Multimedia Systems (2023) 29:3291–3303
https://doi.org/10.1007/s00530-023-01160-4

REGULAR PAPER

Generative adversarial text-to-image generation with style image constraint

Zekang Wang[1] · Li Liu[1] · Huaxiang Zhang[1] · Dongmei Liu[1] · Yu Song[1]

Received: 1 July 2023 / Accepted: 5 August 2023 / Published online: 18 August 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
Most text-to-image generation works focus on the semantic consistency and neglect the style of the generated image. In this paper, a novel text-to-image generation method is proposed to generate image with style image constraint. In order to provide more comprehensive information by mining long–short-range information dependencies, the multi-group attention module is introduced to capture the multi-scale dependency information in the semantic feature. The adaptive multi-scale attention normalization is adopted to pay the multi-scale style feature attention in the style fusion process. The style information related to semantic feature is filtered out by the style feature attention. This selected style information is transferred to the generated results by aligning the mean and variance of the semantic feature and the style feature. Experiments conducted on common datasets show the validity of the proposed approach.

Keywords Text-to-image generation · Style fusion · Generative adversarial networks · Deep feature learning

1 Introduction

With the development of cross-modal research, the construction of other modal information using text modal has become a popular research direction. Compared with the conversion between other modal, the generated image that conforms to text semantics by given text information becomes a more concerned object. The emergence of generative adversarial networks (GANs) [1] promotes the development of text-to-image generation. Stacked networks based on GANs [2] improve the generation effect of the early generative network. Xu et al. [3] introduce the attention mechanism to select important words and refine the image by the selected word feature to obtain high-resolution result. The methods of mirror text comparison [4] and prior knowledge guidance [5] are proposed to obtain the higher quality image. Among them, the mirror text comparison method is realized by using additional semantic alignment between the generated image and the text regenerated by the generated image, and the prior knowledge guidance method is first learning the prior knowledge and then combining the knowledge to generate the image. Ruan et al. [6] improve the image effect by introducing multi-level information. Cheng et al. [7] select and optimize titles from prior knowledge to enrich the given description. This improves the image quality by providing additional visual details. Zhang et al. [8] introduce the idea of comparative learning. It improves image authenticity and semantic fidelity by capturing multiple contrast losses between inter-modality and intra-modality. More recently, single-stage networks are emerging again. Liao et al. [9] utilize semantic mask to promote the fusion of textual and visual information.

The above work mainly starts from truth degree and semantic fidelity of the generated image and achieves relatively successful results. Although most of the simple image generation in Fig. 1a cannot fully meet the expectation of users in a certain area, the utilization of additional constraint as image generation condition gradually comes into view of people. In the generative network based on GANs, various forms of constraints are used to control image generation. Park et al. [10] generate a new image by synthesizing the background of the real image from the dataset and the new object described by the text description of the specific position. Li et al. [11] propose a model that uses additional multiple texts for image group

Communicated by B. Bao.

✉ Li Liu
liuli_790209@163.com

[1] School of Information Science and Engineering, Shandong Normal University, Jinan, China

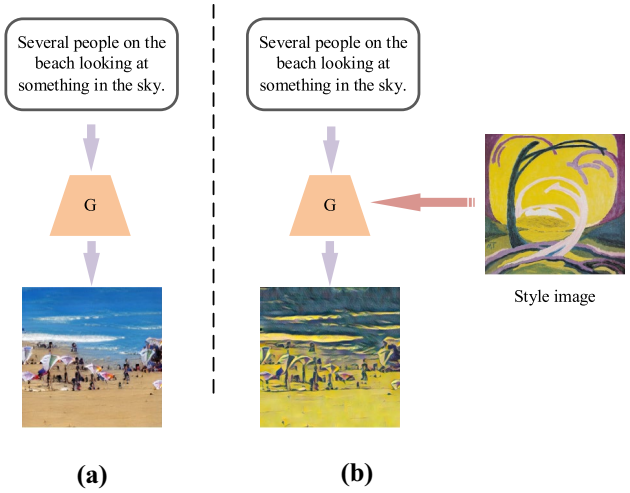

Fig. 1 **a** Text-to-image generation; **b** Style-constrained text-to-image generation

generation. This model visualizes the story content with multiple sentences.

At present, the input for image generation with style image constraints is usually image rather than text. Image generation with style image constraints based on text input is a challenging task. Unlike simple constraints such as adding text boxes in previous work, extracting the feature of style image as constraints affects semantic feature and complicates the generation process, which is not conducive to improving the semantic preservation of the generated image. In addition, it is practical to use style image constraints to achieve oriented generation to meet user needs. The style image constraint network with text as input in Fig. 1b is designed to address the above issues. This network adopts text and style image as constraints to generate image with a specific style and matching text. In this paper, the multi-group attention module and the adaptive multi-scale attention normalization are proposed to support generate image. The multi-group attention module is used to focus on the long–short-range information dependencies. The adaptive multi-scale attention normalization focuses on the important feature relationship between semantic feature and style feature and completes the feature fusion. The contributions of our work can be summarized as follows:

- The generation architecture with multiple constraints based on generative adversarial networks is proposed. This framework completes the text-to-image generation with style image constraint by the dual constraints of text and style image.
- The multi-group attention module is adopted to capture long–short-range information dependencies. This module focuses on multi-scale information to obtain more visual details and improve the semantic consistency between the generated image and the given text.
- The adaptive multi-scale attention normalization module is utilized to achieve the particular style fusion. Based on the attention of the multi-scale feature, this module adjusts the mean and variance of the multi-scale semantic feature and the style feature, and achieves the fusion of semantic feature and the style feature. The multi-scale information is utilized for rich details, enabling the generated image to obtain more corresponding style colors and textures.

The rest of this paper is organized as follows. In Sect. 2, some work in related areas is presented. In Sect. 3, the proposed method is elaborated. In Sect. 4, the experimental result and analysis are provided. Section 5 introduces the conclusion of this paper.

