Neural Processing Letters (2024) 56:11
https://doi.org/10.1007/s11063-024-11503-5

# RII-GAN: Multi-scaled Aligning-Based Reversed Image Interaction Network for Text-to-Image Synthesis

**Haofei Yuan[1] · Hongqing Zhu[1] · Suyi Yang[2] · Ziying Wang[1] · Nan Wang[1]**

Accepted: 25 November 2023 / Published online: 6 February 2024


**Abstract**
The text-to-image (T2I) model based on a single-stage generative adversarial network (GAN) has significantly succeeded in recent years. However, the generation model based on GAN has two disadvantages: the generator does not introduce any image feature manifold structure, which makes it challenging to align the image and text features. Another is the image's diversity; the text's abstraction will prevent the model from learning the actual image distribution. This paper proposes a reversed image interaction generative adversarial network (RII-GAN), which consists of four components: text encoder, reversed image interaction network (RIIN), adaptive affine-based generator, and dual-channel feature alignment discriminator (DFAD). RIIN indirectly introduces the actual image distribution into the generation network, thus overcoming the problem that the network lacks the learning of the actual image feature manifold structure and generating the distribution of text-matching images. Each adaptive affine block (AAB) in the proposed affine-based generator can adaptively enhance text information, establishing an updated relation between original independent fusion blocks and the image feature. Moreover, this study designs a DFAD to capture important feature information of images and text in two channels. Such a dual-channel backbone improves semantic consistency by utilizing a particular synchronized bi-modal information extraction structure. We have performed experiments on publicly available datasets to prove the effectiveness of our model.

✉ Hongqing Zhu
hqzhu@ecust.edu.cn

Haofei Yuan
y30210919@mail.ecust.edu.cn

Suyi Yang
suyi.yang@kcl.ac.uk

Ziying Wang
y10210090@mail.ecust.edu.cn

Nan Wang
wangnan@ecust.edu.cn

[1] School of Information Science and Engineering, East China University of Science and Technology, No.130 Meilong Road, Shanghai 200237, China

[2] Department of Mathematics, Natural, Mathematical and Engineering Sciences, King's College London, Strand, London WC2R 2LS, England, UK

**Keywords** Text-to-image · Single-stage generation · Reversed image interaction network · Adaptive affine-based generator · Dual-channel

## 1 Introduction

Text-to-image (T2I) synthesis is an essential task involving natural language processing and computer vision, which aims to generate semantic consistent images from given text descriptions. Most existing T2I frameworks utilize generative adversarial network (GAN) [5] to synthesize images. Although it has been widely recognized for generating visually realistic images, the T2I model based on GAN still faces some challenges.

Firstly, the training instability of GAN leads to the mode collapse and the lack of diversity of generated images. To address the issues above, recent approaches have developed stacked GANs to generate multi-scale images in a low-to-high resolution manner to supervise image synthesis, such as AttnGAN [32], DM-GAN [41], and SD-GAN [35], etc. These methods have been proven effective in synthesizing high-resolution images and preventing model collapse. However, based on a multi-stage structure, each generator refines the previous stage's results by improving the image resolution and enriching more details. Hence, the final output depends heavily on the image generated in the initial step. To this end, Tao et al. [29] explored a simplified single-stage end-to-end generative model to optimize the training trajectory of GANs.

However, efficiently aligning features between image and text presents a significant challenge. Some methods, such as MirrorGAN [19], SuperGAN [3], etc., utilize the extra inverse inference model to caption generated images. This architecture enhances the generator's capacity to generate semantically consistent images by effectively capturing visual-textual consistency information throughout training. Although this strategy is straightforward, it has certain drawbacks. Notably, it necessitates the pre-training of the image-to-text (I2T) model due to its significant size and complexity, making it challenging to train compared to some end-to-end models.

Another challenge is effectively implementing cross-model transformation in the case of a large semantic gap between text and image domains. Recently, some researchers have considered two strategies, cross-model attention [32] and affine transformation [29], to alleviate this problem. For the former, cross-modal attention calculation usually occurs between the word embedding and the current layer feature map. Therefore, expensive computing costs hinder the model from generating higher-resolution images and performing multi-scale depth fusion. For the latter, the affine transformation is a conditional batch normalization (CBN) [30] layer, which was first introduced in the T2I task of SD-GAN [35]. However, it did not perform multi-scale deep fusion due to the limitations of the stacked network architecture. Based on their work, Tao et al. [29] used affine transformation from sentence embedding to image to achieve multi-scale depth fusion, in which each fusion module is merged independently at different scales. As we know, too many independent modules will inevitably lead to training conflicts. In addition, the fusion module may cause the problem of overlapping or missing input text information because it does not accept the forward calculated features from the generator and lacks self-adaptability when selecting input text information.

The last challenge is that if the T2I model lacks an understanding of the image feature manifold structure, and only relies on the game between the generator and the discriminator to learn the real image distribution, the T2I model may converge slowly.

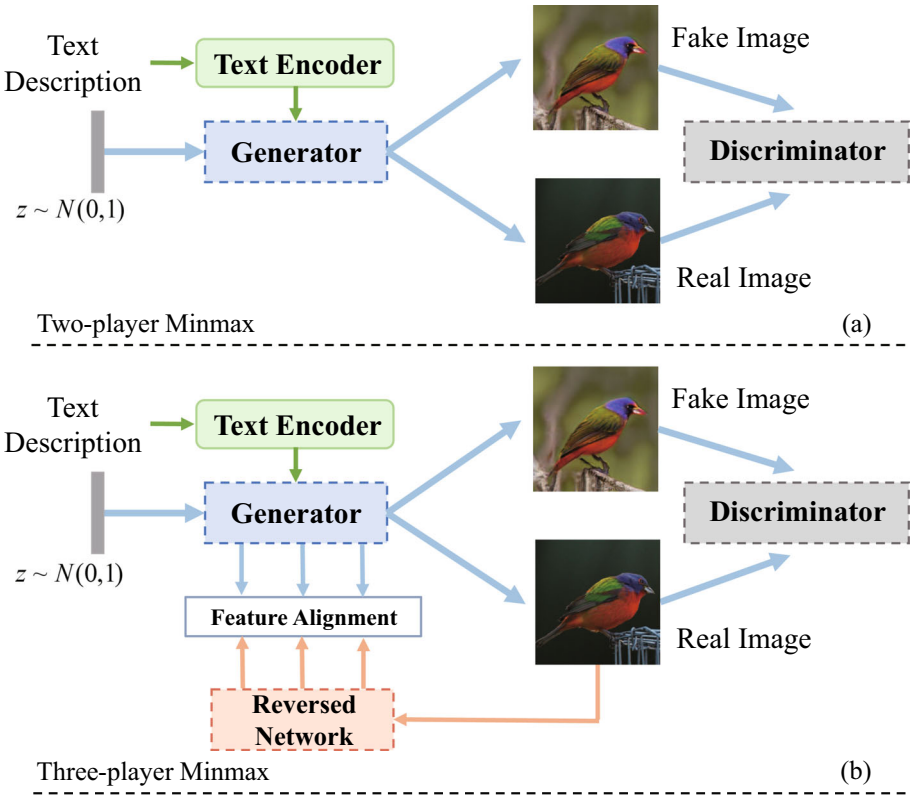


**Fig. 1** Comparison between a convention GAN-based T2I (up) and the proposed RII-GAN (bottom)

This paper proposes a novel end-to-end GAN-based T2I model called reversed image interaction GAN (RII-GAN), which has a single-stage generation structure. RII-GAN indirectly introduces the authentic image feature manifold structure into the generator using feature alignment techniques in the proposed reversed image interaction network (RIIN). Specifically, the architecture of this RIIN is equivalent to the reversed feature transformation of the generator, and the feature alignment constraint conducts in the multi-scale. Figure 1 shows the framework of the conventional GAN-based T2I network and the proposed RII-GAN with the reversed network. RIIN is a symmetrical counterpart to the generator's low-to-high resolution hierarchical structure, which is both lightweight and adaptable. Discarding the consuming strategy of feature alignment within the individual text or image domains, it operates within the intermediate domains between text and image throughout the generation process. This unique reversed interaction mechanism reduces the reliance of the generator on text descriptions and provides it with a supplement pathway to access features of images during training. Meanwhile, the mutual optimization of the generator and RIIN, which begins with a weaker capacity and progressively strengthens alongside the improvement of the generator and discriminator, adapts to the intermediate mixed domain and dynamically evolves throughout the adversarial training process. Moreover, we propose a novel adaptive affine-based generator composed of various adaptive affine blocks (AAB) at different resolutions as text-image fusion blocks. AAB will enhance text information before the affine transformation. It includes an adaptive block to adapt the input text embedding by driving the

information exchange between the text feature information and the feature maps from the current generator forward computation process to avoid overlapping and missing text information. AAB further alleviates the problem of the semantic gap between text and image domains by rebuilding the text embeddings. Finally, we correspondingly enhance the discriminator structure to facilitate adversarial training for the generator that has strengthened generation capability with the introduced reversed interactive scheme. Therefore, this study proposes the dual-channel feature alignment discriminator (DFAD). It achieves simultaneous refinement of text and image features through bi-modal feature extractions in two channels. It guides the discriminator to extract semantically consistent bi-modal features through feature alignment operations. Distinct from mainstream conditional discriminators that concatenate features, our proposed DFAD significantly enhances the processing capability of bimodal features. Consequently, it better aligns images and text features, substantially improving semantic consistency.
The main contributions of this work are summarized as follows.

