200 categories, in which the training set contains 8,855 images and the test set contains 2,933 images. There are ten text descriptions for each image. The MS-COCO [9] dataset is a multi-object dataset, in which the training set contains 82,783 images and the test set contains 40,470 images. There are five text descriptions for each image.

#### 3.2 Implementation details

We use the bidirectional LSTM as text encoder to encode the input text description to obtain the word features and the sentence feature. An image with  $64 \times 64$  resolution is generated at the initial stage, and then refined to generate images with  $128 \times 128$  and  $256 \times 256$  resolution.  $D_v$  and  $D_w$  are set to 64 and 256, respectively.  $N_w$  is set to 64 and  $N_v$  is the resolution of the generated image at each stage, which is set to  $64 \times 64$ ,  $128 \times 128$ ,  $256 \times 256$ . The model is trained on a Nvidia GeForce RTX 2080 Ti GPU. The batch size is set to 16 on the CUB dataset and 12 on the MS-COCO dataset. All models are optimized with the ADAM optimizer [5],  $\beta_1$  and  $\beta_2$  are set to 0.5 and 0.999. The learning rate of generators and discriminators are set to 0.0002. The model is trained for 800 epochs on the CUB dataset and 200 epochs on the MS-COCO dataset. For  $\lambda_1$  and  $\lambda_2$  in Eq. (8),  $\lambda_1$  is set to 1 on the CUB dataset and MS-COCO dataset,  $\lambda_2$  is set to 5 on the CUB dataset and 50 on the MS-COCO dataset.

#### 3.3 Evaluation metrics

We evaluate the model generation performance by generating 30,000 images based on the text descriptions from unseen test set. There are three metrics used for evaluation: IS [16], FID [3], and R-precision [22]. IS [16] is used to evaluate the diversity of the generated images, FID [3] is used to evaluate the reality of the generated images, and R-precision [22] is used to evaluate how well the generated images match the text descriptions.

- The IS [16] metric calculates the Kullback-Leibler divergence between the conditional class distribution and the marginal class distribution. A higher IS means the generated images have more diversity.
- The FID [3] metric calculates the Fréchet distance between synthetic and real-world images. It first uses the Inception v3 network to extract features, then uses a Gaussian model to model the feature space, and finally calculates the distance between the two features. A lower FID means that the generated images are closer to real-world images.
- The R-precision [22] metric measures the cosine similarity between global image features and candidate sentence features. A higher R-precision means that the generated images match the text descriptions better.

### 3.4 Experimental results

The comparison results of our approach with other previous state-of-the-art approaches on the test set are shown in Table 1, Table 2 and Table 3. The following is the analysis report of the comparison results.

Firstly, as seen in the IS performance in Table 1, our method performs only worse than TVBi-GAN [21] and DF-GAN [19] on the CUB dataset, and only worse than SD-GAN [23] on the MS-COCO dataset. Although SD-GAN [23] trains the model with multiple text descriptions, our method only uses a single text description, which may lead to a possible limitation of our method in the diversity of generated images. Moreover, SD-GAN [23] will fail to train if each image in the dataset contains only a single text description. In addition, SD-GAN [23] uses the siamese structure to extract text semantic information, which is more complex than our network and more powerful hardware equipment is required for training.

Secondly, as seen in the FID performance in Table 2, our method performs only worse than TVBi-GAN [21] on the CUB dataset and achieves the best performance on the MS-COCO dataset. Our method performs worse than TVBi-GAN [21] on the CUB dataset, but the CUB dataset is a single-object dataset, while the MS-COCO dataset is a multi-object dataset, and our method achieves the best performance on the MS-COCO dataset which proves that our method performs better in generating multi-object images.

Thirdly, as seen in the R-precision performance in Table 3, our method achieves the state-of-the-art performance on the CUB dataset and performs only worse than Obj-GAN [7] on the MS-COCO dataset. Obj-GAN [7] uses a discriminator based on the Fast R-CNN model to provide rich object-wise discrimination signals, which helps semantic alignment of text descriptions and images. This also complicates the network. Our method does not need to add additional networks, and the performance is very close.

**Table 1.** The performance of IS for our proposed method comparing with other methods on CUB and MS-COCO test sets. Higher IS means better performance.