2 Related work

2.1 GANs for text-to-image generation

In recent years, GANs become a popular method in the field of text-to-image generation. In the generation framework based on GANs, multi-stage refinement frameworks [3, 6, 12, 13] are the mainstream now. This type of framework generates the semantic image through sentence embedding and refines the image through multiple pairs of discriminators and generators. AttnGAN [3] pays attention to fine-grained local information in the process of image refinement to enrich the details of the target object in the image. DAE-GAN [6] uses the object-level semantic extraction method to sketch the whole object with additional sub-global information. These are helpful for layout and detail generation of the whole image. SDGAN [12] uses parallel networks to obtain common semantic information in the textual description and preserves the diversity of information. RiFeGAN [13] performs caption matching to enrich textual information and feeds the enriched feature into the network.

All of the above methods enhance the generating effect of high-resolution image. However, these multi-stage frameworks cause higher computational burden and instability in the training process. Recently, single-stage networks [9, 14] are emerging again to solve the disadvantages of multi-stage GANs. DF-GAN [14] has only a single-stage text-to-image backbone. This backbone consists of a series of cyclic blocks that are responsible for deep fusion of image and text information and upsampling of image feature. SSA-GAN [9] directs the text–image fusion spatially through masks. These single-stage networks can generate image of higher quality and reduce excessive computation. In addition, Ramesh et al. [7] use transformer to model text and image

equally and use large-scale data for training. This makes it possible to achieve good results even in the case of zero shot in other tasks. The large-scale data-trained transformer is utilized as encoder for the proposed network. However, none of the above methods take into account the acquisition of long–short-range information dependencies to enrich the detail, leaving some potential connections unexplored. Instead of the above methods, the multi-group attention method attempts to focus on the long–short-range information relationship in the proposed network.

2.2 Image style transfer

In the early days, image style transfer is solved by using methods that relied on low-level statistics: non-parametric sampling [15], and non-photo-realistic rendering [16]. These methods often rely on low-level statistics and fail to capture semantics. Gatys et al. [17] iteratively update the base image formed by random noise to make it resemble to the style image in style texture and the original image in content, which is very slow. Johnson et al. [18] add an autoencoder-shaped feedforward network to fit the process of style transfer, which improved the speed to some extent, but each model could only transfer one style image. Huang et al. [19] believe that different styles are determined by the variance and mean of the feature, so that style transfer can be realized by making the variance and mean of the generated image the same as that of the style image. The affine parameter of instance normalization layer is adjusted adaptively to realize the transformation of arbitrary style. Park et al. [20] realize efficient style transfer by mapping the relationship between content and style image through the attention mechanism. Liu et al. [21] adopt a multi-level strategy that combines adaptive instance normalization and attention mechanism to further enhance the effect of style transfer. Recently, Park et al. [22] utilize the transformer structure to achieve efficient style transfer. Different from the above methods, we aim to introduce the principle of multi-group attention into the construction of style transfer, and try to improve the key component of style transfer called adaptive instance normalization as part of style-constrained image generation.

3 The proposed method

The proposed method generative adversarial text-to-image generation with style image constraint (SC-GAN) is shown in Fig. 2. The generation process includes two stages: semantic image generation and style-constrained image generation. In the semantic image generation stage, the semantic image is obtained by feature extraction, feature fusion, and feature upsampling. In the style-constrained image generation stage, the semantic feature is transformed into style-constrained image by style fusion and feature upsampling.

3.1 Semantic image generation

In this part, the process from text description to semantic image generation is realized. First, text information is

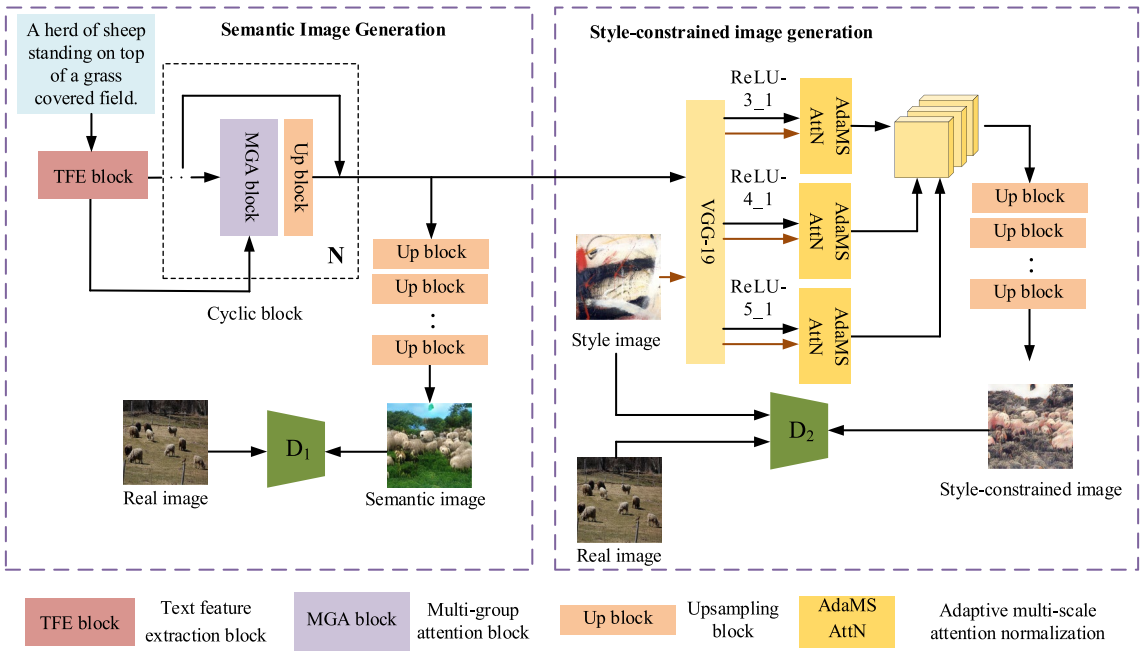

Fig. 2 The generative adversarial text-to-image generation with style image constraint

extracted from the text description by the text feature extraction module. Then, the multi-scale attention block is used to focus on the embedded feature, and the embedded feature is upsampled by the upsampling block. Finally, multiple convolution layers are used to generate semantic image.

3.1.1 Text feature extraction module

To obtain a suitable text embedding, text feature extraction block is utilized to process natural language descriptions. This block contains a text encoder and a component for frequency reduction of high-frequency data.