- We propose a reversed image interaction network to help the generator learn the authentic image feature manifold structure through multi-scaled feature alignment constraints. This alleviates the reliance of the generator on input text description and provides an alternative generation scheme for GAN-based multimodal generation tasks.
- We explore an adaptive affine-based (AAB) generator which designs an adaptive block and establishes an adaptive updating scheme for text-image fusion.
- We propose a dual-channel feature alignment discriminator, which allows simultaneous extraction of textual and visual features in each modality and implements an advanced feature alignment strategy to improve semantic consistency.

## 2 Related Works

### 2.1 GAN-Based Text-to-Image Synthesis

Reed et al. [22] first demonstrated that conditional GAN (cGAN) [15] can synthesize credible images from text descriptions. To generate the $256 \times 256$ resolution images, Zhang et al. [36] developed a stacked structure GAN (StackGAN) by several generator-discriminator pairs. Xu et al. [32] extended it and introduced cross-modal spatial attention to calculating the attention matrix for text-to-image fusion. Zhu et al. [41] explored a dynamic memory scheme for improving T2I fusion process. Li et al. [10] introduced channel attention to exploring the relationship between the image feature channel and word context vector. To improve the semantic consistency between the generated images and text, Yin et al. [35], and Tan et al. [27] utilized Siamese structures with contrastive learning loss to enhance semantic consistency. MirrorGAN [19] leverages an additional inverse inference model to caption the generated images, enhancing the generator's ability to create semantically consistent images by capturing visual-textual consistency. However, it requires pre-training of the I2T model. Due to its substantial size and complexity, the method is relatively challenging in training and potentially less efficient than end-to-end models. To improve the generation quality, Yang et al. [33] explored the semantic common between different sentences describing the same image by utilizing multiple single-sentence generation and multi-sentence discrimination (SGMD) modules. Considering the properties of image complexity and text generality, Tan et al. [28] designed a regularized GAN framework to distinguish the critical and unimportant

information of generated images, making the network more inclined to focus on the necessary semantic parts of the feature map.

Recently, some works have simplified the stack structure into a single generator-discriminator pair. For example, Zhang et al. [38] chose the one-stage generative structure as the backbone for XCM-GAN through five contrastive learning pairs between image, sentence, and word embeddings. Tao et al. [29] also chose a single-stage GAN-based T2I model to achieve text-image fusion. They employed sentence embedding as conditional input, making fusion more effective. For making the model pay more attention to semantic focus areas, inspired by [29], Liao et al. [12] introduced a weakly-supervised mask prediction for the fusion module. In [11], Li et al. provided the image feature manifold structure directly to the generator by a memory bank, which previously stores the authentic image feature maps. Unlike them, the proposed RII-GAN improves the quality of generated images by exploiting a combination of a simple single-stage generation structure and a new reversed image interaction mechanism.

### 2.2 Text-to-Image Fusion

The early T2I model achieved text-to-image fusion by directly concatenating text and image feature maps. Subsequently, some works such as AttnGAN [32] and StackGAN++ [37] used the cross-modal attention mechanism to improve the fusion effect. However, the introduction of an attention module inevitably increases computational consumption. To solve this problem, Yin et al. [35] introduced condition batch normalization (CBN) into text-image fusion. It fuses text and image features by disentangling text embedding and performing the affine transformation. Afterward, Tao et al. [29] proposed a deep fusion module, which achieves an excellent result by adopting sentence embedding to achieve multi-scale fusion. Liao et al. [12] used masks to pay more attention to the generation of crucial feature regions, thus improving the fusion effect. Ye et al. [34] found that mutually independent fusion modules conflict with each other during training. To address this issue, they use a bidirectional long short-term memory network (BiLSTM) [25] to establish a long-term dependency on the independent fusion blocks, which helps to reduce the difficulty of model training. Our framework introduces an adaptive affine transformation, which also provides a solution to achieve long-term dependency of fusion blocks while implementing textual information augmentation by adaptively considering the feature maps computed forward by the current generator.

### 2.3 Feature Alignment in Text and Image

Image-text and image-image feature alignment schemes are the research focus of T2I generation tasks. Xu et al. [32] utilized BiLSTM [25] and Inception-v3 model [26] to build the deep attentional multimodal similarity model (DAMSM), which is applied to vary text embedding and image features to the same dimension and perform feature alignment to improve the image-text consistency. To reduce the randomness of the generated images, Li et al. [10] developed perceptual loss to align the features of the generated and authentic images. Yin et al. [35] found that the T2I model based on GAN is challenging to distinguish the subtle differences between each representation of the same image. Therefore, they use Siamese structure to align different generated images with corresponding text descriptions but the same reference object. Zhang et al. [38] improved their work using five pairs of contrastive learning. Afterward, Tan et al. [28] argued that the crucial to generating high-quality images is whether the generator can identify critical features in the forward calculation process.

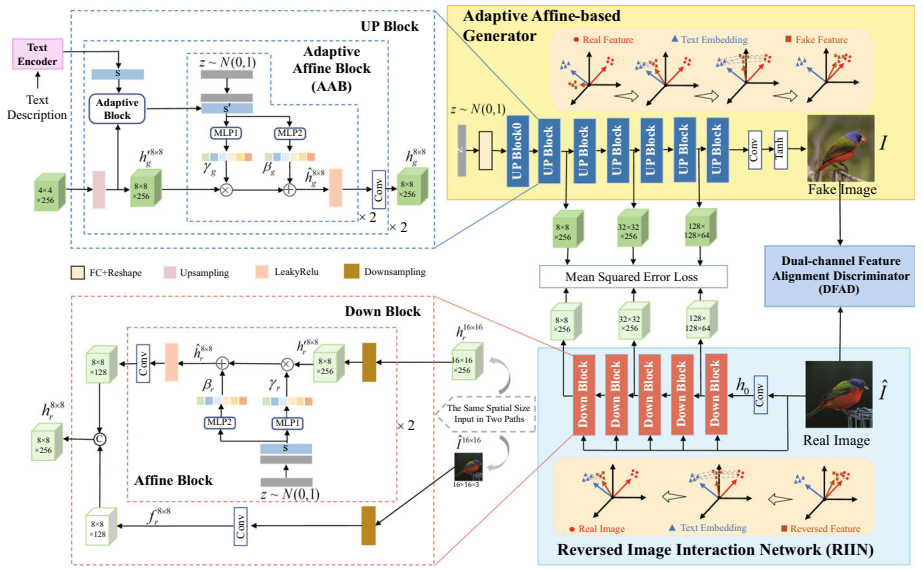


**Fig. 2** The framework of the proposed RII-GAN consists of four modules: text encoder, adaptive affine-based generator, reversed image interaction network and dual-channel feature alignment discriminator

Therefore, they used the semantic disentangling module (SDM) to disentangle the generated foreground image features to perform feature alignment with real image features. We propose a novel RIIN for multi-scale image feature alignment with the generator. It can provide rich image distribution information to the generator while enhancing semantic consistency and mitigating mode collapse.

## 3 Methodology

The proposed reverse interaction scheme with RIIN mutual constraints and optimization with the generator greatly enhance the generation ability of this network. To facilitate adversarial training of the enhanced generator, we correspondingly design a more powerful discriminator. This study adopts a single-stage generative structure as the basic framework illustrated in Fig. 2, which contains a text encoder, an adaptive affine-based generator, a dual-channel feature alignment discriminator, and a reversed image interaction network. Text encoder is a pre-trained BiLSTM [25] for text representation learning, which is responsible for extracting sentence vectors. The text encoder is a pre-trained BiLSTM [25] for text representation learning, responsible for extracting sentence vectors. It is trained with the help of DAMSM [32]. The DAMSM module maps images and sentences to a shared semantic space, measuring image-text similarity at an identical dimension. This method enhances the precision of the alignment process between the BiLSTM and the generated image features during training. As for the generator, we select six UP Blocks and 1 UP Block0 for obtaining $256 \times 256$ images. Each UP Block contains two AABs, which will be introduced in Sect. 3.2. The UP Block0 accepts the noise vector sampled from the normal distribution. The proposed RIIN contains five Down Blocks. It reversely transmits the feature information of the actual image to the generator through feature alignment of the inverse transformed feature map,

thus avoiding model collapse. The proposed DFAD conducted adversarial training with the generator and RIIN.

### 3.1 Reversed Image Interaction Network: RIIN

The optimization goal of GAN is to achieve the Nash equilibrium of the generator and discriminator, which usually leads to mode collapse due to unstable training. Although some methods [1, 2, 20] attempt to improve the stability of GAN training, few reports focus on modifying network architecture. To maximize the prior knowledge of text-image training pairs and let T2I models eliminate the dilemma that lacks the learning opportunities of image feature manifold structure, we propose the reversed image interaction network in this work. This study has changed the basic structure of the generator-discriminator pair in traditional GAN by embedding the proposed RIIN so that the RIIN can indirectly provide actual image prior knowledge. RIIN incorporates a custom-designed convolutional network, which synchronously shapes with the generator's output during the forward computation process. This unique feature allows modification of the output value of the intermediate feature map via affine transformation. The alignment feature in our model is domain-agnostic, focusing instead on the intermediate domains between text and image throughout the image generation process. RIIN will update the generator by reversed transformation for cooperation with the GAN mechanism. Thus, during the training process, RIIN can provide a stable image manifold structure based on the capabilities of the improved generator and discriminator. If the effect of RIIN is only to provide image feature manifolds, pre-training models with fixed parameters migrated from some large datasets, such as ResNet-101 [6], ViT-L/16 [4], can improve the ability to extract image feature manifolds at the early stage of training. However, the large pre-trained model is built on a single image modality. If the generation network features are forced to align with those of the pre-trained model, the results may not be satisfactory. Since the T2I model generator involves cross-modal transformation, there must be an intermediate domain in the transformation process from the text domain to the image domain. The intermediate domain is a potential spatial domain with text and image manifolds but does not favor either side. In the training stage, the intermediate domain cannot be controlled through external supervision. The intermediate domain of T2I generation is part of the latent space learned by GAN. This is the main reason many works use the GAN mechanism for implicit learning of image distributions. Monitoring the learning results of the intermediate domain is challenging because we need valid labels and metrics. Therefore, traditional GANs are prone to mode collapse, gradient disappearance, etc. Fortunately, in the training phase, we can easily access three information sources related to the intermediate domain: random noise distribution, text distribution, and actual image distribution. With the help of the GAN mechanism, we can obtain weakly-supervised labels from the proposed RIIN for intermediate domain learning by effectively using these three information sources. We develop the CBNs-based Affine Block in RIIN as shown in Fig. 2. It aims to make the network adaptively close to the intermediate domain representation and provide image information of feature maps from large to small. Moreover, this is a cross-mode task, the information contained in the generator's feature map at different resolutions should be different. A larger feature map ($128 \times 128$) carries more image information. Correspondingly, a smaller feature map ($8 \times 8$) takes more text information and noise distribution. The generator and RIIN are mutually constrained and optimized for each other.

