The Visual Computer manuscript No.
(will be inserted by the editor)

Stacked Deep Fusion GAN for Enhanced Text-to-Image Generation

Wen-li Chen · Ya-qi Sun · Paul L. Rosin · Yu-kun Lai

Abstract Generating high-quality, semantically consistent images from text descriptions remains a challenging task in computer vision. Current methods often struggle with effectively integrating textual information into the image generation process, resulting in images that lack realism or contain significant artifacts. To address these issues, we propose SDeep, a novel framework utilizing a Generative Adversarial Network (GAN) architecture with a channel-attention mechanism. SDeep deepens the text-to-image fusion process through Staked Deepening Blocks (SD Blocks) and enhances image detail through Multilayer Channel Attention (MLCA Attention). Extensive experiments on the CUB and COCO datasets demonstrate that SDeep outperforms state-of-the-art methods in terms of image quality and semantic alignment with text descriptions. Our approach not only generates more realistic images but also better preserves the semantic consistency between text and generated images, marking a significant advancement in text-to-image synthesis. Code can be found at https://github.com/zxcnmmmmm/SDeep.

Keywords Text to Image, Image Generation, Generative Adversarial Network, Feature extraction.

1 Introduction

With a great deal of attention, the field of vision processing is growing by leaps and bounds. In terms of applications, deep learning techniques have been used by many researchers to accomplish tasks such as natural language processing [1], semantic segmentation [2,3], remote sensing [4] and model compression [5,6]. Among them, image synthesis, as a research direction with great potential, has absorbed the attention of many researchers.

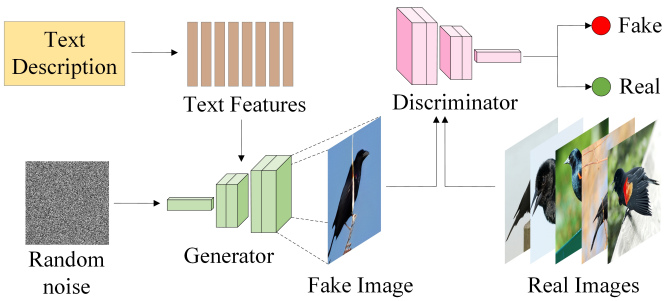

Fig. 1 Simplified architecture of a GAN for text-to-image generation. The Generator and Discriminator in the figure can be one or more groups.

Text to Image (T2I) is achieved by using technical methods to produce real images that match a given text description. After the first T2I generation using a GAN in 2016 was implemented by Reed et al. [7], GANs occupy an important place in the work on image generation. Regardless of the type of GAN used, text description is first processed into textual features, which are then used to constrain the subsequent image generation process. As shown in Fig. 1, random noise and text information are used as the original inputs to determine the generation of the images, and then the real images and fake (synthesized) images will be identified in the discriminator. Based on the discriminated results, the generator then creates a more realistic image again, and the discriminator discriminates the new results once again. This process iterates until the net-

Wen-li Chen and Ya-qi Sun
Hengyang Normal University, School of Computer Science and Technology, Hengyang, China.

Paul L. Rosin and Yu-kun Lai
School of Computer Science & Informatics, Cardiff University, UK

work converges. The generator and discriminator are updated based on the adversarial loss. Up to now, T2I models using GANs as frameworks have been able to synthesize high-resolution images. However, further improvements are required for complex scene datasets, as the generated images still lack fine-grained detail and clarity.

In T2I works, a key research focus in T2I works is the fusion of text and image information. In the current models, three common approaches are feature concatenation, cross-modal attention [8], and Conditional Batch Normalization (CBN) [9]. Feature concatenation is a linear transformation of text information into the feature vector size required for image information, followed by a simple concatenation operation of the two. Cross-modal attention achieves the mapping of a word to a sub-region in an image through a fusion approach based on an attention mechanism. Thus different regions of the image are generated. CBN is a specialized scaling and shifting operation applied to the general feature map, enhancing its visual semantic embedding.The current methods still cannot deeply and effectively fuse textual information with image information, producing a situation where the generated image does not match the textual information.

To study the above problems, we propose a new approach to text-to-image synthesis. In general, we make the following contributions:

(1) We propose a new framework called SDeep. We use it to accomplish the transformation of text and image information and generate high-quality realistic images.

(2) We propose the Stakced Deepening Block(SD Block) for deepening text and image feature fusion. This block makes the generated images more relevant to the textual descriptions and also results in a better fusion of textual information and image features.

(3) We propose Multilayer Channel Attention (MLCA), a novel channel attention mechanism, for effective channel feature learning. It is used in the generator and enhances image feature aggregation to achieve images with richer detail.

The structure of this article is as follows. In Section 2, the work related to the Text to Image task is presented. In Section 3, the general framework of the proposed model is introduced. In Sections 4 and 5, the relevant features and details of the proposed method are clearly defined as follows. In Section 6, experiments on various aspects of the proposed method are conducted. Finally, our conclusions are given in Section 7.