The text encoder based on transformer [23] is adopted as the text encoder of the proposed network. This encoder contains the n-pair multi-head self-attention block and the feed forward block. Compared to the text encoder in previous networks, the semantic feature extracted by this encoder is more easily mapped to the image semantic space of SC-GAN. Before the text is input to the encoder, a text description in the corresponding text group of each image is randomly selected to form a text–image pair with the image. This text–image pair is used for the subsequent multimodal semantic alignment work. As shown in Fig. 3, this text description is fed into the encoder. The resulting output is obtained as a sentence embedding $e \in R^{512}$ of the text. In addition, the output of a hidden layer is used as an additional text cue $w \in R^{n\times 512}$ for SC-GAN to enrich the visual detail. The combined use of global and local information guides and enriches the generation of visual detail. According to the global and local feature of image, the two text embeddings are used for the semantic alignment of the global and local text vision.

Since the convergence rate of high-frequency data is lower than that of low-frequency data [24], in order to achieve consistent and rapid convergence of data with different frequencies, it is attempted to stretch the different frequency data of the text embedding e into consistent low-frequency data. The process of unifying data of different frequencies into low-frequency data is the process of assigning different coefficients to data. In the concrete implementation, the noise vector sampled in Gaussian distribution z is combined with text embedding e to increase the diversity of image generation. Then the spliced vector is multiplied by the different coefficients 1, 2, 3, and 4. Finally, a set of text vectors S is obtained and used for the image generation process.

3.1.2 Multi-group attention module

Most of the previous work has often overlooked the joint multi-scale information, so it is conducive to make full use of the multi-scale feature information in feature fusion. By using convolution kernels of different sizes, the multi-scale information is obtained by mining long–short-range information dependencies in the proposed multi-group attention (MGA) model, and the importance of multi-scale information is selected. In other words, convolution kernels of different sizes correspond to different attention groups, and the features by different convolution kernels are focused separately. The selected information is used for feature fusion. Given the advantages of the deep fusion network, multiple cyclic blocks are configured as an important part of our backbone network. The N-pair multi-group attention block and the upsampling blocks are adopted as cyclic block in the process of dimension raising of the text vector group. The cyclic blocks pay attention to the generated feature for N times of multi-scale information and upsample the generated feature. This process implements detailed text semantic embedding of different dimensional features.

The specific structure of the multi-group attention module is shown in Fig. 4, the input of the multi-group attention block is the text feature t_i and the image feature v_i, where the text feature is the local text embedding w by the dimension transformation. The text feature t_i and the image feature v_i pass through a convolution layer whose convolution kernel is 5×5. In order to use receptive fields of different scales to fuse multi-scale context information, text feature and image feature are convolved with convolution layer group of size $(1\times7, 7\times1)$, $(1\times11, 11\times1)$, $(1\times21, 21\times1)$ to obtain text feature $T(t_i^7, t_i^{11}, t_i^{21})$ and image feature $V(v_i^7, v_i^{11}, v_i^{21})$ with context information of different scales. The three text features and image features are merged respectively, and the attention map can be expressed as

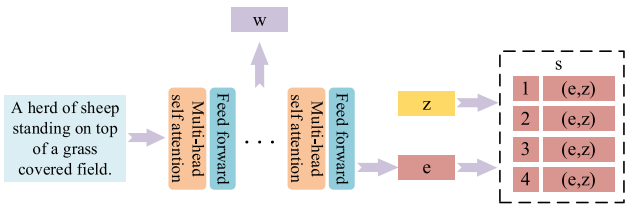

Fig. 3 The text feature extraction module

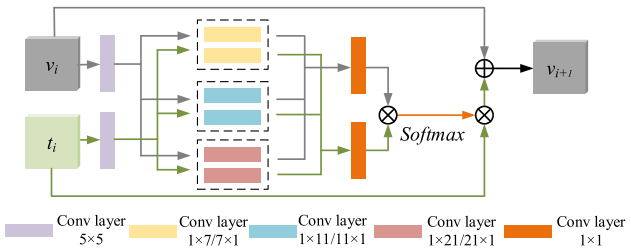

Fig. 4 The multi-group attention module

$$M = Softmax(f(t_i^{7:11:21})^T \otimes f(v_i^{7:11:21})), \quad (1)$$

where $f(*)$ represents the convolution layer with convolution kernel size of 1×1, $t_i^{7:11:21}$ and $v_i^{7:11:21}$ represent the text merged feature and the image merged feature respectively. $\otimes$ represents the matrix multiplication, and $Softmax(*)$ represents the softmax function. The resulting image feature can be expressed as

$$v_{i+1} = (M \otimes t) \oplus v_i, \quad (2)$$

where $\oplus$ the represents matrix addition. v_{i+1} is the image feature that passes through the i-th multi-group attention block. The i+1-th cycle block input is the output from the i-th cycle block and the local text feature of the corresponding dimension. By deep feature fusion through multiple cycle blocks, the semantic correlation between the generated image feature v and the text feature e is improved. Before and after each pair of multi-group attention block and upsampling block, the skip connection is used to solve the gradient vanishing problem.

3.2 Style-constrained image generation

In this part, the semantic feature is stylized to achieve image generation with style image constraint. Firstly, the style feature from the style image and the semantic feature from the semantic generation part are extracted through the pre-trained network. Then the extracted feature is input into the adaptive multi-scale attention normalization for feature style fusion. Finally, the convolution layers are used to generate the image with style image constraint.

To better extract feature for the style-constrained image generation part, the pre-trained VGG-19 network [25] is utilized to extract multi-scale feature. The semantic feature from the semantic image generation part and the style feature with dimension adjustment are input into VGG for feature extraction. To fully utilize of the multi-scale information of the deep network, the output features $(c_{i\_1}, s_{j\_1}|i,j \in 3,4,5)$ of layers $ReLU-3\_1$, $ReLU-4\_1$, and $ReLU-5\_1$ of VGG network are used for feature fusion. s_{i_1} represents the style feature from $ReLU-i\_1$ layer, and c_{i_1} represents the semantic feature from $ReLU-i\_1$ layer.