In the proposed RIIN, the actual image $\hat{I}(256 \times 256)$ is sent to a convolution layer to obtain a $256 \times 256$ feature map $h_0 = conv(\hat{I})$. $h_0$ will then be conveyed to five Down Blocks

in turn. To quickly approach the intermediate domain features in the generator and make better use of the real image prior, as shown in the lower right corner of Fig. 2, each Down Block has four inputs: the output feature $h_r$ from the upper Down Block, the real image $\hat{I^{M\times M}}$, the sentence vector $s$, and the noise distribution $z$. Because we want to approach the implicit space with text feature, image feature, and input random noise distribution, this study considers incorporating all these three information source feature components and one upper-level feature into Down Block. The affine layer is a CBN framework inspired by [29], which takes the sentence vector $s$ as input and fuses the textual information into the feature map. Several parameters of CBN are defined as follows.

$$\gamma_r = \text{MLP1}(s, z), \quad \beta_r = \text{MLP2}(s, z), \quad \hat{h}_r \quad = \gamma_r \times h'_r + \beta_r, \tag{1}$$

where $\gamma_r \in \mathbb{R}^{N\times C_{in}}$ is the channel-wise scaling parameter, $\beta_r \in \mathbb{R}^{N\times C_{in}}$ is the shifting parameter. $h'_r \in \mathbb{R}^{N\times C_{in}\times H\times W}$ and $\hat{h}_r \in \mathbb{R}^{N\times C_{in}\times H\times W}$ are the input and output of the affine layer. $\hat{h}_r$ will be used to concatenate with the real image feature map to obtain the final Down Block output feature $h_r \in \mathbb{R}^{N\times C_{out}\times H\times W}$. The dimensions of RIIN output features are respectively $8\times 8$, $32\times 32$ and $128\times 128$, and these features will align with generator features of the same size. The feature alignment loss computes the mean square error between features for shortening the feature distance. This loss is used to establish a bi-directional constraint between the generator and RIIN, which is defined by

$$\mathcal{L}_{G_f} = \|h_r^{8\times 8} - h_g^{8\times 8}\|_2 + \lambda_1\|h_r^{32\times 32} - h_g^{32\times 32}\|_2 + \lambda_2\|h_r^{128\times 128} - h_g^{128\times 128}\|_2, \tag{2}$$

where $h_r^{M\times M}$ and $h_g^{M\times M}$ represent the $M \times M$ feature map of the RIIN and the generator, respectively. $\|\cdot\|_2$ represents the mean square error loss. Later experiments will discuss the impact of two hyper-parameters $\lambda_1$ and $\lambda_2$.

### 3.2 Adaptive Affine-Based Generator

According to [34], the isolated fusion modules conflict with each other, increasing the training difficulty. Therefore, this study designs an adaptive affine-based generator consisting of six UP Blocks and one UP Block0. We developed two AABs for each UP Block. When the proposed adaptive affine-based generator performs forward computation, the text embedding will be updated by the Adaptive Block in the UP Block (Fig. 2), and input into the subsequent affine layers. Figure 3 shows the architecture of Adaptive Block. This adaptive structure links isolated affine layers to the network backbone. It provides the affine transformation with the path to obtain the current layer feature and establishes long-term dependence on other network modules to reduce the training difficulty. The information enhancement of the AAB is mainly completed in the Adaptive Block. As shown in Fig. 3, the Adaptive Block has two inputs: the current layer feature map $h_g^{'M\times M}$ and the sentence vector $s$. The convolution kernel size is the same as the dimension of the feature map. $h_g^{'1} \in \mathbb{R}^{N\times 100}$ is the depth-wise convolution transformation output. The semantic condition $s_{con} \in \mathbb{R}^{N\times 356}$ is the result of concatenation between $h_g^{'1}$ and $s$, which will convey into the multilayer perceptron (MLP). We use Tanh as the final activation function to compute an attention map within values in (-1, 1). Finally, we adopt a residual structure to acquire the adaptive information-enhancing sentence vector $s'$ using

$$s' = s \times s_{attn} + s. \tag{3}$$

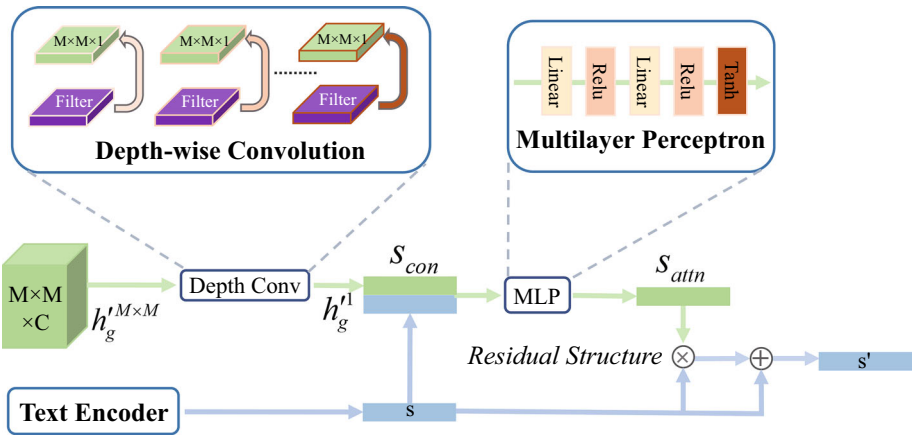


**Fig. 3** The architecture of Adaptive Block in AAB. Adaptive Block extracts feature $h_g^{'M\times M}$ through depth-wise convolution. By analyzing the matching of semantics and $h_g^{'M\times M}$, the attention map $s_{attn}$ of sentence vector $s$ is obtained. The block utilizes $s_{attn}$ to acquire information-enhancing sentence vector $s'$

$s'$ then is used to conduct the affine transformation. The channel-wise scaling parameter $\gamma_g \in \mathbb{R}^{N\times C_{in}}$ and the shifting parameter $\beta_g \in \mathbb{R}^{N\times C_{in}}$ are defined as follows

$$\gamma_g = \text{MLP1}(s', z), \quad \beta_g = \text{MLP2}(s', z), \quad \hat{h}_g = \gamma_g \times h'_g + \beta_g \tag{4}$$

where $h'_g \in \mathbb{R}^{N\times C_{in}\times H\times W}$ and $\hat{h}_g \in \mathbb{R}^{N\times C_{in}\times H\times W}$ are the input and output of affine layer, respectively.

### 3.3 Dual-Channel Feature Alignment Discriminator: DFAD

This study proposes a dual-channel feature alignment discriminator to enhance the semantic consistency between generated images and text descriptions. Since the major role of the discriminator in the T2I task is to capture the feature matching degree between the information of text and image feature map, the model should perform internal unwrapping and alignment operations on image and text features. The problem of extracting the matching information of two modes should be related. For example, extracting matched text information with fixed image features inevitably introduces the independent semantic part (redundant data) of image features. And extracting image-related features in the case of fixed text embedding will also have similar problems. Therefore, the extraction of matching information of the two modes should be carried out jointly to consider each other. Yin et al. [35] utilized Siamese structure and contrastive loss to forcibly improve the matching degree of image feature and text embedding features. Their strategy consumed enormous computing resources. Therefore, this study designs a new discriminator framework to improve the disentangling capabilities of the bi-modal information embedding of image and text. In the proposed discriminator, we need image and text feature information that can match each other. Different objects need to match different key information, which leads to the discriminator in the feature disentangling obtain some unmatched features if without feature selecting performance. So, feature alignment is required to guide feature extraction. Each channel of the dual-channel consists of the disentangling feature of one mode (text/image) and the original feature of the other (image/text). To improve the ability of the model to select correctly matched text and