Methods	CUB [20]	MS-COCO [9]
GAN-INT-CLS [15]	2.88±.04	7.88±.07
GAWWN [14]	3.62±.07	-
StackGAN [24]	3.70±.04	8.45±.03
StackGAN++ [25]	4.04±.05	-
HD-GAN [26]	4.15±.05	11.86±.18
AttnGAN [22]	4.36±.03	25.89±.47
MirrorGAN [13]	4.56±.05	26.47±.41
ControlGAN [6]	4.58±.09	24.06±.60
LeicaGAN [12]	4.62±.06	-
SEGAN [17]	4.67±.04	27.86±.31
ObjGAN [7]	-	30.29±.33
DM-GAN [27]	4.75±.07	30.49±.57
SD-GAN [23]	4.67±.09	35.69±.50
DF-GAN [19]	4.86±.04	-
TVBi-GAN [21]	5.03±.03	31.01±.34
<b>Ours</b>	<b>4.79±.05</b>	<b>31.22±.58</b>

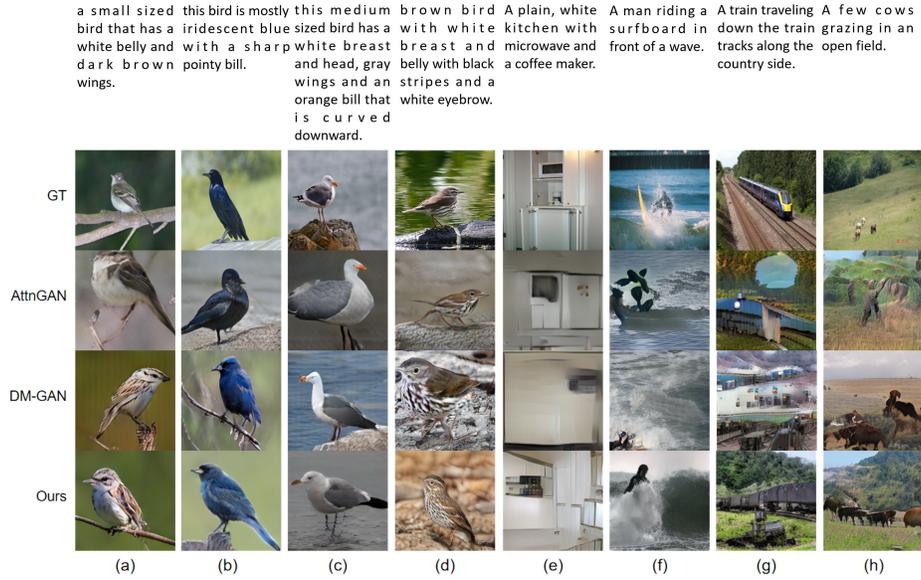
**Table 2.** The performance of FID for our proposed method comparing with other methods on CUB and MS-COCO test sets. Lower FID means better performance.

Methods	CUB [20]	MS-COCO [9]
StackGAN [24]	51.89	74.05
AttnGAN [22]	23.98	35.49
SEGAN [17]	18.17	32.28
DM-GAN [27]	16.09	32.64
TVBi-GAN [21]	11.83	31.97
Obj-GAN [7]	-	25.64
KT-GAN [18]	17.32	30.73
<b>Ours</b>	<b>15.16</b>	<b>25.49</b>

**Table 3.** The performance of R-precision for our proposed method comparing with other methods on CUB and MS-COCO test sets. Higher R-precision means better performance.

Methods	CUB [20]	MS-COCO [9]
AttnGAN [22]	67.82	85.47
MirrorGAN [13]	57.67	74.52
ControlGAN [6]	69.33	82.43
DM-GAN [27]	72.31	88.56
Obj-GAN [7]	-	91.05
<b>Ours</b>	<b>78.07</b>	<b>90.97</b>

Lastly, we compare our method with SEGAN [17] and DM-GAN [27]. Firstly, as seen in the IS performance in Table 1, our method achieves 2.57% higher