As shown in Fig. 5, the adaptive multi-scale attention normalization (AdaMSAttN) is used to combine the semantic features $(c_{i\_1}|i \in 3,4,5)$ and the style features $(s_{i\_1}|i \in 3,4,5)$. The features extracted $(c_{i\_1}, s_{j\_1}|i,j \in 3,4,5)$ from VGG network are computed with multi-group attention mechanism. The multi-scale style features $(ms_{i\_1}|i \in 3,4,5)$ and semantic features $(mc_{i\_1}|i \in 3,4,5)$ are obtained by convolution of features at different scales using the convolution group specified in the multi-group attention module. The multi-group calculation map Map_i can be expressed as

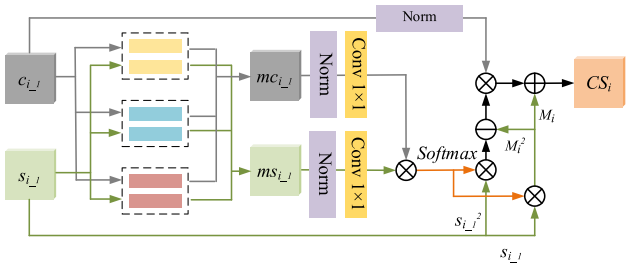

Fig. 5 The adaptive multi-scale attention normalization module

$$Map_i = Softmax(f(Norm(ms_{i\_1}))^T \otimes f(Norm(mc_{i\_1}))), \quad (3)$$

where $Norm(*)$ represents the mean–variance normalization in terms of channels, $f(*)$ represents convolution layer with convolution kernel size of 1×1, and ms_{i_1} and mc_{i_1} represent the multi-scale style feature and the image feature, respectively. The maximum value of i is 5, corresponding to the three groups of features from the VGG network. $\otimes$ represents the matrix multiplication. According to adaptive attention normalization [21], the generated feature is processed to obtain the weighted mean of attention M and the weighted standard deviation of attention S, which are, respectively, expressed as follows:

$$M_i = s_{i\_1} \otimes Map_i^T, \quad (4)$$

$$S_i = sqrt\left(\left(s_{i_1} \cdot s_{i_1}\right) \otimes Map_i^T - M_i \cdot M_i\right), \quad (5)$$

where $\cdot$ represents the element multiplication, and $sqrt(*)$ represents the square root calculation. Since the variance of the variable is equal to that the expectation of its square minus the square of its expectation, the above S is obtained. The fusion feature CS_i is obtained by adaptive instance normalization of the multi-scale semantic feature, and the formula is as follows:

$$CS_i = S_i \cdot Norm(c_i) + M, \quad (6)$$

where the fusion feature CS_i is upsampled to obtain the image with style image constraint. In the image style transfer domain, the multi-scale features of different depth networks are extracted for adaptive instance normalization. In contrast to this "deep" multi-scale exploration, the adaptive multi-scale attention normalization module additionally performs "breadth" multi-scale feature extraction on specific feature. Based on the extraction of features from different layers of the VGG network, the adaptive multi-scale attention normalization uses the multi-group attention mechanism to calculate the multi-scale target feature distribution and increases the richness and diversity of the extracted feature information at this layer.

3.3 Loss function

The whole network architecture contains two pairs of generators and discriminators. The total generator loss and total discriminator loss consist of two generator losses and two discriminator losses, respectively.

Generator G_1 loss: The generator loss L_{G_1} includes the unconditional loss to verify the authenticity of semantic feature, the conditional loss to verify the semantic consistency of text feature and semantic feature, and text–image similarity attention loss. The formula for the loss of L_{G_1} is

$$L_{G_1} = -\frac{1}{2}E_{x_s\sim p_{G_1}}[\log(D_1(x_s))] - \frac{1}{2}E_{x_s\sim p_{G_1}}\left[\log\left(D_1\left(x_s, e\right)\right)\right] - L_{SA}, \tag{7}$$

where x_s is from the output of the semantic image generation part and e represents the text embedding.

To further the semantic consistency of text and semantic image, inspired by DAMSM t3], the text–image similarity attention loss L_{SA} based on the transformer encoder is built. The matching text–image pair is input into the text encoder and the image encoder, respectively. The text feature t and the image feature v are, respectively, obtained from the middle layer of the text encoder and image encoder [23]. The two features are transformed into two groups of features $T(t_7 : t_{11} : t_{21}) \in R^{3N\times N}$ and $V(v_7 : v_{11} : v_{21}) \in R^{3N\times N}$ with different scale-context information, and then the attention map of T and V is calculated. The weighted vector I_n in the n-th part can be expressed as the following formula:

$$\begin{aligned} I_n &= \sum_{m=1}^{3N} A_m V_m, \qquad A_m = \frac{\exp\left(a_{n,m}\right)}{\sum_{k=1}^{3N}\exp\left(a_{n,k}\right)}, \\ a_{n,m} &= \frac{\exp\left(V_n^T \otimes T_m\right)}{\sum_{k=1}^{N}\exp\left(V_k^T \otimes T_m\right)}, \end{aligned} \tag{8}$$

where A_m is the attention weight of the m-th part of the text, and $a_{n,m}$ is the dot product similarity between the m-th part of text feature and the n-th part of image feature. T_m and V_n represent the m-th part of the text feature and the n-th part of the image feature, respectively. Text–image similarity attention loss L_{SA} can be expressed as

$$L_{SA} = L_1^w(I, T) + L_2^w(I, T), \tag{9}$$

where $L_1^w(*)$ is the cross entropy loss for calculating the similarity between feature I and feature T. The similarity between I and T is calculated by using the cosine similarity between their subparts, and $L_2^w(*)$ is handled similarly.

Discriminator D_1 loss: The discriminator D_1 is set to monitor the text–image semantic alignment of the semantic feature and the similarity between the semantic feature and the real image feature. The discriminator D_1 loss is as follows:

$$\begin{aligned} L_{D_1} = &-\frac{1}{2}E_{x_1\sim p_{real}}\left[\log\left(D_1\left(x_1\right)\right)\right] - \frac{1}{2}E_{x_1\sim p_{real}}\left[\log\left(D_1\left(x_1, e\right)\right)\right] \\ &- \frac{1}{2}E_{x_s\sim p_{G_1}}\left[\log\left(1 - D_1\left(x_s\right)\right)\right] - \frac{1}{2}E_{x_s\sim p_{G_1}}\left[\log\left(1 - D_1\left(x_s, e\right)\right)\right], \end{aligned} \tag{10}$$

where x_1 is from the real image distribution p_{real} of p_{G_1} corresponding scale.