2 Related Work

Text to Image. In T2I works, the main objective is to transform the descriptive text into a realistic image that matches its semantics. This process uses natural language processing and image synthesis techniques to first extract text features containing important details from textual descriptions, and then transform the text features into visual image pixels. The final image information obtained should be authentic in nature. The current methods are Variational Auto-Encoder (VAE) [10], Deep Recurrent Attention Writer (DRAW), Generative Adversarial Network (GAN) [11, 12], Diffusion Model [13,14], Cross-modal pre-training model based on contrast picture-text: CLIP [15], etc. VAE is a variant of the self-encoder that can be used to learn data generation distributions as well as random samples from the latent space to generate similar images containing the features of the training network. DRAW utilizes recurrent neural networks to generate images by focusing on sequences of regions of visual attention based on a variational autoencoder. The diffusion model is an emerging network model that starts with an image composed entirely of noise, predicts the noise filtered out at each step, and iteratively denoises to obtain a high-quality image sample. The contribution of CLIP [15] is the ability to generate corresponding text descriptions based on image content or to use text descriptions to match images without having to tune the model to the new category.

Text to Image based on Generative Adversarial Network. As a powerful deep learning model, a GAN generates a complete set of real data by learning the data distribution of the original real sample set. In 2016, Reed et al. first applied a GAN to T2I. After this, Reed et al. built on the previous work by proposing the GAWWN model [16], which improved the image resolution from 64×64 to 128×128. The ability of GANs to generate specific and attractive images has captured the attention of many researchers. And the field of T2I has also gained many development opportunities. In 2017, PPGAN [17] generated 227×227 high-resolution images by training the GAN and then iteratively optimizing it to get a better result. The same year, Zhang et al. successively proposed StackGAN [18] and StackGAN++, which introduced a stack structure in the GAN and used multiple GAN architectures to generate 256×256 images with realistic details. The improvement also provided some basis for much subsequent work. In 2018, Xu et al. sparked attention with AttnGAN [8] using an attention mechanism that pays more attention to the details of the generation situation between text and images. In 2019, the concept of a mirror structure was also

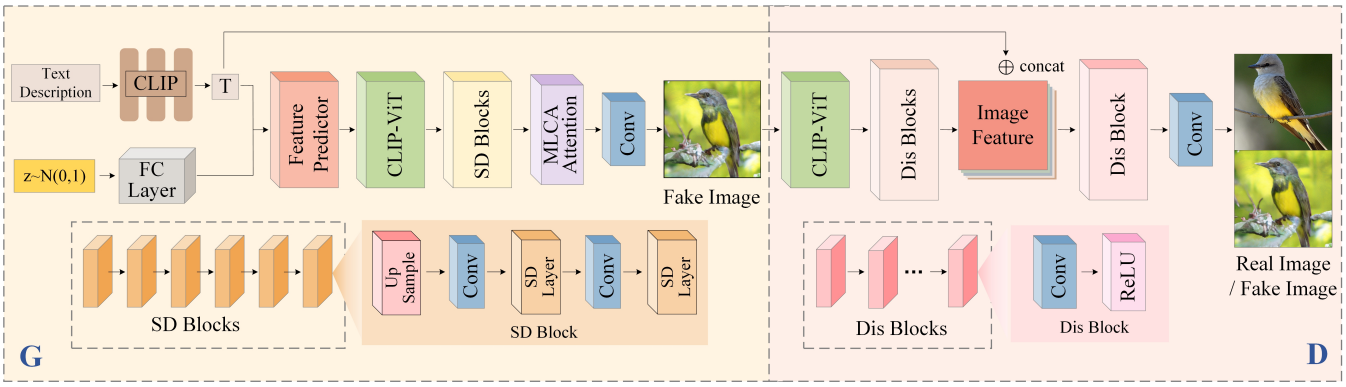

Fig. 2 The network structure of SDeep. It is a single-stage GAN structure combined with CLIP. Among them, the FC Layer is the fully connected layer used to reconstruct the dimensions of the random noise vectors. T is the sentence vectors generated by the pre-trained CLIP text encoder.

introduced into GANs. MirrorGAN [19] used a cyclic structure to propose a 'text-image-text' framework to regenerate text descriptions from the generated images, enhancing the consistency of text descriptions and visual content. In the same year, DMGAN [20] used dynamic memory networks to select text content and introduced dynamic storage modules to refine blurred image content. In 2020, CookGAN [21] investigated the synthesis from recipes to images from a causal perspective, using cooking recipe descriptions to generate food images. In 2022, DFGAN [22] used conditional affine variation to propose deep fusion generative adversarial networks to synthesize high-resolution images directly in a more efficient way. Most of the above methods continue to evolve by enhancing image resolution and quality, yet they still have certain limitations. They tend to focus on the synthesis of single objects, such as birds and flowers. For complex image synthesis tasks, the synthesized objects can easily be placed in various unreasonable positions in the image, which means the layout structure can be easily confused. The generated image scenes for complex tasks suffer from incoherence, disharmony, and discrepancies between the image content and the proposed text.