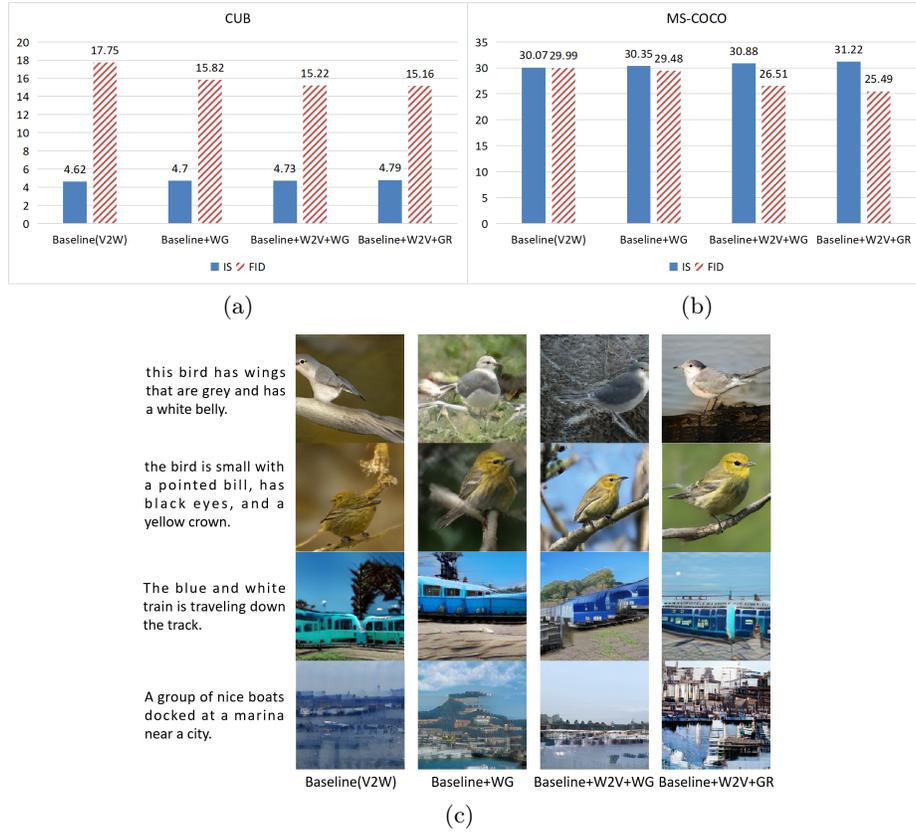
**Fig. 2.** The performance comparison of ground truth images and images generated by AttnGAN [22], DM-GAN [27] and our method. The four columns on the left are from the CUB [20] dataset, and the four columns on the right are from the MS-COCO [9] dataset.

than SEGAN [17], 0.84% higher than DM-GAN [27] on the CUB dataset, and 12.06% higher than SEGAN [17], 2.39% higher than DM-GAN [27] on the MS-COCO dataset. Secondly, as seen in the FID performance in Table 2, our method achieves 16.56% lower than SEGAN [17], 5.77% lower than DM-GAN [27] on the CUB dataset, and 21.03% lower than SEGAN [17], 21.90% lower than DM-GAN [27] on the MS-COCO dataset. Thirdly, as seen in the R-precision performance in Table 3, our method also achieves better performance than DM-GAN [27].

As can be seen from Fig. 2, for single-object generation and multi-object generation, the shapes of the generated images are more realistic and the generated images also have more details, such as the black stripes and white eyebrow in Fig. 2(d) and the microwave in Fig. 2(e). This verifies that our method can generate more realistic images with more details.