Generator G_2 loss: The style loss and the content loss are added to the generator loss. The generator formula is as follows:

$$L_{G_2} = -E_{x_{cs}\sim p_{G_2}}\left[\log\left(D_2\left(x_{cs}\right)\right)\right] + \gamma L_{style} + \alpha L_{content}, \tag{11}$$

where x_{cs} is from the output of the style-constrained image generation part, and γ and α are the hyperparameters, $L_{content}$ is the content loss, and L_{style} is the style loss. The two losses can be expressed as follows:

$$\mathrm{L_{content}} = \sum_{\mathrm{x=2}}^{5} \|\mathrm{Norm(SC_i)} - \mathrm{Norm(C_i)}\|_2, \tag{12}$$

$$L_{style} = L_g + L_l, \tag{13}$$

where SC_i represents the fusion feature from the $ReLU - i\_1$ layer in the VGG network, C_i represents the generated feature from the $ReLU - i\_1$ layer in the VGG network, L_g is the global style loss, and L_l is the local style loss. The formulas of the two style losses are as follows:

$$L_g = \sum_{x=2}^{5}(\|\mu\left(SC_i\right) - \mu\left(S_i\right)\|_2 + \|\sigma\left(SC_i\right) - \sigma\left(S_i\right)\|_2, \tag{14}$$

$$L_l = \sum_{x=3}^{5}(\|SC_i - AdaMSAttN\left(S_i, C_i\right)\|_2, \tag{15}$$

where $\mu(*)$ represents the mean, $\sigma(*)$ represents standard deviation, and $AdaMSAttN(*)$ represents adaptive multi-scale attention normalization.

Discriminator D_2 loss: The discriminator D_2 is set to preserve the semantics integrity of the style-constrained feature. The discriminator loss is as follows:

$$L_{D_2} = -E_{x_2\sim p_{real}}[\log(D_2(x_2))] - E_{x_{cs}\sim p_{G_2}}[\log\left(1 - D_2\left(x_{cs}\right)\right)], \tag{16}$$

where x_2 is from the real image distribution p_{real} of x_{cs} corresponding scale.

3.4 Algorithm

Algorithm 1

Input: Text Feature Extraction Module F_{TFE}; Multi-group Attention Module F_{MGA}; Adaptive Multi-scale Attention Normalization Module F_{AMAN}; Pretrained VGG Network F_{VGG}; Text-real-style pairs $\{X^{txt}, X^{real}, X^{sty}\}$; Learning rate N; Batch size M.

Output: $\hat{X}_{sem}$ and $\hat{X}_{sty}$.

1 **repeat**

2 Sample text-real-style pair $\{X^{txt}, X^{real}, X^{sty}\}$ and generated random noise vector z;

3 Generate text semantic embedding by f_0, $w \leftarrow F_{TFE}(X^{txt}, z)$;

4 **for** $i \in [0,1...N-1]$ **do**

Generate semantic feature by $f_{i+1} \leftarrow F^i_{MGA}(f_i, w)$;

end for

5 Generate semantic image by $(\hat{X}_{sem}) \leftarrow G_1(f_N)$;

6 Generate extracting feature $\left(f^c_1, f^c_2, f^c_3, f^s_1, f^s_2, f^s_3\right) \leftarrow F_{VGG}\left(f_N, X^{sty}\right)$;

7 Generate style feature $f_{cs} = F_{AMAN}\left(f^c_1, f^c_2, f^c_3, f^s_1, f^s_2, f^s_3\right)$;

8 Generate style image by $(\hat{X}_{sty}) \leftarrow G_2(f_{cs})$;

9 Calculate the discriminative loss L_{D_1}, L_{D_2} by (10) and (16) respectively;

10 Update parameters of discriminator D_1, D_2 by AdamW optimizer;

11 Calculate the generator loss L_{G_1}, L_{G_2} by (7) and (11) respectively;

12 Update the parameters of the generators (G_1, G_2) by AdamW optimizer;

13 **until** Reach maximum epochs number;

14 **return** $\hat{X}_{sem}, \hat{X}_{sty}$.

In Algorithm 1, some details of the proposed SC-GAN are shown. The main modules involved in this network include the text feature extraction module F_{TFE}, the multi-group attention module F_{MGA}, and the adaptive multi-scale attention normalization module F_{AMAN}. Text–real–style pairs $\{X^{txt}, X^{real}, X^{sty}\}$ are formed by matching the corresponding text, real image, and style image in the dataset. The outputs of the network are semantic image $\hat{X}_{sem}$ and style-constrained image $\hat{X}_{sty}$.

In the process of semantic feature generation, there are N cyclic blocks containing the multi-group attention module, which corresponds to the N in the fourth line of the algorithm. Here N is set to 3. The generator and the discriminator are trained alternatively. The parameters of the first pair of generator and discriminator are updated 5 times faster than those of the second pair.

4 Experiments

To evaluate the effectiveness of the proposed SC-GAN, extensive experiments are performed on the MS COCO [26] and the WikiArt [27] datasets. The performance is compared with current mainstream generation methods and style transfer methods.

4.1 Datasets

The MS COCO and the WikiArt datasets are used for image generation. The MS COCO dataset as the real image dataset provides the basic semantic feature of the generated image. The WikiArt dataset as the style image dataset provides the style feature of the generated image. The MS COCO dataset and the WikiArt dataset are used for the training and testing of the generated model. The MS COCO dataset contains 123,287 images, including 82,783 training images and 40,504 test images. Each image in the MS COCO dataset has 5 corresponding textual descriptions. A total of 42,129 training images and 10,628 test images are used from the WikiArt dataset.

The whole network is trained in stages under the constraints of hardware to reduce the training time. The semantic image generation part is trained separately, and then the trained model is used as a pre-trained model for the style-constrained image generation part. To improve the generalization ability of the model, the real image dataset is divided into two parts, more than 60,000 images are used for the pre-training stage and another 20,000 images are used for the style fusion training stage.

4.2 Implementation details

Since there is a many-to-one correspondence between text and image in the MS COCO dataset, in order to form a one-to-one text–image pair, the corresponding multiple texts of an image are randomly selected to form a text–image pair with this image. The AdamW optimizer [28] is used for training, where the learning rate is set to 0.0001. The style-constrained image output size is 256×256. The model is trained for 120 epochs on common dataset. The hyperparameters for the conditional loss and unconditional loss of the discriminator D_1 are set to 2 and 1, respectively. The unconditional loss of the discriminator D_2 is set to 0.5. The hyperparameters for style loss and content loss are set to 1 and 1.5, respectively.