Fusion of text to image information. The initial way of processing textual and image information is feature stitching, such as GAN-INT-CLS [7], StackGAN [18], StackGAN++ [23]. They both combine textual information in the depth direction with the feature vectors obtained by convolving the original image, but this approach only achieves integration by simple manipulation and the connection between the parameters is very loose. The fusion approach based on attention mechanisms, represented by AttnGAN [8], implements a mapping of words to a sub-region in a picture at the word level. MirrorGAN [19] reintroduces global attention on the basis of AttnGAN, combining it with the local attention in the previous AttnGAN so as to focus on the details of the image and semantic generation. In DRGAN [24], both sentence features and image features go through a separate layer of self-attention to extract key information to enhance image semantic coherence. SDGAN [25] first applied CBN to image generation, embedding word-level and sentence-level CBNs in image feature maps. In addition, DFGAN [22], DTGAN [26], SSAGAN [27] and RATGAN [28] also employ CBN and CIN to integrate textual information into the synthesized image. This is accomplished by encoding the text into a vector representation and subsequently embedding it into the image feature map. The process involves scaling and shifting the visual features to enable the merging of text and image attributes. While CBN methods are effective for localized, independent fusion of textual and visual data, they have limited utility in capturing global distributions. Additionally, there exists a significant structural disparity between text and image modalities. Cross-modal image generation models are susceptible to challenges such as overfitting or instability, often resulting in irregular object shapes or poorly rendered details.

3 Model Overview

As shown in Figure 2, the proposed SDeep is a single-stage text generation image backbone. First, the text encoder processes the text information, and the generator generates images. Then, the Match-Aware Gradient Penalty [22] strategy is used to improve the discriminative ability of our model. Through iterative training, the generator and discriminator are continuously updated until the network converges.

First, the pre-trained CLIP text encoder is used to extract sentence vectors from text descriptions. The text description is encoded in the text encoder and then

fed into the generator together with noise vectors sampled from a Gaussian distribution. In the generator, images are generated using Feature Predictor [29], Clip-VIT [29], SD Blocks, MLCA Attention, and a convolutional layer. The fully connected layer resizes the noise vectors, and then processing image features is achieved by Feature Predictor, Clip-VIT, and SD Blocks, as well as the information fusion of the text vectors and the image features. Then MLCA Attention is used to learn the importance of different feature channels in the feature map. Finally, the image features are converted into an image.

The discriminator aims to be capable of discerning the authenticity of the input image. In the discriminator, the fake image is first converted into image features by the CLIP-VIT and Dis Blocks. The fake image and real image are then discriminated and the adversarial loss is generated, which drives the generator to continuously improve and generate outputs that are closer to the true image distribution. Through iterative training in this way, the discriminator's power to discriminate and classify is continuously improved, thus providing a better direction for the generator to optimize.

In SDeep, the loss function of the generator is (1).

$$\begin{aligned} L_{Gen} = & - E_{G(n,t_v)\sim P_{gen}}[D(C(G(n,t_v)),t_v)] \\ & - \lambda E_{G(n,t_v)\sim P_{gen}}[S(G(n,t_v),t_v)] \end{aligned} \tag{1}$$

where n is a noise vector sampled from a Gaussian distribution. t_v is a vector of textual information. G is the generator of SDeep. D is the discriminator and C is the CLIP-ViT in the discriminator. S denotes the cosine similarity between the encoded visual features and the text features. λ is the text image similarity coefficient. P_{gen} denotes the synthetic data distribution.

In the discriminator, the loss function is calculated by (2).

$$\begin{aligned} L_{Dis} = & - E_{x\sim P_{real}}[\min(0,-1+D(C(x),t_v))] \\ & - (1/2)\, E_{G(n,t_v)\sim P_{gen}}[\min(0,-1-D(C(G(n,t_v)),t_v))] \\ & - (1/2)\, E_{x\sim P_{\mathrm{err}}}[\min(0,-1-D(C(x),t_v))] \\ & + bE_{x\sim P_r}\left[\left(\|\nabla_{C(x)}D(C(x),t_v)\| + \|\nabla_t D(C(x),t_v)\|\right)^d\right] \end{aligned} \tag{2}$$

where P_{gen} denotes the synthetic data distribution. P_r denotes the true data distribution. P_{err} denotes the mismatched data distribution. b and d are the two parameters that balance the gradient.

The entire training process of SDeep is shown in Algorithm 1.

Algorithm 1 Training Algorithm of SDeep.