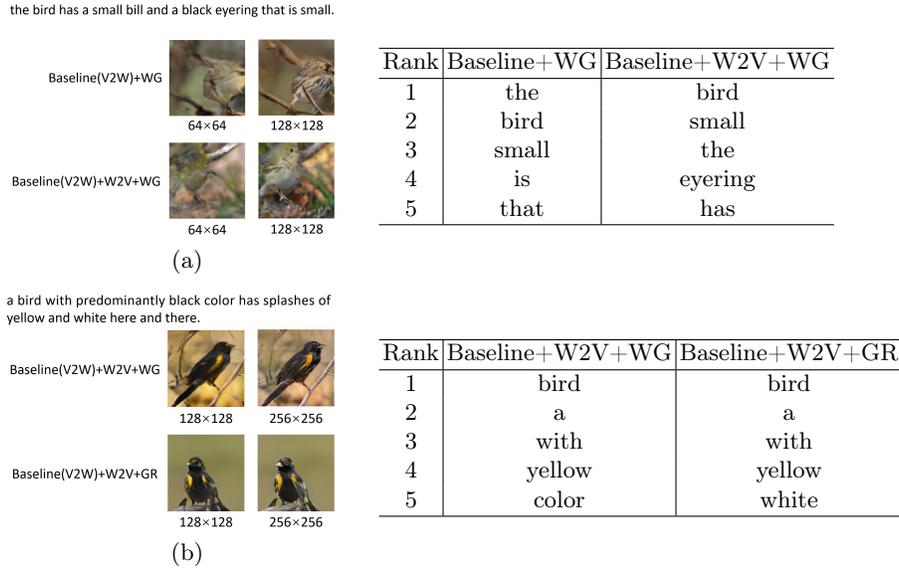
### 3.5 Ablation study

Our work is to improve the V2W attention by integrating W2V attention and gated refinement, which enables V2W attention to pay more attention to important words. We conduct an ablation study to gradually integrate various components and evaluate the model performance using IS and FID based on the CUB and MS-COCO dataset. As shown in Fig. 3(a) and Fig. 3(b), the perfor-



**Fig. 3.** (a) and (b) are the results of the ablation study on the CUB and MS-COCO datasets. (c) is the visualization of the ablation study. Baseline denotes that only V2W attention is integrated, WG denotes writing gate [27], W2V denotes our proposed W2V attention, and GR denotes our proposed gated refinement.

mance of our model on IS and FID improved progressively with each component being integrated, which demonstrates the effective contribution of each component. We also show the images by gradually integrating various components. As shown in Fig. 3(c), for the first text description, the generated objects obviously do not yet have the correct shape, and the important words “bird”, “wings”, and “belly” have not been accurately positioned and highlighted when only the V2W attention is integrated; the shape of bird is highlighted after the writing gate with fixed word features is integrated; the shape of bird is formed and the important words “wings” and “belly” have been accurately positioned and highlighted after the W2V attention is integrated; the image details corresponding to “bird”, “wings”, and “belly” have been significantly enhanced after the gated refinement with refined word features is integrated. For the third text description, the shape of the object is more realistic after the writing gate with fixed word features is



**Fig. 4.** (a) shows the refinement effect of W2V attention on the initial image with  $64 \times 64$  resolution. (b) shows the refinement effect of gated refinement on the intermediate image with  $128 \times 128$  resolution. The tables show the top-5 words that V2W attention pays attention to.

integrated; the color of the object is more accurate after the W2V attention is integrated; the object has the most realistic shape and the most accurate color after the gated refinement with refined word features is integrated.

To further verify how our proposed W2V attention and gated refinement improve V2W attention, we visualize the top-5 words that V2W attention pays attention to. As shown in Fig. 4(a), the attention weights of important words “bird”, “small” and “eyering” are improved after the W2V attention is integrated, which means that V2W attention pays more attention to these important words. We can also see that the shape of the object in the initial image can be effectively improved after W2V attention is integrated, which is due to the fact that W2V attention can focus on the relevant subregions of the image to select important words, instead of treating each subregion of the image equally in the writing gate [27]. As shown in Fig. 4(b), V2W attention can still pays more attention to important words “bird” and “yellow” after gated refinement is integrated. In addition, the attention weight of the important word “white” is improved, and we can also see that the details on the wings of the object in the final image are richer, which is due to the fact that gated refinement can retain important word information selected at the previous stage.

## 4 Conclusions

A new Txt2Img synthesis approach has been proposed in this paper by incorporating a gated cross word-visual attention unit into the multiple-stage GAN-based image generation framework. Our approach reconstructs images with better quality and visually realistic images, as verified in our qualitative and quantitative results using two benchmark datasets.