Table 1 Performance of different methods on COCO dataset

Methods	IS ↑	FID ↓	R-Precision ↑
Baseline	24.24 ± 0.39	37.27	80.33 ± 0.26
Baseline + MGA	32.44 ± 0.36	27.33	90.14 ± 0.22

4.3 Ablation study

To further investigate the influence of various components proposed by the network, the proposed network SC-GAN performs ablation study on the MS COCO dataset. Since the proposed network utilizes style constraints for image generation, the influence of the multi-group attention module on the generation quality is mainly studied. The baseline is set to the semantic image generation part, which lacks the multi-group attention module. The results of the proposed method in Inception Score (IS) [29], Fréchet Inception Distance (FID) [30], and R-Precision [3] are shown in Table 1. The higher the value of IS, the better the sharpness and diversity of the image. The lower the value of FID, the closer the generated image is to the real image. R-Precision is used to evaluate the visual–semantic similarity between the generated image and the corresponding text description, and its value is proportional to the visual–semantic similarity. As can be seen, after adding the multi-group attention module, the proposed network gains increase of 8.2 on IS, decrease of 9.94 on FID, and increase of 9.81 in R-Precision. This shows that the network with multi-group attention module can capture different scales of information and rich visual details, so the generated image has better diversity and authenticity. The experiment proved that the multi-group

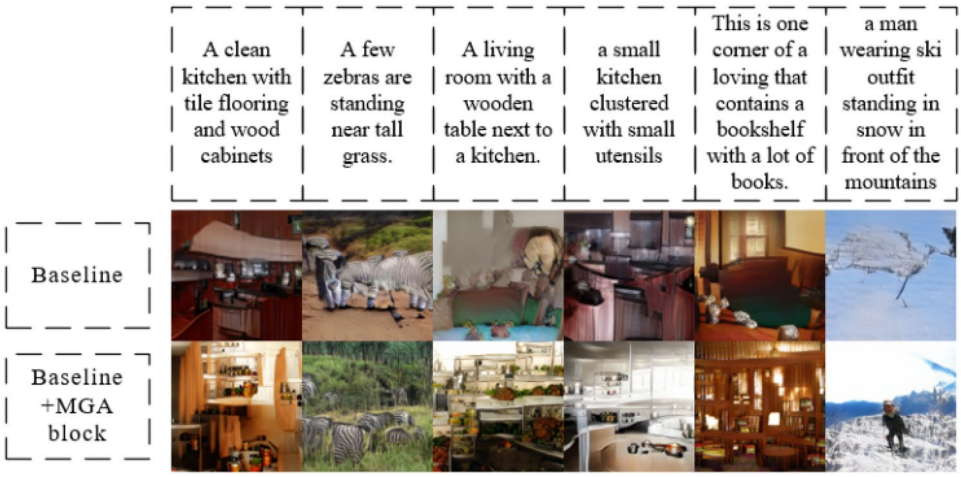

Fig. 6 Qualitative comparison between different modules of the proposed method for ablation

Fig. 7 Visualization results of the multi-group attention module in the semantic generation section

attention module has greatly improved the baseline method performance.

As for the visual effect generated with or without the multi-group attention module, it can be seen in Fig. 6 that the visual effect generated by the method with the multi-group attention module is obviously better than the baseline method. Compared with the images in the second line, the details generated by the objects in the third line are clearer, and the content of the image is more consistent with the text description. This shows that the multi-group attention module helps generate important details during the generation process. At the same time, the layout of the whole image tends to be more realistic. This proves that the multi-group attention module effectively captures the dependencies of long–short-range information.

To visually demonstrate the function of the proposed multi-group attention module, the semantic image is visualized with attention. Four groups of examples are shown in Fig. 7, each group containing a text–image pair and a set of attention visualizations that generated image. The words that appear in the attention visualization in each text description are shown in red. In the attention visualization, the words that are most relevant to the allocation of attention are highlighted. In the upper left group example, object names like "woman" and "water" receive valid attention. This shows that the multi-group attention module can help the network define important objects and background information in the image. In the lower left example, the highlighted section corresponding to "street" is not purely continuous and does not cover the red object. This shows that the proposed module is beneficial for the association of non-neighborhood information.

The adaptive multi-scale attention normalization method is mainly used for the style fusion of the generated image, so the generation comparison between methods is not performed in the ablation experiment.

4.4 Comparison experiments

Since there are almost no references on relevant issues, this paper compares the generation method and style transfer method with other mainstream methods, respectively, so as to prove the effectiveness of the proposed method.

4.4.1 Generation methods comparison

In the generation task, the visual effect of the image is an important reference for the generated image. To demonstrate the effect of generation, the generated architecture without style image constraint (SC-GAN-w) is used for comparison with the current advanced generation network [3, 6, 9, 14, 31]. SC-GAN-w includes all of the semantic image generation parts of SC-GAN. As shown in Fig. 8, the image details of SSA-GAN [9] and DF-GAN [14] are obviously better than those of AttnGAN [3] and DM-GAN [31], and the generation effects of the proposed method are close to those of SSA-GAN [9] and DF-GAN [14]. In the sixth line, it can be seen that there is no identifiable target corresponding to the text "a man" in the other methods, while the easily identifiable target can be found in the first column. This proves that the proposed method achieves better results in text–image semantic alignment, and the text–image semantic consistency of the generated image is higher than other methods. In addition, it can be seen that the proposed method generates recognizable "character" in the third line images. This shows that the proposed method obtains more visual details through multi-scale attention, which has certain advantages in image detail generation compared to other methods.