Required: input text description t; random noise n; Generator G; Discriminator D.
$n \leftarrow$ random noise, $n \sim N(0,1)$;
$t_w \leftarrow$ pretrained CLIP text encoder for t;
$n' \leftarrow$ fully connected layer(n);
// Generate initial image in G
$Img = G(n', t_w)$;
  Obtain $bridge\_fea$ in Feature Predictor;
  $vis\_c \leftarrow$ CLIP-ViT-G$(n', t_w, bridge_fea)$;
  for each SD Block i in 1 to 6 **do**
    $Fea\_sd \leftarrow$ UpSample$(n', t_w, vis_c, bridge_fea)$;
    $Fea\_sd \leftarrow$ Conv(Fea_sd);
    $Fea\_sd \leftarrow$ SD Layer(n', t_w, Fea_sd);
  end for
 $F\_attn \leftarrow$ MLCA Attention(Fea_sd);
 $generated\_image \leftarrow$ Conv2D(F_attn);
// Evaluate loss from D
$Loss = D(t_w, generated\_image)$;
  $feature\_e \leftarrow$ CLIP-ViT-D$(generated_image)$;
  for each Dis Block j in 1 to 6 **do**
    $Fea\_dis \leftarrow$ Conv$(t_w, feature_e)$
    $Fea\_dis \leftarrow$ ReLU(Fea_dis)
  end for
  $Total\_Loss \leftarrow loss_g(n, t_w) + loss_d(Fea\_dis, t_w)$
Output $generated\_image, Total\_Loss$

4 SD Block

4.1 Staked Deepening Block

Many researchers have proposed valuable approaches [30] to address the problem of better fusion of features from different perspectives [31]. Initially, text and image information was processed by direct feature concatenation, and the fusion effect became increasingly effective with the development of approaches based on attention mechanisms. Later, this further improved with the use of CBN and CIN to achieve the fusion of text information into image generation. In the SDeep we propose the SD Block, which is an extension of the use of text fusion blocks to achieve text-to-image fusion using the affine transformation method. Compared to previous fusion approaches, SD Block can further deepen the fusion process for T2I.

As shown in Figure 3, there are six SD Blocks in the generator, each of which is made up of an upsam-

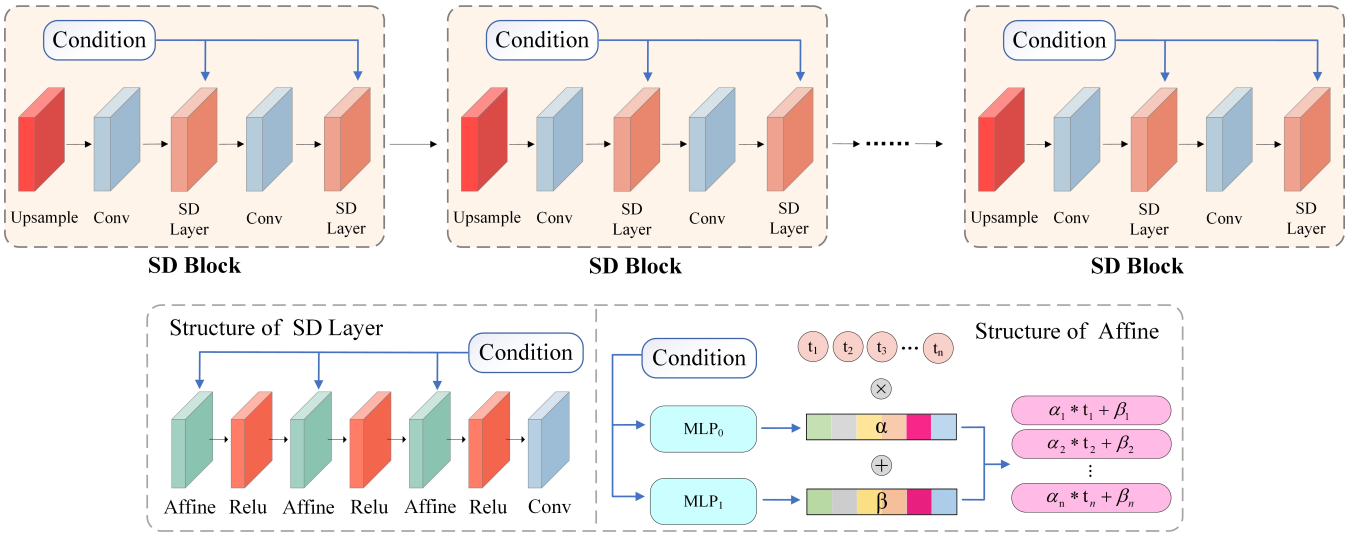

Fig. 3 There are six SD Blocks in the generator. An SD Block contains two SD Layers. An SD Layer consists of three affine transformation layers and ReLU layers stacked alternately, followed by a convolution layer.

pling, two convolution layers, and two SD Layers. Each SD Layer utilizes the Affine Transform layer, ReLU layer, and convolution layer, with a total of three Affine Transform and ReLU layers stacked. We accomplish the fusion of textual information to images by means of superimposed affine transformations.