**Acknowledgement.** The work described in this paper is supported by China GDSF No. 2019A1515011949.

## References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y.: Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. pp. 2672–2680. Montreal, Canada (Dec 2014)
2. Gregor, K., Danihelka, I., Graves, A.: DRAW: A recurrent neural network for image generation. In: *Int. Conf. on Machine Learning*. pp. 1462–1471. Lille, France (Jul 2015)
3. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale upyear rule converge to a local nash equilibrium. In: *Advances in Neural Information Processing Systems*. pp. 6626–6637. Long Beach, CA, USA (Dec 2017)
4. Huang, L., Wang, W., Chen, J., Wei, X.: Attention on attention for image captioning. In: *IEEE/CVF Int. Conf. on Computer Vision*. pp. 4633–4642. Seoul, Korea (Oct 2019)
5. Kinga, D., Adam, J.B.: A method for stochastic optimization. In: *International Conference on Learning Representations*. vol. 5. San Diego, CA (May 2015)
6. Li, B., Qi, X., Lukasiewicz, T., Torr, P.: Controllable text-to-image generation. In: *Advances in Neural Information Processing Systems*. pp. 2065–2075. Vancouver, BC, Canada (Dec 2019)
7. Li, W.: Object-driven text-to-image synthesis via adversarial training. In: *IEEE Int. Conf. on Computer Vision and Pattern Recognition*. pp. 12166–12174. Long Beach, CA, USA (Jun 2019)
8. Li, W., Zhu, H., Yang, S., Zhang, H.: DADAN: dual-path attention with distribution analysis network for text-image matching. *Signal, Image and Video Processing* **16**(3), 797–805 (2022)
9. Lin, T., Maire, M.: Microsoft COCO: Common objects in context. In: *European Conf. on Computer Vision*. pp. 740–755. Zurich, Switzerland (Jun 2014)
10. Lu, Z., Chen, Y.: Single image super-resolution based on a modified U-net with mixed gradient loss. *Signal, Image and Video Processing* **16**(5), 1143–1151 (2022)
11. Pathak, D., Krahenbuhl, P., Donahue, J.: Context encoders: Feature learning by inpainting. In: *IEEE Int. Conf. on Computer Vision and Pattern Recognition*. pp. 2536–2544. Las Vegas, NV, USA (Jun 2016)
12. Qiao, T., Zhang, J., Xu, D., Tao, D.: Learn, imagine and create: Text-to-image generation from prior knowledge. In: *Neural Information Processing Systems* (2019)
13. Qiao, T., Zhang, J., Xu, D., Tao, D.: MirrorGAN: Learn-ing text-to-image generation by redescription. In: *IEEE Int. Conf. on Computer Vision and Pattern Recognition*. pp. 1505–1514. Long Beach, CA, USA (Jun 2019)

14. Reed, S., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. *New Republic* (2016)
15. Reed, S.E., Akata, Z., Yan, X.: Generative adversarial text to image synthesis. In: *Int. Conf. on Machine Learning*. pp. 1060–1069. New York City, NY, USA (Jun 2016)
16. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Xi, C.: Improved techniques for training GANs. In: *Advances in Neural Information Processing Systems*. pp. 2234–2242. Barcelona, Spain (Dec 2016)
17. Tan, H., Liu, X., Li, X., Zhang, Y., Yin, B.: Semantics-enhanced adversarial nets for text-to-image synthesis. In: *IEEE/CVF Int. Conf. on Computer Vision*. pp. 10500–10509. Seoul, Korea (Oct 2019)
18. Tan, X., Liu, M., Yin, B., Li, X.: KT-GAN: Knowledge-transfer generative adversarial network for text-to-image synthesis. *IEEE Trans. on Image Processing* **30**, 1275–1290 (Oct 2021)
19. Tao, M., Tang, H., Wu, S., Sebe, N., Wu, F., Jing, X.Y., Bao, B.: DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865* (2020)
20. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. *Tech. Rep. CNS-TR-2011-001*, California Institute of Technology (2011)
21. Wang, Z., Quan, Z., Wang, Z.J., Hu, X., Chen, Y.: Text to image synthesis with bidirectional generative adversarial network. In: *IEEE Int. Conf. on Multimedia and Expo*. pp. 1–6. London, UK (Jul 2020)
22. Xu, T.: AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: *IEEE Int. Conf. on Computer Vision and Pattern Recognition*. pp. 1316–1324. Salt Lake City, UT, USA (Jun 2018)
23. Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. In: *IEEE Int. Conf. on Computer Vision and Pattern Recognition*. pp. 2322–2331. Long Beach, CA, USA (Jun 2019)
24. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: *IEEE Int. Conf. on Computer Vision*. pp. 5907–5915. Venice, Italy (Oct 2017)
25. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **41**(8), 1947–1962 (Aug 2019)
26. Zhang, Z., Xie, Y., Yang, L.: Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018)
27. Zhu, M., Pan, P., Chen, W., Yang, Y.: DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In: *IEEE Int. Conf. on Computer Vision and Pattern Recognition*. pp. 5795–5803. Long Beach, CA, USA (Jun 2019)