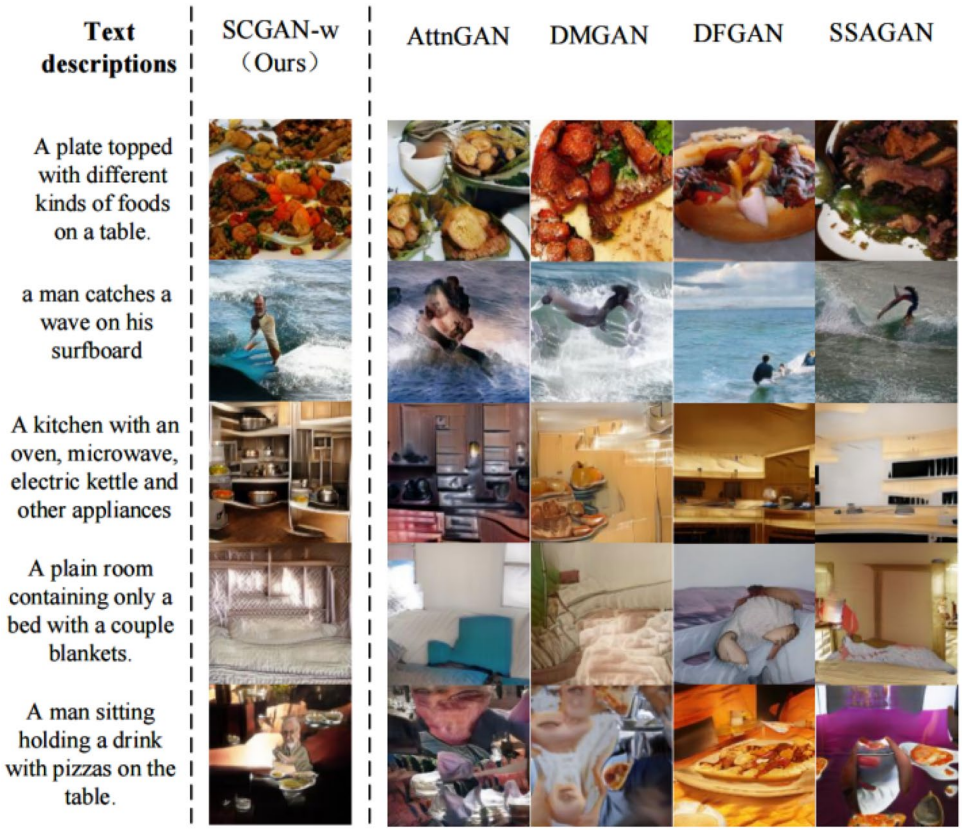

Fig. 8 Comparison of semantic images using different methods on COCO dataset

Table 2 Performance of different methods on COCO dataset

Methods	IS ↑	FID ↓	R-Precision ↑
StackGAN++[2]	8.30±0.03	51.62	72.83±3.17
AttnGAN [3]	25.89±0.47	35.49	85.47±3.69
ControlGAN [32]	24.06±0.47	–	82.43±2.43
MirrorGAN [4]	26.47±0.41	–	74.52
DM-GAN [31]	30.49±0.57	32.64	88.56±0.28
DAE-GAN [6]	35.08±1.16	28.12	92.61±0.50
SC-GAN-w	32.44±0.36	27.33	90.14±0.22

As shown in Table 2, SC-GAN-w is used for comparison with the current mainstream generation method in quantitative research. Compared with the current mainstream methods including StackGAN++[2], AttnGAN [3], ControlGAN [32], MirrorGAN [4], DM-GAN [31], and DAE-GAN [6], the proposed method is still competitive in the evaluation methods of image generation quality Inception Score (IS) [29], Fréchet Inception Distance (FID) [30], and R-Precision. The proposed method values are both better than StackGAN++[2], AttnGAN [3], and DM-GAN [31] in IS, FID, and R-Precision. The proposed method SC-GAN-w method is better than ControlGAN [32] and MirrorGAN [4] in data performance. Although SC-GAN-w is inferior to DAE-GAN [6] in IS and R-Precision, it still achieves good

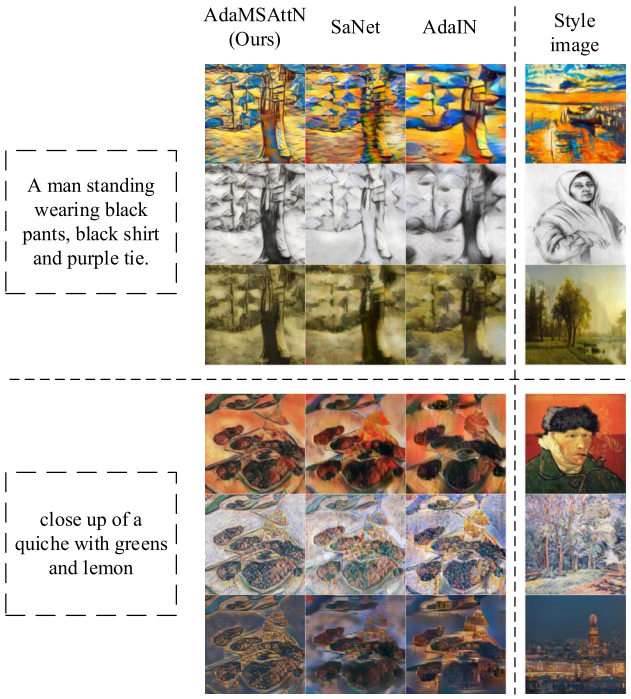

Fig. 9 Comparison of style transfer components

Table 3 Content difference and style difference of different methods

	AdaMSAttN	SANet	AdaIN
L_c	1.96	2.32	2.21
L_s	2.41	2.45	3.17

Fig. 10 Generated images with a specified style

values in FID. This shows that the proposed method can generate image with higher semantic consistency.

4.4.2 Style transfer methods comparison

To verify the validity of the proposed component AdaMSAttN, the proposed style transfer module AdaMSAttN is visually contrasted with the modern mainstream style transfer network [19, 20]. The specific approach is using the SC-GAN architecture, respectively, equipped with AdaMSAttN, AdaIN, and SANet key style transfer components to complete the comparison. The comparison of the style-constrained image generated by text and style image is shown in Fig. 9. In the figure, the first column text is the text description, the last column images are the given style image, and the middle images are the comparison of three different style transfer components. In the image corresponding to the first text, the images with the style transfer method AdaMSAttN still maintain good texture, and the object in the image is clearer, while the target identification in the image generated by AdaIN is low. This shows that the proposed method makes the image have high semantic retention through multi-scale attention. It can be seen that the proposed method AdaMSAttN in this paper can transfer the color and other elements in the style image to the generated image in a more detailed way, so the proposed method AdaMSAttN has better recognition ability and semantic retention ability than other methods.