$$\alpha = MLP_1(t) = \phi(W_1 t + b_1) \tag{3}$$

$$\beta = MLP_2(t) == \phi(W_2 t + b_2) \tag{4}$$

4.2 The introduction of nonlinear transformations

In particular, for the Affine Transformation layer, two Multilayer Perceptrons (MLPs) are used to complete the key processing steps. MLPs learn and generate scaling parameters α and the shifting parameters β for language conditional channels from input text information, which are used to guide the transformation of image features. These two parameters are represented as Eq.(3) and Eq.(4) respectively, used to characterize the calculation process of scaling and shifting. Specifically, when inputting the feature map X, the scaling parameter α is first used to scale each channel of the feature map X to adjust the importance of the channels. Subsequently, the shift parameter β is used to perform a translation operation on each channel of the feature map X to further enrich the expressive power of the features. This series of processing procedures is represented by Eq.(5). By combining text information with image information in the above way, text information can directly guide the representation learning of image features. This not only enhances the diversity results of visual features but also dynamically expands the representation space based on different text descriptions by adjusting scaling and shifting parameters, effectively capturing different visual features.

$$Affine(x_i|t) = \alpha_i \cdot x_i + \beta_i \tag{5}$$

where t is the textual information, α is the scale factor and β is the shifting factor. Affine denotes the affine transform, and x_i is the i-th channel information of the feature map.

Affine transformation is essentially a linear operation, and although it can change the distribution of input, its representational power is limited. Adding a ReLU layer (non-linear activation function) after affine transformation can introduce non-linear characteristics, thus compensating for the shortcomings of simple linear transformation. This allows for more flexible learning of complex feature relationships when fusing text and image information, reducing the constraints of linear limitations on the fusion effect. Using only a single affine transformation layer significantly restricts the generator's conditional representation space, limiting its ability to map text descriptions to image representations. This constrained representation space is inadequate for distinguishing diverse text descriptions, and therefore cannot generate images with significant differences. Stacking three layers of affine transformations and alternately adding ReLU layers in the middle can significantly expand the conditional representation space. This enhances the adaptability of the generator

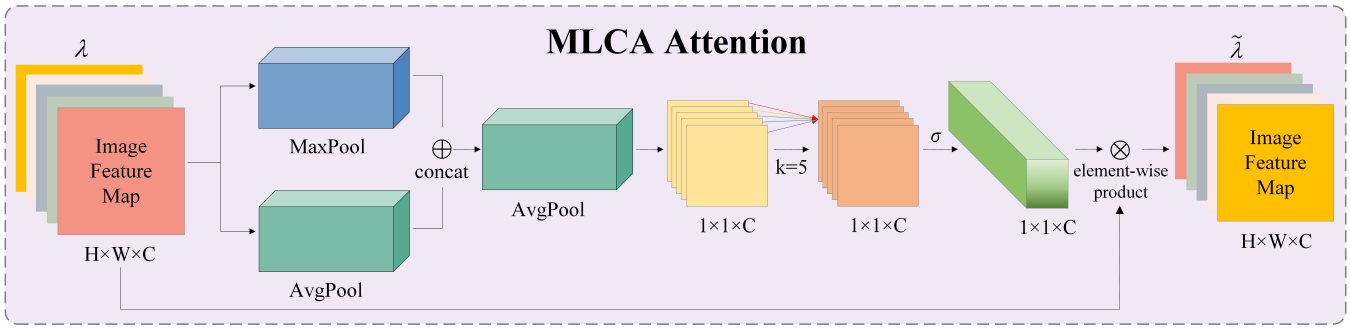

Fig. 4 The architecture of MLCA Attention. This module uses the channel attention mechanism to learn the importance of each feature channel of the feature map, suppressing ineffective information and making the generated images more detailed.

to complex text descriptions, enabling it to generate more diverse and semantically relevant images. Meanwhile, this also deepens the fusion process of text and image features. Deep fusion helps capture finer semantic features, thus improving the quality and authenticity of the final image generation.

5 MLCA Attention

5.1 Channel feature learning for avoiding dimensionality reduction

In previous works on the text-to-image task, many approaches used a stack structure to ensure image quality and resolution, using multiple generators and discriminators divided into multiple stages to generate images. However, the stacking structure partially depends on the image generated in the first stage. In other words, when poor images are generated in the initial stage, the subsequent refinement stages struggle to function effectively. This often leads to a final output that resembles a mere patchwork of disconnected fragments rather than a cohesive image. While discarding the stacking structure, it is important to investigate how to improve the sharpness and fine-grained detail of the resulting images.

In SDeep, we propose MLCA Attention, specifically designed as a channel attention approach. The structure is illustrated in Figure 4. In MLCA Attention, the process begins by taking the feature map of size $H \times W \times C$, followed by applying both global maximum pooling(GMP) [32] and global average pooling(GAP) [33] to produce two distinct feature maps. Then they are concatenated based on channel dimension to increase the number of channels. We obtain a feature map with more feature representations. The resulting feature maps are then subjected to spatial feature compression after GAP to produce feature maps of size $1 \times 1 \times C$. Next, the compressed feature map undergoes channel-wise learning, where a 1×1 convolution is applied to capture the importance of each channel. This results in a $1 \times 1 \times C$ output feature map. Finally, the channel attention feature map is combined with the original input feature map, and element-wise multiplication is performed across the channels. The final output is a feature map enhanced with channel attention.