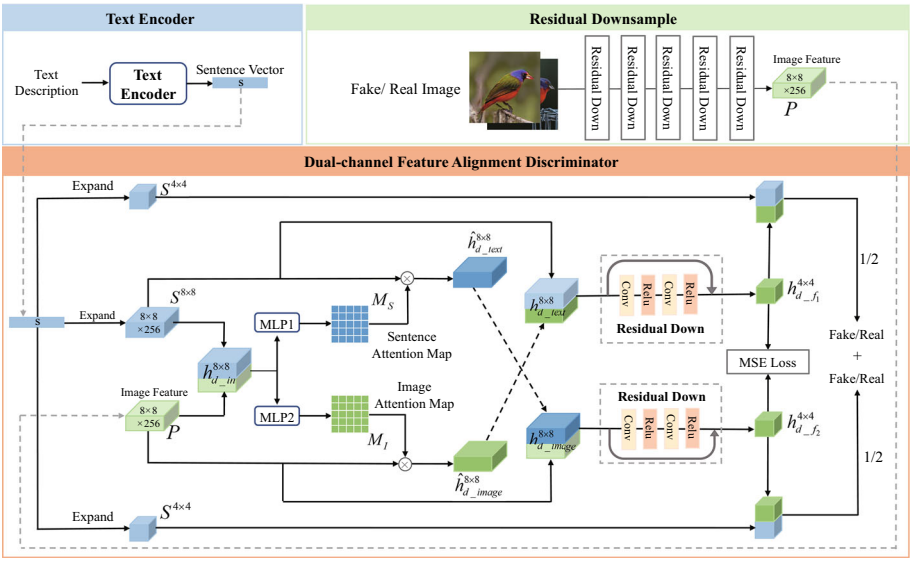


**Fig. 4** The architecture of the proposed DFAD. $h^{8\times8}_{d\_text}$ is the concatenation of raw expanded sentence embedding $S^{8\times8}$ and disentangling image feature $\hat{h}^{8\times8}_{d\_image}$. $h^{8\times8}_{d\_image}$ is the concatenation of raw image feature $P$ and disentangling expanded sentence embedding $\hat{h}^{8\times8}_{d\_text}$. $h^{8\times8}_{d\_image}$ and $h^{8\times8}_{d\_text}$ are delivered to the Residual Down block to align the dual-channel features behind

image features, we perform feature alignment constraint operations on the final disentangling fused features in two channels, as shown in Fig. 4.

The structure of the proposed DFAD is shown in Fig. 4. Firstly, one obtains the initial image feature map $P$ through a series of Residual Down blocks. The $P$ and the extended sentence embedding $S^{8\times8}$ are concatenated and sent respectively to the multilayer perceptron of two different channels (MPL1 and MPL2) to obtain the upper channel's sentence attention map $M_S$ and the lower channel's image spatial attention map $M_I$. We utilize $M_S$ and $M_I$ to obtain two disentangling feature maps $\hat{h}^{8\times8}_{d\_image}$ and $\hat{h}^{8\times8}_{d\_text}$ using

$$\begin{aligned} M_S &= MLP1(h^{8\times8}_{d\_in}), \quad M_I = MLP2(h^{8\times8}_{d\_in}) \\ \hat{h}^{8\times8}_{d\_image} &= M_S \times S^{8\times8}, \quad \hat{h}^{8\times8}_{d\_text} = M_I \times P \end{aligned} \tag{5}$$

Then, we respectively send the concatenation between $\hat{h}^{8\times8}_{d\_image}$ and $S^{8\times8}$, $\hat{h}^{8\times8}_{d\_text}$ and $P$ to two different residual blocks to obtain the feature map $h^{4\times4}_{d\_f_1}$ and $h^{4\times4}_{d\_f_2}$. These feature maps are used to align features. This study uses the mean square error loss $\mathcal{L}_{D_f}$ to represent the feature distance.

$$\mathcal{L}_{D_f} = \left\| h^{4\times4}_{d\_f_1} - h^{4\times4}_{d\_f_2} \right\|_2 \tag{6}$$

Since the discriminator adopts a dual-channel structure, the result of the forward computation is the average of the dual-channel output, as shown in Fig. 4. This study uses the adversarial

hinge loss $\mathcal{L}_{D_{adv}}$ [13] and MA-GP loss $\mathcal{L}_{D_{MA-GP}}$ [29] for the proposed discriminator.

$$\begin{aligned}\mathcal{L}_{D_{adv}} = &- \mathbb{E}_{\hat{I}\sim P_r}[\min(0, -1 + D(G(\hat{I}, s)))] - \mathbb{E}_{G(z)\sim P_g}[\min(0, -1 - D(G(z), s))] \\ &- \mathbb{E}_{\hat{I}\sim P_{mis}}[\min(0, -1 - D(\hat{I}, s))]\end{aligned} \tag{7}$$

$$\mathcal{L}_{D_{MA-GP}} = \mathbb{E}_{\hat{I}\sim P_r}[(||\nabla_{\hat{I}} D(\hat{I}, s)||_2 + ||\nabla_s D(\hat{I}, s)||_2)^p] \tag{8}$$

Generator adversarial loss $\mathcal{L}_{G_{adv}}$ is the cross-entropy loss defined by

$$\mathcal{L}_{G_{adv}} = -\mathbb{E}_{G(z)\sim \mathrm{P}_g}[D(G(z), s)] \tag{9}$$

Finally, we obtain the loss functions of discriminator, generator, and RIIN as follows:

$$\mathcal{L}_D = \mathcal{L}_{D_{adv}} + \mathcal{L}_{D_{MA-GP}} + \mathcal{L}_{D_f} \tag{10}$$

$$\mathcal{L}_G = \mathcal{L}_{G_{adv}} + \mathcal{L}_{G_f}, \quad \mathcal{L}_{RIIN} = \mathcal{L}_G. \tag{11}$$

To better understand the training process, we present the implementation details of the proposed RII-GAN framework with the following pseudo-code.

**Algorithm 1** Implementation details of RII-GAN.

**Input:** Text-image pairs $\{(t^{(1)}, \hat{I}^{(1)}), (t^{(2)}, \hat{I}^{(2)}), ..., (t^{(m)}, \hat{I}^{(m)})\}$;
**Input:** Noise prior: $\{z^{(1)}, z^{(2)}, ..., z^{(n)}\}$;
**Initialize:** Affine-based generator $G(\cdot)$, dual-channel feature alignment discriminator $D(\cdot)$, and reversed image interaction network RIIN$(\cdot)$;
Prepare pre-trained text encoder BiLSTM;
**repeat**
  Load one batch: $z$, $t$ and $\hat{I}$;
  Obtain sentence vectors: $s = \text{BiLSTM}(t)$;
  Determine attention map: $s_{attn} = \text{MLP}(s)$;
  Adaptive sentence vectors: $s' = s \times s_{attn} + s$;
  Get scaling and shifting parameters: $\gamma_g$ =MLP1$(s', z)$, $\beta_g$ =MLP2$(s', z)$;
  Intermediate generator feature map: $\hat{h}_g = \gamma_g \times h'_g + \beta_g$;
  Generate fake images and resolution-specific feature maps:

$$I, h_g^{8\times8}, h_g^{32\times32}, h_g^{128\times128} = G(\hat{h}_g);$$

  Propagate through RIIN for feature maps:

$$h_r^{8\times8}, h_r^{32\times32}, h_r^{128\times128} = \text{RIIN}(z, \hat{I}, s);$$

  Generate bimodal feature maps: $h_{d\_image}^{8\times8}, h_{d\_text}^{8\times8}$ using (5);
  Determine alignment loss: $\mathcal{L}_{D_f} = \left\| h_{d\_f_1}^{4\times4} - h_{d\_f_2}^{4\times4} \right\|_2$;
  Compute discriminator loss using (7), (8): $\mathcal{L}_D = \mathcal{L}_{D_{adv}} + \mathcal{L}_{D_{MA-GP}} + \mathcal{L}_{D_f}$;
  Optimize $D(\cdot)$ with gradients from $\mathcal{L}_D$;
  Compute feature alignment loss:

$$\mathcal{L}_{G_f} = \|h_r^{8\times8} - h_g^{8\times8}\|_2 + \lambda_1 \|h_r^{32\times32} - h_g^{32\times32}\|_2 + \lambda_2 \|h_r^{128\times128} - h_g^{128\times128}\|_2;$$

  Compute generator loss using (9): $\mathcal{L}_G = \mathcal{L}_{G_{adv}} + \mathcal{L}_{G_f}$;
  Update $G(\cdot)$, RIIN$(\cdot)$ with gradients from $\mathcal{L}_G$.
**until** iter times
Return updated weights: $D(\cdot)$, $G(\cdot)$, RIIN$(\cdot)$.

## 4 Experimental Results

In this section, we first introduce two benchmark datasets, evaluation metrics, and implementation details. Then, we conducted some experiments to compare the proposed RII-GAN with some SOTA GAN-based T2I algorithms, including PCCM-GAN [18], DF-GAN [29], DM-GAN [41], DTGAN [39], SAM-GAN [16], MirrorGAN [19], DR-GAN [28], DiverGAN [40], KD-GAN [17], SSA-GAN [12], DAE-GAN [23], AttnGAN [32], and StackGAN++ [37]. Section 4.4 and Sect. 4.9 discuss the parameter selection and a series of ablation studies.

### 4.1 Datasets

We conduct all experiments on two standard T2I datasets: CUB-Bird [31] and MS-COCO [14]. The former has 2,933 test birds (50 categories) and 8,855 training birds (150 categories). Each bird has ten English sentences to describe the fine-grained visual scenes. The latter consists of 82,783 training images and 40,504 test images. Each image has five sentence annotations. Compared with CUB-Bird dataset, images in MS-COCO present complex visual scenes, which make the T2I generation tasks more challenging.

### 4.2 Evaluation Metrics

We use three broadly utilized metrics CLIPScore (CS) [7], Inception Score (IS) [24], and Fréchet Inception Distance (FID) [8] to quantify the performance of all approaches in terms of image-text alignment and image quality.

*Image-text alignment:* CS adopts a pre-trained CLIP model [21] to map image and text into the same feature space and calculates the cosine similarity between the text feature and image feature. A larger CS indicates that the generated image has a more significant semantic similarity to the text.