The content difference L_c in this paper is the difference between the generated result with style image constraint and the semantic image. The smaller the content difference, the higher the semantic information retention degree of the semantic features. The style difference L_s represents the difference between the generated result with style image constraint and the style image. The smaller the difference, the stronger the style fusion ability of the style transfer component. The content difference and style difference mentioned above are two indirect indicators of style fusion by style transfer components. For AdaMSAttN, AdaIN, and SANet methods, the content difference and style difference are calculated according to (12) and (14), respectively. As shown in Table 3, the content difference and style difference of the proposed method AdaMSAttN are lower than other methods, which indicates that the proposed method can effectively retain the input semantic information and fully integrate the given style.

4.4.3 Generation images with constraints

In the generation image guided by text, style is first used as a constraint in the image generation of the proposed network SC-GAN. In Fig. 10, representative style images generated by the proposed method are shown, where the first column is the text and the other column is the generated image with style image constraint. The generated images in column 2 and column 3 inherit the main hue of "red" and "blue" of the style image, respectively, and the generated images in line 3 and line 7 obtain a small amount of texture features of the style image. This shows that the proposed method can perform effective style-specific fusion. As you can see, the content of the generated image in the middle corresponds to the text semantics, and the color and texture details of the style image have basically not changed the image semantics and are integrated into the generated image. This proves that the proposed method can generate better image with both text and style constraints.

5 Conclusion

The generative adversarial text-to-image generation with style image constraint is proposed innovatively, which makes the generated image have a specific style through the process of style fusion. The multi-group attention module is used for the semantic alignment of multi-scale information and richer diversity of images. At the same time, multi-group attention module is applied in the adaptive instance normalization to

realize the fusion of style features. The proposed method generates image with good visual effects under the double constraints of text and style image and verifies the feasibility of the proposed method in the direction of text-to-image generation with style image constraint. In the future, new ideas are tried to simplify the model size and multi-style constraints of the text-to-image generation direction.

Acknowledgements This work is partially supported by the National Natural Science Foundation of China (Nos. 62076153, 62176144), the major fundamental research project of Shandong, China (No. ZR2019ZD03), and the Taishan Scholar Project of Shandong, China (No. ts20190924).

Author contributions ZW wrote the main manuscript text and prepared figures. All authors reviewed the manuscript.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
2. 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. Pattern Anal. Mach. Intell. **41**(8), 1947–1962 (2018)
3. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: 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)
4. Qiao, T., Zhang, J., Xu, D., Tao, D.: Mirrorgan: learning text-to-image generation by redescription. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1505–1514 (2019)
5. Qiao, T., Zhang, J., Xu, D., Tao, D.: Learn, imagine and create: text-to-image generation from prior knowledge. In: Proceedings of Advances in Neural Information Processing Systems, pp. 885–895 (2019)
6. Ruan, S., Zhang, Y., Zhang, K., Fan, Y., Tang, F., Liu, Q., Chen, E.: Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 13960–13969 (2021)
7. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever I.: Zero-shot text-to-image generation. In: International Conference on Learning Representations, pp. 8821–8831 (2021)
8. Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., Yang, Y.: Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 833–842 (2021)
9. Liao, W., Hu, K., Yang, M. Y., Rosenhahn, B.: Text to image generation with semantic-spatial aware GAN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 18187–18196 (2022)
10. Park, H., Yoo, Y., Kwak, N.: Mc-gan: multi-conditional generative adversarial network for image synthesis. arXiv preprint arXiv: 1805.01123 (2018)
11. Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., Gao, J.: StoryGAN: A Sequential Conditional GAN for Story Visualization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6329–6338 (2019)
12. Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2327–2336 (2019)
13. Cheng, J., Wu, F., Tian, Y., Wang, L., Tao, D.: Rifegan: rich feature generation for text-to-image synthesis from prior knowledge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10911–10920 (2020)
14. Tao, M., Tang, H., Wu, S., Sebe, N., Jing, X., Wu, F., Bao, B.: Df-gan: deep fusion generative adversarial networks for text-to-image synthesis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 16515–16525 (2022)
15. Efros A. A., Freeman W. T.: Image quilting for texture synthesis and transfer. In: Annual Conference on Computer Graphics and Interactive Technique, pp. 341–346 (2001)
16. Kyprianidis, J.E., Collomosse, J., Wang, T., Isenberg, T.: State of the "art": a taxonomy of artistic stylization techniques for images and video. IEEE Trans. Vis. Comput. Graphics **29**(5), 866–885 (2013)
17. Gatys, L., Ecker, A., Bethge M.: A neural algorithm of artistic style. arXiv preprint arXiv: 1508.06576 (2015)
18. Johnson, J., Alahi, A., Li, F.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp. 694–711 (2016)
19. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
20. Park, D. Y., Lee, K. H.: Arbitrary style transfer with style-attentional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5880–5888 (2019)
21. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6649–6658 (2021)
22. Park, J., Kim, Y.: Styleformer: transformer based generative adversarial networks with style vector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8983–8992 (2022)
23. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021)
24. Liu, Z., Cai, W., Xu, Z.J.: Multi-scale deep neural network (MscaleDNN) for solving Poisson-Boltzmann equation in complex domains. Commun. Comput. Phys. **28**(5), 1970–2001 (2020)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Lin, T, Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L.: Microsoft coco: common objects in context. In: European conference on computer vision, pp. 740–755 (2014)
27. Phillips, F., Mackintosh, B.: Wiki art gallery, inc.: a case for critical thinking. Issues Account. Educ. **26**(3), 593–608 (2011)
28. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. arXiv preprint arXiv:1711.05101, (2017)

29. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Proceedings of Advances in Neural Information Processing Systems, pp. 2234–2242 (2016)
30. 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: Proceedings of Advances in Neural Information Processing Systems, pp. 6626–6637 (2017)
31. 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 Conference on Computer Vision and Pattern Recognition, pp. 5802–5810 (2019)
32. Li, B., Qi, X., Lukasiewicz, T., Torr, P.: Controllable text-to-image generation. In: Proceedings of Advances in Neural Information Processing Systems, pp. 2063–2073 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.