5.2 Use of Dynamic Convolutional Kernel

In the previous attention mechanism, global learning occurs when the input channel feature maps are processed through a fully connected layer. When 1×1 convolution is used, the information learned is limited and only the information between the local channels can be acquired. After considering the fact that different input feature maps will extract different ranges of features, and also in the operation of convolution, the sensory field is affected by the size of the convolution kernel. MLCA Attention leverages a dynamic convolution kernel to perform 1×1 convolution, enabling the model to learn the relative importance of different channels. The dynamic convolution kernel adjusts its size adaptively based on a predefined function. For layers with a higher number of channels, the convolution kernel becomes larger, facilitating enhanced cross-channel interactions during the 1×1 convolution process. In layers with a smaller number of channels, a smaller convolution kernel is used to do 1×1 convolution, making fewer cross-channel interactions. MLCA Attention obtains the kernel size by expanding the linear function to a nonlinear function, and since the number of channels usually is set to the power of 2, there is the following formula:

$$C = \phi(k) = 2^{(\gamma * k - b)} \tag{6}$$

where C denotes the channel dimension and ϕ is the mapping relation between k and C expressed as a nonlinear function.

The convolution kernel adaptive function is as follows:

$$k_s = \psi(C) = \left| \frac{\log_2(C)}{\eta} + \frac{d}{\eta} \right|_{odd} \tag{7}$$

where k_s denotes the convolutional kernel size, C denotes the number of channels, $|\cdot|_{odd}$ means k_s takes an odd number, and η and d are set to 2 and 1.

Compared with SE-Net [34], we use 1×1 convolution to learn the channel information, avoiding the loss of a part of the feature expression caused by dimensionality reduction. Compared with ECA-Net [35], the operations of GMP Eq.(8) and GAP(Eq.(9))are added based on the H and W dimensions when compressing spatial features, where $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ representing input feature maps. The joint use of the two poolings can improve the representation of the network. In contrast, we utilize MLCA Attention to focus more on local features compared to the way that focuses on global features. In the process of generating images, MLCA Attention acquires feature maps with more feature representations. It is able to perform channel feature enhancement on the input feature maps and does not change the size of the input feature maps. It mainly uses the channel attention mechanism to learn the importance of each feature channel of the feature map, captures the information between different channels, and highlights the significant feature channels while suppressing the ineffective information. This makes the details of the generated image richer and improves the quality of the generated images.

$$y_c = \max_{i=1,\ldots,H} \max_{j=1,\ldots,W} F_[c,i,j] \tag{8}$$

where y_c denotes the maximum value of the c-th channel output after the pooling operation. $F_[c,i,j]$ denotes the value of the c-th channel in the input feature map at position (i,j). H and W denote the height and width of the feature map.

$$y_c = \frac{1}{H\times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_[c,i,j] \tag{9}$$

where y_c denotes the average value of the c-th channel of the output. $F_[c,i,j]$ denotes the value of the c-th channel in the input feature map at position (i,j). H denotes the height of the input feature map. W denotes the width of the input feature map.

6 Experiments

In this section, we will introduce the datasets, training details, and evaluation metrics used in the experiments, and perform quantitative evaluation, qualitative evaluation, and ablation study on the proposed model, respectively.

Datasets. To evaluate our method, we used two challenging datasets for our experiments, CUB [36] and COCO [37]. The CUB bird dataset [36] contains 200 bird subclasses and 312 attribute descriptions, with 11788 bird images, each of which corresponds to 10 linguistic descriptions. The dataset is divided into 8,855 training images and 2,933 test images. The COCO dataset [37] consists of more than 80,000 training images and more than 40,000 test images, with a total of 80 annotated classes, each of which corresponds to five linguistic descriptions.

Training details. In this study, we use pytorch to build the framework of our model. We use a pre-trained model based on CLIP as a text encoder to extract semantically rich text features and ensure high-quality expression of text features. In the generator, the random noise with an input dimension of 100 is a vector sampled from a standard Gaussian distribution, used to initialize the generation process and introduce diversity. In addition, our network optimization uses the Adam optimizer, where β_1 and β_2 are set to 0.0 and 0.9 respectively to ensure model stability and convergence. During training, all models are experimented with a Nvidia RTX 4090 GPU (24gb memory). The batch size is set to 32, the dimensionality of the noise vector is set to 100, and the size of the generated images is set to 256×256. The models are trained on the CUB and COCO datasets until the models converge. The maximum training epoch on the CUB dataset is 1300 and the training time is about 1.5 days. The maximum training epoch on the COCO dataset is 1500 and the training duration is about 16.5 days.