*Image quality:* IS evaluates image quality by calculating KL divergence between marginal distribution (authentic image) and conditional distribution (generated image). The larger IS represents that the generated image is high quality in authenticity and diversity. FID is a metric used to measure the distribution consistency between generated and authentic images. A lower FID means the generated image is close to the authentic image.

### 4.3 Implementation Details

Our proposed model RII-GAN is implemented using the PyTorch toolbox. The model training consists of two distinct phases. In the initial phase, the two independent BiLSTMs [21] are trained as text encoders, leveraging the DAMSM approach for computing image-text similarity [2]. Specifically, we employ DAMSM to train a pair of text and image encoders that can align with each other, thereby obtaining an efficient text encoder. Each BiLSTM is trained using CUB-Bird and MS-COCO datasets on a single RTX 3090 GPU. The BiLSTM outputs global sentence vectors with 256 dimensions used as text embeddings for the subsequent generator, RIIN, and the discriminator. In the second phase, we freeze the text encoder and train the GAN. We assign a batch size of 32 for both CUB-Bird and MS-COCO datasets. The generator and RIIN in our framework are simultaneously trained using Adam optimizer [9], with parameters $\beta_1$ and $\beta_2$ set to 0.5 and 0.999, respectively. The optimizer is initiated with a learning rate $2e-4$, which starts to linearly decays over the final one-third of the

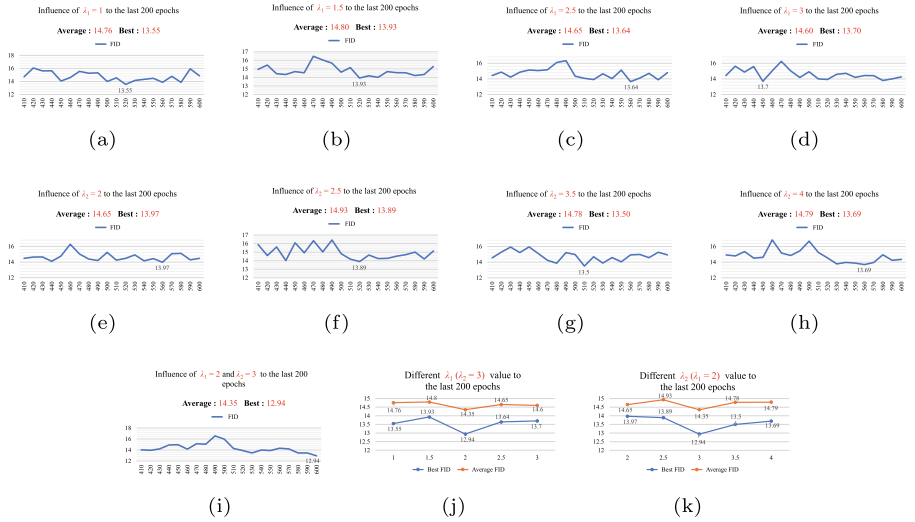


**Fig. 5** FID metrics of our RII-GAN under different $\lambda_1$ and $\lambda_2$ on CUB-Bird dataset in the last 200 training epochs (total of 600 training epochs). The horizontal axis denotes the training epochs. The first row evaluates the effect of $\lambda_1$ when $\lambda_2$ is fixed as 3, and the second row investigates the impact of $\lambda_2$ when $\lambda_1$ is fixed as 2; **i** FID for $\lambda_1 = 2, \lambda_2 = 3$; **j** the best and average FID ($\lambda_2 = 3$); **k** the best and average FID ($\lambda_1 = 2$)

training epochs approaching $5e - 5$. In parallel, the discriminator also employs the Adam optimizer with $\beta_1$ and $\beta_2$ set to 0.5 and 0.999, respectively and a fixed learning rate of $4e-4$. Throughout the second phase, RIIN training status is synchronized with the generator and alternately trained with the discriminator. The hyperparameters $\lambda_1$ and $\lambda_2$ are respectively set to 2 and 3. The subsequent experimental sections present a comprehensive discussion of the choice of these parameters. The model undergoes training for 600 epochs on CUB-Bird dataset and 200 on MS-COCO dataset. The image generation process only retains the generator component during the testing phase.

### 4.4 Parametric Sensitivity Analysis

In this study, $\lambda_1$ and $\lambda_2$ are two important hyper-parameters in the feature alignment loss $\mathcal{L}_{G_f}$ of the generator and RIIN. According to (2), these parameters will guide the generator and RIIN to pay more attention to the feature corresponding to the larger parameter. Figure 5 shows the FID scores in the last 200 training epochs on CUB-Bird dataset with different $\lambda_1$ and $\lambda_2$. The first row evaluates the effect of $\lambda_1$ when $\lambda_2$ is fixed as 3, and the second row investigates the impact of $\lambda_2$ when $\lambda_1$ is fixed as 2. The best FID score with different parameters is marked with red numerals. Empirically, we found from the first and the second rows of Fig. 5 that when $\lambda_1 = 1, 1.5, 2.5, 3 (\lambda_2 = 3)$ or $\lambda_2 = 2, 2.5, 3.5, 4 (\lambda_1 = 2)$, the FID scores are unstable and fluctuates obviously with the increase of epochs. Figure 5j and k summarize the optimal and average FID scores obtained in the last 200 training epochs when $\lambda_1$ and $\lambda_2$ take different values. We observe that when $\lambda_1 = 2$, $\lambda_2 = 3$, our model's average, and optimal FID scores reach the bottom, indicating that the network reaches a local minimum. As shown in Fig. 5i, taking $\lambda_1 = 2, \lambda_2 = 3$, the model obtains the minimum FID score under relative stability, especially in the last 100 epochs. Therefore, parameters $\lambda_1$ and $\lambda_2$ are set to 2 and 3 in all experiments.

**Table 1** Quantitative comparison of different models on CUB-Bird and MS-COCO datasets

| Methods | CUB-Bird | | | MS-COCO | |
|---|---|---|---|---|---|
| | FID↓ | IS↑ | CS↑ | FID↓ | CS↑ |
| AttnGAN (CVPR'18) | 23.98 | 4.26±.03 | 70.83 | 35.49 | 52.58 |
| StackGAN++ (TPAMI'18) | 15.30 | 4.04±.06 | – | 81.59 | – |
| MirrorGAN (CVPR'19) | 18.34 | 4.56±.05 | – | 34.71 | – |
| DM-GAN (CVPR'19) | 16.09 | 4.75±.07 | ***71.91*** | 32.64 | 53.23 |
| PCCM-GAN (Neuro'21) | 22.15 | 4.65±.20 | – | 33.59 | – |
| SAM-GAN (Neural Network'21) | 20.49 | 4.61±.03 | – | 33.41 | – |
| DTGAN (IJCNN'21) | 16.35 | 4.88±.03 | – | 23.61 | – |
| DAE-GAN (ICCV'21) | 15.29 | 4.42±.04 | 71.8 | 28.12 | 54.04 |
| KD-GAN (TMM'21) | 13.89 | 4.90±.06 | – | 23.92 | – |
| SSA-GAN (CVPR'22) | 15.61 | 5.17±.08 | – | 19.37 | – |
| DR-GAN (TNNLS'22) | 14.96 | 4.90±.05 | – | 27.8 | – |
| DF-GAN (CVPR'22) | 14.81 | 5.10±. – | 70.63 | 19.32 | ***61.91*** |
| DiverGAN (Neuro'22) | 15.63 | 4.98±.06 | – | 20.52 | – |
| RII-GAN (Ours) | ***12.94*** | ***5.41±.02*** | 70.85 | ***19.01*** | 60.28 |

The italic values denotes the best result

### 4.5 Quantitative Results

Table 1 shows the quantitative results of our RII-GAN and several advanced T2I models on CUB-Bird and MS-COCO datasets. The results of these T2I models on two datasets come from their publicly available codes on the web. This table shows that our RII-GAN obtains the third-highest CS, lower than the baseline DM-GAN. The DM-GAN achieves the best CS (71.91), indicating better text-image consistency maintenance on CUB-Bird dataset. Our RII-GAN on CUB-Bird dataset achieves the lowest FID scores (12.94) and the highest IS (5.41). These results show that the images generated by our RII-GAN are closer to the actual image distribution in terms of maintaining image quality. On the large-scale MS-COCO dataset, our RII-GAN also achieves competitive results. The statistics in the second last column in Table 1 show that our RII-GAN achieves the lowest FID score of 19.01 among current state-of-art methods, which indicates the ability to generate more realistic images. The CS of our RII-GAN is much higher than that of AttnGAN, DM-GAN, and DAE-GAN and is almost on the same level as the state-of-the-art DF-GAN. The score is relatively much better than the CS scores in CUB-Bird dataset. It illustrates that our RII-GAN is much superior to AttnGAN, DM-GAN, and DAE-GAN in terms of semantic consistency on large datasets with the help of RIIN, AAB, and DFAD. Three metrics on both datasets demonstrate that our RII-GAN can generate higher-quality images.