Evaluation indicators. We employ the widely used Fréchet Inception Distance (FID) [43] and CLIP-SIM (CS) [44] to evaluate the performance of the model. FID relates to the problem of comparing the distribution between real images and generated images, evaluating the quality of the generated image. It is used to measure the distance between the distribution of real images and the distribution of generated images. The covariance is used to take the diversity into account when computing the distance. When the score of FID is lower, it shows that the two distributions are more similar and the quality of the generated image is higher. CLIPSIM is a method for measuring the relevance of images and text. It consists of a CLIP model

Table 1 Comparison of metrics with alternatives on the CUB and COCO. The best results are shown by bolding the data in black.

Model	FID↓		CS↑	
	CUB	COCO	CUB	COCO
DMGAN [20]	16.09	32.64	–	–
DAEGAN [38]	15.19	28.12	–	–
XMC-GAN [39]	–	9.33	–	–
DFGAN [22]	14.81	19.32	0.2920	0.2972
DR-GAN [22]	14.96	27.80	–	–
SSAGAN [27]	15.61	19.37	–	–
LAFITE [40]	**10.48**	8.12	0.3125	**0.3335**
ALRGAN [41]	15.14	29.04	–	–
GigaGAN [42]	–	9.09	–	–
Ours	10.49	**7.84**	**0.3188**	0.3297

that extracts image and text features and computes the semantic similarity between them. This metric is typically used for text-conditioned generation or editing tasks.

6.1 Quantitative Evaluation

We compared our method with various current state-of-the-art methods, including DMGAN [20], DAE-GAN [38], XMC-GAN [39], DFGAN [22], DR-GAN [22], SSAGAN [27], LAFITE [40], ALRGAN [41], GigaGAN [42]. As shown in Table 1, our method was compared with other methods on the CUB and COCO datasets. The FID scores indicate that the images generated by our method on both datasets demonstrate excellent authenticity. In addition, our method has also achieved significant advantages in the semantic relevance index CS score. Especially on the CUB, our method achieved a performance of 0.3188, significantly higher than the other methods. This indicates that our method not only generates high-quality images, but also better preserves semantic consistency between text and images.

From the computational perspective, we compare the number of parameters between the SDeep and diffusion models and the autoregressive model. As shown in Table 2, the results are derived from relevant paper data and replication validation. Based on a large number of predictions from autoregressive models and diffusion, they require a much larger GPU cluster for retraining and development. Compared to them, SDeep is lower than the number of parameters by one or two orders of magnitude. However, SDeep's FID score is very close to them. Therefore, SDeep has a better learning representation with fewer parameters, which will be more user-friendly and flexible.

Table 2 Comparison of our method with autoregressive models and diffusion based models. Displayed the number of parameters and FID score based on the COCO.

Model	Type	FID↓	params(B)↓
DALL-E [45]	AR	27.50	12
CogView [46]	AR	27.10	4
CogView2 [47]	AR	24.00	6
Parti [48]	AR	7.23	20
GLIDE [49]	DM	12.24	3.5
DALL-E2 [50]	DM	10.39	3.5
Imagen [51]	DM	7.27	3.4
SDeep	GAN	7.84	**0.18**

6.2 Qualitative Evaluation

For qualitative evaluation, we conducted experiments with five models, AttnGAN [8], DMGAN [20], SSAGAN [27], DFGAN [22] and GALIP [29], and presented multiple sets of experimental results for comparison based on visualization results.

The experimental results of the COCO are shown in Figure 5. In the nine sets of experiments presented, the results of AttnGAN, DMGAN, and SSAGAN are similar to the collocation of some simple objects. The image subject is not clear, and even the content to be expressed cannot be recognized, nor can the content of the text description be effectively expressed. DFGAN and GALIP, compared to these three methods, can express the image scene more effectively and embody the main object of the image. They all have their problems with the presentation of the image contents. The generated image has only a rough frame outline, and the specific content is ambiguous. Our proposed method, however, has significantly improved results, which expresses the image subject more effectively and is more appropriate to the text description, such as 'white tile' in column (1), 'lake' in column (4), 'street sign' in column (6), 'cows' in column (7), and 'boat' in column (9). All in all, compared to other methods, our method generates images with better textures, more relevant colors, less distorted shapes, and more relevant content. Moreover, from a visual perception perspective, the generated im-