### 4.6 Qualitative Results

In this section, we visualized some synthesized images by our RII-GAN and three other advanced models: AttnGAN [32], DM-GAN [41], and DF-GAN [29] on CUB-Bird and MS-COCO datasets shown in Fig. 6. AttnGAN and DM-GAN are two classical multi-stage generation methods, and DF-GAN and our RII-GAN are one-stage generation methods. The

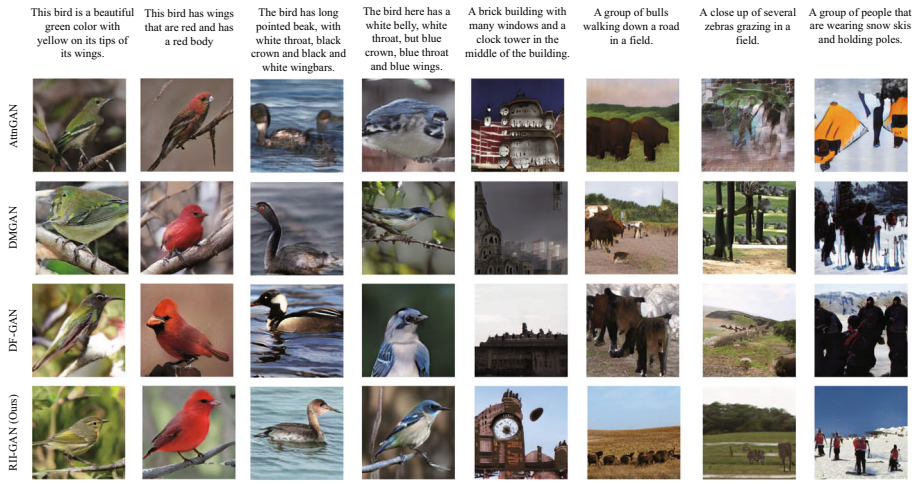


**Fig. 6** Visual comparisons on CUB-Bird (the first four column) and MS-COCO (the last four column) datasets

images in each column are generated by different approaches based on the scene's text at the top of each column. These text descriptions are randomly selected from datasets. According to semantic consistency, our method captures more text detail descriptions. As shown in Fig. 6, in the first column, our RII-GAN captures the "yellow on its tips of its wings" detail in "green color with yellow on the tips of its wings" and reflects it in the generated image, which fails in the other methods. In the third column, the generated images of AttnGAN, DMGAN, and DF-GAN show a lot of overlap between the water scene and the bird body. By contrast, the images generated by our RII-GAN method can reflect the details and colors of birds on the lake. For a more challenging MS-COCO dataset, it can be observed that our RII-GAN can still generate images of the complex scene, e.g., RII-GAN successfully focuses on the "clock tower" keyword information. In contrast, AttnGAN, DMGAN, and DF-GAN only generated objects with inaccurate shapes through the keyword "tower" (the fifth column). In columns 6 and 7, the "bulls" and "zebras" are visible in our method but hard to recognize in others. As shown in the 8th column, the proposed model better reflects the meaning of the text, such as "a group of people". Generally, these subjective visual comparison results confirm the effectiveness of the proposed RII-GAN.

### 4.7 User Evaluation

Since there is no adequate convincing criterion to judge the consistency of image-text semantics of the generated image, we conducted a human test shown in Fig. 7. The experiment randomly selected 100 text descriptions from CUB-Bird dataset and then invited 30 volunteers to compare the results of our RII-GAN and two popular baselines AttnGAN [32] and DF-GAN [29]. In this test, the user receives three images generated from three models and a corresponding text description. Each user sorts these images according to realism and text alignment. As summarized in Fig. 7, our RII-GAN achieved the highest ranking of 50.46% and 63.08%, respectively, regarding text alignment and realism. The conclusion also shows that our RII-GAN has obtained better approval in human perception.

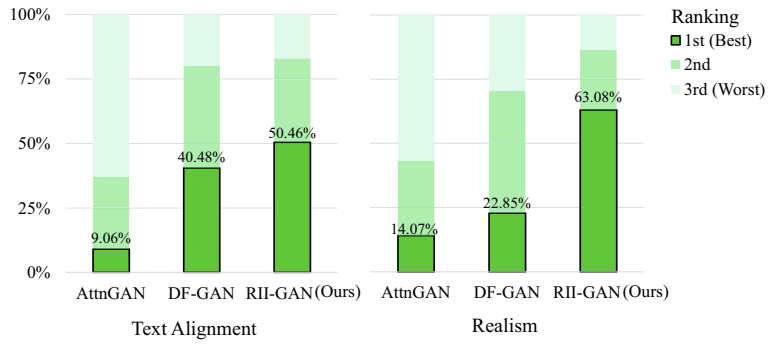


**Fig. 7** Human evaluation on CUB-Bird dataset for text alignment and realism

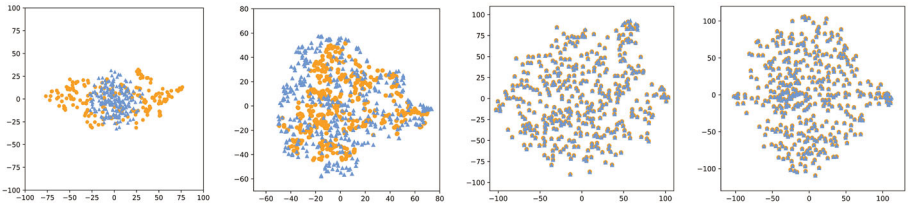


**Fig. 8** Visualization of feature alignment between generator (blue triangles) and our RIIN (orange circles) for intermediate features during forward computation on CUB-Bird dataset, **a** 2D distribution of noise $z \sim N(0, 1)$ and real images; **b**, **c**, and **d** are feature distributions with the scale of $128 \times 128$, $32 \times 32$, and $8 \times 8$, respectively

### 4.8 Performance Analysis

*Visualization by t-SNE analysis.* To investigate the feature distribution between the RIIN and the proposed adaptive affine-based generator, we utilize the t-SNE algorithm to visualize the feature alignment between the generator (blue triangles) and our RIIN (orange circles) on CUB-Bird dataset. Figure 8a shows the two-dimensional distribution of noise $z \sim N(0, 1)$ and real images before forward computing. Figures 8b, c, d show feature distributions of different scales. We found that the feature distance between the generator and RIIN is large in the initial stage. However, as the forward computing of RIIN gets deep into the reversed image interaction process (from $128 \times 128$ to $8 \times 8$). The feature distributions of both modules successfully converge on the same intermediate domain (see Figs. 8c, d).

*Image diversity* The randomness of the noise $z$ controls the diversity of the generated images. $z$ is sampled from a Gaussian distribution $N(0, 1)$. To explore the influence of $z$ on the generated images, we conducted control-variable experiments on the text input and random noise $z$ of our RII-GAN, as shown in Fig. 9. The random noise $z$ is fixed at the same level. The sentence embedding of the input model varies by changing the specific words in the text description. From the generation images shown in Fig. 9, we found when the color attribute in the description changes, the images generated by RII-GAN can keep other elements except for the color unchanged (such as the bird's pose and background information). This result demonstrates the high controllability of our method for generating images. If the text embedding is fixed under the same vertical line, changing the random sampling noise $z$, we found that different $z$ significantly impacts the object pose and background of the gen-


**Fig. 9** Generated samples using our RII-GAN by controlling different text inputs and randomly sampled noise $z$

erated image. By varying the input $z$, the generator synthesizes bird images with different angles and backgrounds. Therefore, the results of Fig. 9 demonstrate that our RII-GAN can effectively disentangle the attribute of the input text description and control the modeling of the sample relevant regions while not affecting the generation of non-relevant regions during the disentanglement process. Figure 10 presents more examples of images synthesized by the proposed RII-GAN on CUB-Bird dataset.

### 4.9 Ablation Study

*Impact of RIIN, DFAD and AAB.* The proposed framework comprises three main modules: AAB, RIIN, and DFAD. We carried out ablation experiments to validate the impact of each module in our RII-GAN on CUB-Bird dataset. The baseline model is a one-stage GAN-based T2I network. The evaluation metrics are FID and IS summarized in Table 2. The role of AAB is to use the current feature layer to adaptively provide a basis for updating text embedding. Without (w/o) AAB, the FID and IS scores are 13.25 and 5.37±0.03 on CUB-Bird dataset while the FID value on MS-COCO dataset is 21.23. It verifies that adding AAB into the framework helps generate more realistic images. We also investigate our model with (w/) AAB and RIIN, and without DFAD, we find the FID increases from 12.94 to 14.03, and IS decays from 5.41 to 5.39 on CUB-Bird dataset and the FID increases from 19.01 to 20.03 on MS-COCO dataset. It can be concluded from the table that the use of the dual-channel feature alignment structure is beneficial for semantic extraction and matching of the discriminator. Finally, we also evaluate the effect of the RIIN module. When RIIN is removed from our model, the FID has significantly increased by 9.89%, and IS has decreased by 2.96% on CUB-Bird dataset. On the other hand, the FID increased substantially by 11.68% on MS-COCO dataset. The results in Table 2 reflect that the proposed RIIN has a more significant impact on the performance of our framework than AAB and DFAD.

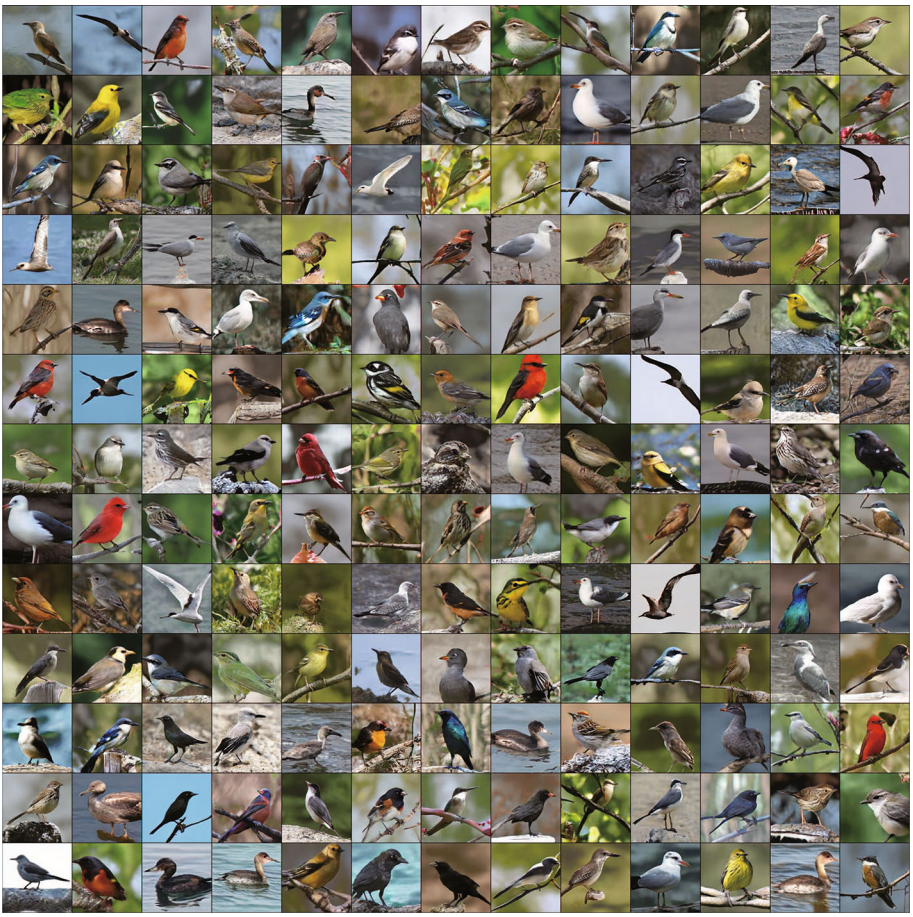

**Fig. 10** Images synthesized by the proposed RII-GAN on CUB-Bird dataset by controlling different text inputs and randomly sampled noise $z$

**Table 2** The performance of different components of our model on CUB-Bird and MS-COCO datasets, in which "AAB" means adaptive affine block, "DFAD" represents dual-channel feature alignment discriminator, and "RIIN" stands for reversed image interaction network

| Components | | | CUB-Bird | | MS-COCO |
|---|---|---|---|---|---|
| RIIN | DFAD | AAB | FID↓ | IS↑ | FID↓ |
| w/o | w/o | w/o | 15.89 | 5.12±0.01 | 23.47 |
| w/ | w/ | w/o | 13.25 | 5.37±0.03 | 21.23 |
| w/ | w/o | w/ | 14.03 | 5.39±0.03 | 20.03 |
| w/o | w/ | w/ | 14.22 | 5.25±0.02 | 19.84 |
| w/ | w/ | w/ | ***12.94*** | ***5.41±0.02*** | ***19.01*** |

The best results are highlighted in italic

In Fig. 11, we show the generated images of the Baseline (Base.), RII-GAN w/o RIIN, RII-GAN w/o DFAD, RII-GAN w/o AAB, and RII-GAN on CUB-Bird dataset ((a)-(e)) and MS-COCO dataset((f)-(j)). As shown in Fig. 11, our RII-GAN performs the best in image generation quality and semantic consistency, which is more evident on CUB-Bird dataset. By ablating each module of AAB, DFAD, and RIIN, it can be observed that model w/o RIIN has

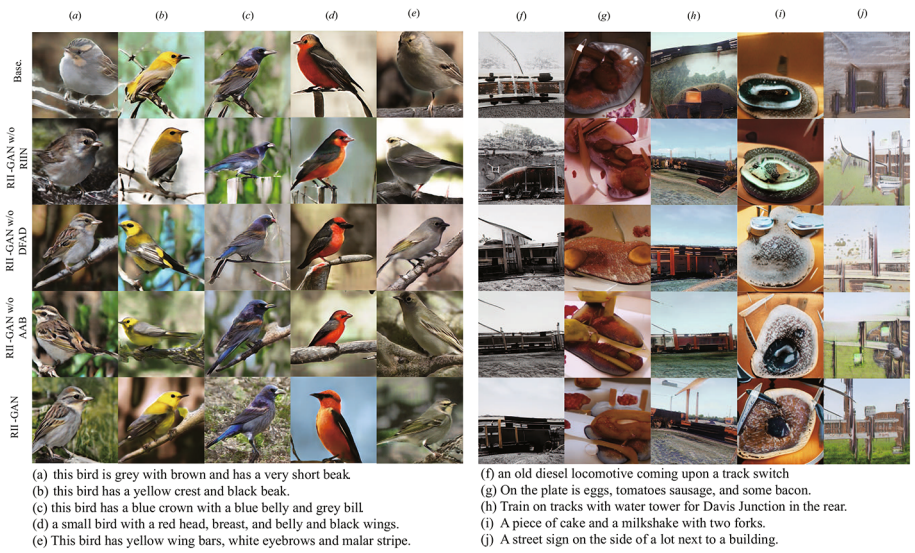


**Fig. 11** The figure provides a comparative study of image generation from identical text descriptions across different models: the Baseline (Base.), RII-GAN without (w/o) RIIN, RII-GAN w/o DFAD, RII-GAN w/o AAB, and the RII-GAN. Each column of images corresponds to a single text description, forming ten groups labeled (**a**)–(**j**). Text descriptions displayed on the left (**a**-**e**) are from CUB-Bird dataset, and text descriptions on the right (**f**-**j**) are from MS-COCO dataset

the greatest effect on the quality of the generated images (eg. (b), (c), (j) of Fig. 11). This is because RIIN directly interacts with the generator to learn the real image feature manifold. In detail, it reversely encodes the features of the real image, and they are aligned with the features forward encoded by the generator. Besides, AAB and DFAD also significantly contribute to the model's ability to generate images that are more realistic and richer in semantic detail ("tomatoes, egg" in (g), "forks" in (i), and "lot" in (j) of Fig. 11) and object layout ("water tower" in (h) of Fig. 11). Specifically, AAB is used to adaptively update the text embedding, which facilitates the affine module to capture the text information acquired by the generator, thereby enriching the semantic detail of the model. Additionally, the dual-channel structure of DFAD enhances the robustness of the discrimination process as it more effectively identifies the critical matched semantics of text and image and dismisses redundant information.

**The dual-channel feature selection scheme of DFAD.** To further explore the impact of dual-channel feature selection and feature alignment of DFAD on the generated images, we conducted the experiments shown in Table 3. When the alignment structure of the discriminator is removed, the FID score increases from 12.94 to 15.92 with no change in IS. This result demonstrates that the feature alignment process significantly affects the quality of generated images. We also found that removing different modal feature channels (image channel/text channel) has a significant impact on the model's performance. As indicated in Table 3, the FID score increases by 0.63, and the IS decreases by 0.11 (compared to our DFAD in the first row) if only the image channel is removed. The FID score increases by 3.97, but the IS increases by 0.06 after removing the text channel. The experimental results demonstrate that the joint action of the dual-channel structure and feature alignment work together to achieve better results.

**Table 3** Evaluate the dual-channel feature selection scheme in DFAD

| Architectures | FID ↓ | IS ↑ |
|---|---|---|
| DFAD | ***12.94*** | 5.41±0.02 |
| DFAD w/o Aligning | 15.92 | 5.41±0.01 |
| DFAD w/o Image Channel | 13.57 | 5.30±0.03 |
| DFAD w/o Text Channel | 16.91 | ***5.47±0.05*** |

The best results are highlighted in italic

**Table 4** Evaluation of multi-scale feature alignment mechanism in RIIN

| Architecture | FID ↓ | IS ↑ |
|---|---|---|
| RIIN | ***12.94*** | ***5.41±0.02*** |
| RIIN w/o 8 × 8 branch | 13.12 | 5.37±0.04 |
| RIIN w/o 32 × 32 branch | 13.45 | 5.32±0.01 |
| RIIN w/o 128 × 128 branch | 14.10 | 5.27±0.01 |

The best results are highlighted in italic

**Table 5** Evaluation of how the performance is affected by different numbers of AABs used in the RII-GAN. The stages indicate the number of UP Blocks that contain AABs in the proposed RII-GAN

| Architecture | Stages | FID ↓ | IS ↑ |
|---|---|---|---|
| AAB | 1 | 13.33 | 5.32±0.04 |
| | 2 | 13.43 | 5.36±0.03 |
| | 3 | 13.02 | 5.29±0.01 |
| | 4 | 13.15 | 5.39±0.04 |
| | 5 | 13.01 | 5.38±0.02 |
| | 6 | ***12.94*** | ***5.41±0.02*** |

The best results are highlighted in italic

*Multi-scale feature alignment mechanism of RIIN*. In addition, we investigated the effect of the multi-scale feature alignment scheme in RIIN. We carried out different combinations of feature alignment and summarized the results in Table 4. Row 2-4 represents the absence of a branch in RIIN. According to comparison results, we found that the effect of feature alignment in terms of optimization capability can increase sequentially with increasing spatial resolutions. This phenomenon demonstrates that the reversed interaction process in large resolutions helps the network to generate more realistic images.

*The configuration and quantities of AAB in UP Blocks.* Then, as presented in Table 5, we evaluate the impact of different numbers of AABs in RII-GAN on CUB-Bird dataset. Column 'stages' in Table 5 corresponds to the number of UP Blocks with AABs, where in this experiment, we comprise two AABs in each UP Block as shown in Fig. 2. UP Blocks are accumulated in ascending resolution order and so does the AABs in each UP Block. The generator needs 6 UP Blocks as images would need to be generated from $4 \times 4$ to $256 \times 256$. Hence, we conduct experiments for stages 1 to 6. For example, the sub-experiment for stage 5 has only the first 5 UP Blocks containing AABs. It can be observed that as the quantity of UP Blocks that contain AABs increases, there is a consequent decrease in the model's FID and an increase in the IS. This trend indicates a continuous enhancement in model performance by an increased number of AABs up to the maximum, which proves its effectiveness.

Another experiment is conducted to evaluate the number of AABs in each UP Block of the generator, as presented in Table 6. Controlling the total number of affine blocks as 4 in each UP Block, which already ensures a sufficient text-to-image fusion [34] [29], we change the number of AABs while appropriately allocating base affine blocks to figure out the optimal

**Table 6** Evaluation of how the performance is affected by different numbers of AABs used in each UP Block

| Architecture | Numbers | FID ↓ | IS ↑ |
|---|---|---|---|
| AAB | 1 | 14.10 | 5.12±0.03 |
| | 2 | ***12.94*** | ***5.41±0.02*** |
| | 3 | 13.13 | 5.22±0.04 |
| | 4 | 13.72 | 5.19±0.01 |

The best results are highlighted in italic

configuration. The experimental models are i) one AAB within each UP Block, which consists of four base affine blocks; ii) two AABs within each UP Block which consists of two base affine blocks; iii) three AABs within each UP Block, where the first two AABs incorporate single base affine block and the last AAB consists of two base affine blocks; iv) four AABs within each UP Block, each incorporating one base affine. From the experimental results, we observed that combining two AABs in each UP Block gives the optimal quantitative results for both FID and IS. FID of the optimal configuration reaches 12.94, slightly lower than the second-best 13.13 of three AABs, IS scores of the other three models are approximately at a similar level, while the chosen model is relatively high, reaching 5.41. This is mainly because this architecture potentially establishes a balance capable of preserving model depth while avoiding excessive complexity in the network.

## 5 Limitation and Discussion

Our RII-GAN presents notable advantages over existing methods, but there are a few limitations that may lead to bad cases. To reduce computational overhead, our approach favors using global sentence embeddings by BiLSTM [21] over word embeddings for text-to-image fusion. This, however, may hinder semantic disentangling during the generation of images corresponding to complex scene text. Specifically, in generation tasks involving complex scene text, there are several main causes of bad cases. Firstly, complex scene text descriptions may consist of multiple discrete objects, actions, or concepts interacting in varied ways. In this case, global embedding might not preserve the unique semantics of each component, leading to images generated that are blurred or lacking precise representation. In addition, despite theoretically mitigating the vanishing gradient problem, BiLSTM can also exhibit a diminished capacity for capturing long-range dependencies in practice. Critical semantic connections between words or phrases situated far apart in the text might not be sufficiently integrated into the global sentence embedding. As shown in Fig. 12, we can see that for complicated sentences, there may be some issues, such as the omitted words “orange beak” and “table” (columns 1 and 4), blurred background (column 2), and artifacts (column 3). These issues more often occur on MS-COCO, where sentences describe more complex scenes. For further improvement, alternative strategies such as incorporating attention mechanisms or transformer models that can more effectively handle complex semantics and long-range dependencies could be applied.

The proposed auxiliary network module RIIN enhances the generator training path by processing actual image distributions. It is designed with a relatively basic architecture taking into consideration computational resources. Amplifying RIIN would more precisely capture and encode real image features at various resolutions. Thus, the network could better align these features with the intermediate layer features of the generator, further integrating true image feature information into the optimization process of the generator. Later studies would


**Fig. 12** Failure examples generated by our methods on CUB-Bird (the first two columns) and MS-COCO datasets (the last two columns)

like to improve feature alignment and overall model performance and may focus on this structure while it also requires increasing computational resources.

## 6 Conclusion

In this work, we propose a novel T2I framework to synthesize realistic images semantically consistent with the given text description. The proposed RII-GAN is different from the previous method. It uses RIIN to optimize the training path of the generator and provides the generator with a reverse image feature manifold structure so that the generator has more opportunities to learn the actual image distribution. AAB modules in the proposed adaptive affine-based generator establish a compelling connection between independent affine modules and improve the existing T2I affine transformation methods, which lack interaction with the feature information of the generator. Moreover, we developed a DFAD to capture the key-matching semantics of text embeddings and image features. The qualitative and quantitative analyses of two standard datasets demonstrate that our RII-GAN can generate more semantic-related and diversified images. Future work may include two directions: (i) synthesize objects with complex scenes; (ii) cross-linguistic T2I synthesis based on the reversed image interaction process.

**Acknowledgements** The authors would like to thank the anonymous reviewers and the associate editor for their insightful comments that significantly improved the quality of this paper. This work was supported by the National Nature Science Foundation of China under Grant 61872143.

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


1. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, pp 214–223
2. Berthelot D, Schumm T, Metz L (2017) BEGAN: boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717
3. Chen Z, Luo Y (2019) Cycle-consistent diverse image synthesis from natural language. In: 2019 IEEE international conference on multimedia & expo workshops (ICMEW), IEEE, pp 459–464
4. Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
5. Goodfellow I, Pouget-Abadie J, Mirza M et al (2020) Generative adversarial networks. Commun ACM 63(11):139–144
6. He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778
7. Hessel J, Holtzman A, Forbes M, et al (2021) Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718
8. Heusel M, Ramsauer H, Unterthiner T, et al (2017) GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems, pp 662–6637
9. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
10. Li B, Qi X, Lukasiewicz T, et al (2019) Controllable text-to-image generation. In: Advances in neural information processing systems, pp 2065–2075
11. Li B, Torr PHS, Lukasiewicz T (2022) Memory-driven text-to-image generation. arXiv preprint arXiv:2208.07022
12. Liao W, Hu K, Yang MY, et al (2022) Text to image generation with semantic-spatial aware GAN. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 18,166–18,175
13. Lim JH, Ye JC (2017) Geometric GAN. arXiv preprint arXiv:1705.02894
14. Lin TY, Maire M, Belongie S, et al (2014) Microsoft COCO: common objects in context. In: European conference on computer vision, pp 740–755
15. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784
16. Peng D, Yang W, Liu C et al (2021) SAM-GAN: self-attention supporting multi-stage generative adversarial networks for text-to-image synthesis. Neural Netw 138:57–67
17. Peng J, Zhou Y, Sun X et al (2022) Knowledge-driven generative adversarial network for text-to-image synthesis. IEEE Trans Multimed 24:4356–4366
18. Qi Z, Sun J, Qian J et al (2021) PCCM-GAN: photographic text-to-image generation with pyramid contrastive consistency model. Neurocomputing 449:330–341
19. Qiao T, Zhang J, Xu D, et al (2019) MirrorGAN: Learning text-to-image generation by redescription. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 1505–1514
20. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434
21. Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: Proceedings of the 38th international conference on machine learning, pp 8748–8763
22. Reed S, Akata Z, Yan X, et al (2016) Generative adversarial text to image synthesis. In: Proceedings of the 33rd international conference on machine learning, pp 1060–1069
23. Ruan S, Zhang Y, Zhang K, et al (2021) DAE-GAN: Dynamic aspect-aware gan for text-to-image synthesis. In: 2021 IEEE/CVF international conference on computer vision (ICCV), pp 13,940–13,949
24. Salimans T, Goodfellow I, Zaremba W, et al (2016) Improved techniques for training GANs. In: Advances in neural information processing systems, p 2234-2242
25. Schuster M, Paliwal K (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681
26. Szegedy C, Vanhoucke V, Ioffe S, et al (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2818–2826
27. Tan H, Liu X, Li X, et al (2019) Semantics-enhanced adversarial nets for text-to-image synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 10,500–10,509

28. Tan H, Liu X, Yin B et al (2022) DR-GAN: distribution regularization for text-to-image generation. IEEE Trans Neural Netw Learn Syst 34:1–15
29. Tao M, Tang H, Wu F, et al (2022) DF-GAN: A simple and effective baseline for text-to-image synthesis. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 16,494–16,504
30. de Vries H, Strub F, Mary J, et al (2017) Modulating early visual processing by language. In: Advances in neural information processing systems, pp 6594–6604
31. Wah C, Branson S, Welinder P, et al (2011) The caltech-ucsd birds-200-2011 dataset
32. Xu T, Zhang P, Huang Q, et al (2018) AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 1316–1324
33. Yang Y, Wang L, Xie D et al (2021) Multi-sentence auxiliary adversarial networks for fine-grained text-to-image synthesis. IEEE Trans Image Process 30:2798–2809
34. Ye S, Liu F, Tan M (2022) Recurrent affine transformation for text-to-image synthesis. arXiv preprint arXiv:2204.10482
35. Yin G, Liu B, Sheng L, et al (2019) Semantics disentangling for text-to-image generation. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 2322–2331
36. Zhang H, Xu T, Li H, et al (2017) StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: 2017 IEEE international conference on computer vision (ICCV), pp 5908–5916
37. Zhang H, Xu T, Li H et al (2019) StackGAN++: realistic image synthesis with stacked generative adversarial networks. IEEE Trans Pattern Anal Mach Intell 41(8):1947–1962
38. Zhang H, Koh JY, Baldridge J, et al (2021) Cross-modal contrastive learning for text-to-image generation. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 833–842
39. Zhang Z, Schomaker L (2021) DTGAN: Dual attention generative adversarial networks for text-to-image generation. In: 2021 International joint conference on neural networks (IJCNN), pp 1–8
40. Zhang Z, Schomaker L (2022) DiverGAN: an efficient and effective single-stage framework for diverse text-to-image generation. Neurocomputing 473:18–198
41. Zhu M, Pan P, Chen W, et al (2019) DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5795–5803