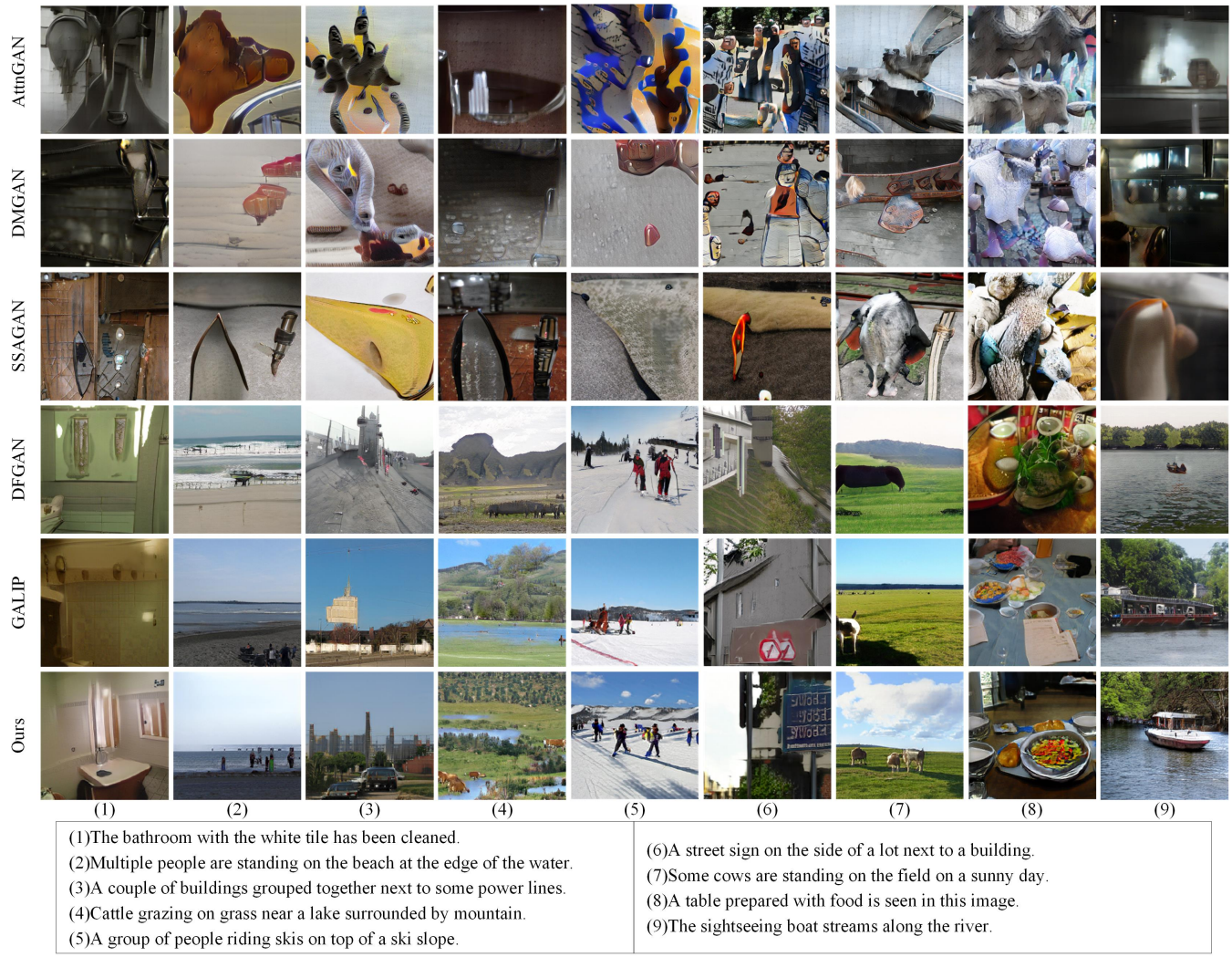

Fig. 5 Qualitative comparison between our method and AttnGAN [8], DMGAN [20], SSAGAN [27], DFGAN [22] and GALIP [29] on some of the test set of the COCO dataset. The text description and the generated image correspond to the same numeric label. The text descriptions corresponding to different numeric labels are given in the boxes below, and the corresponding images generated by the same text description using different methods are shown in the same column.

ages exhibit a well-balanced color distribution, greater overall coherence, and enhanced realism.

Figure 6 shows the experimental results of the CUB dataset. It can be seen that the experimental results of AttnGAN, DMGAN, and SSAGAN contain the subject-object of 'bird', but there are differences in the text descriptions. In the text descriptions corresponding to the experimental results, 'dark brown', and 'speckled' in column (1), 'brown with black' in column (2), 'short', 'stout', and 'curved' in column (3), 'black bill' in column (4), 'blue and white' in column (5), 'white with grey' in column (6) and (7), 'white belly' in column (8), 'red with black' in column (9), and none of them are shown. However, our method can significantly solve these problems, and the relevant text descriptions are present in the resultant images. Although they are also able to be represented in DFGAN and GALIP, the quality of ours is clearly superior to DFGAN and GALIP, The results generated by our method have clear details such as feathers, beak, eyes, and feet. In addition, the background is also more regular in our results, especially the tree branches, and their textures are also naturally generated. While presenting the bird as a subject-object, it can reflect the relevant specific details in the textual description and the presented images are superior.

In Figure 7, the experimental results of our method and the large model methods (Stable diffusion [13] and Genmo) are shown. On the whole, it is obvious that the images generated by Stable diffusion [13] and Genmo lack realism. In the large model methods, such as the images in the first, third, fourth, and fifth columns, the

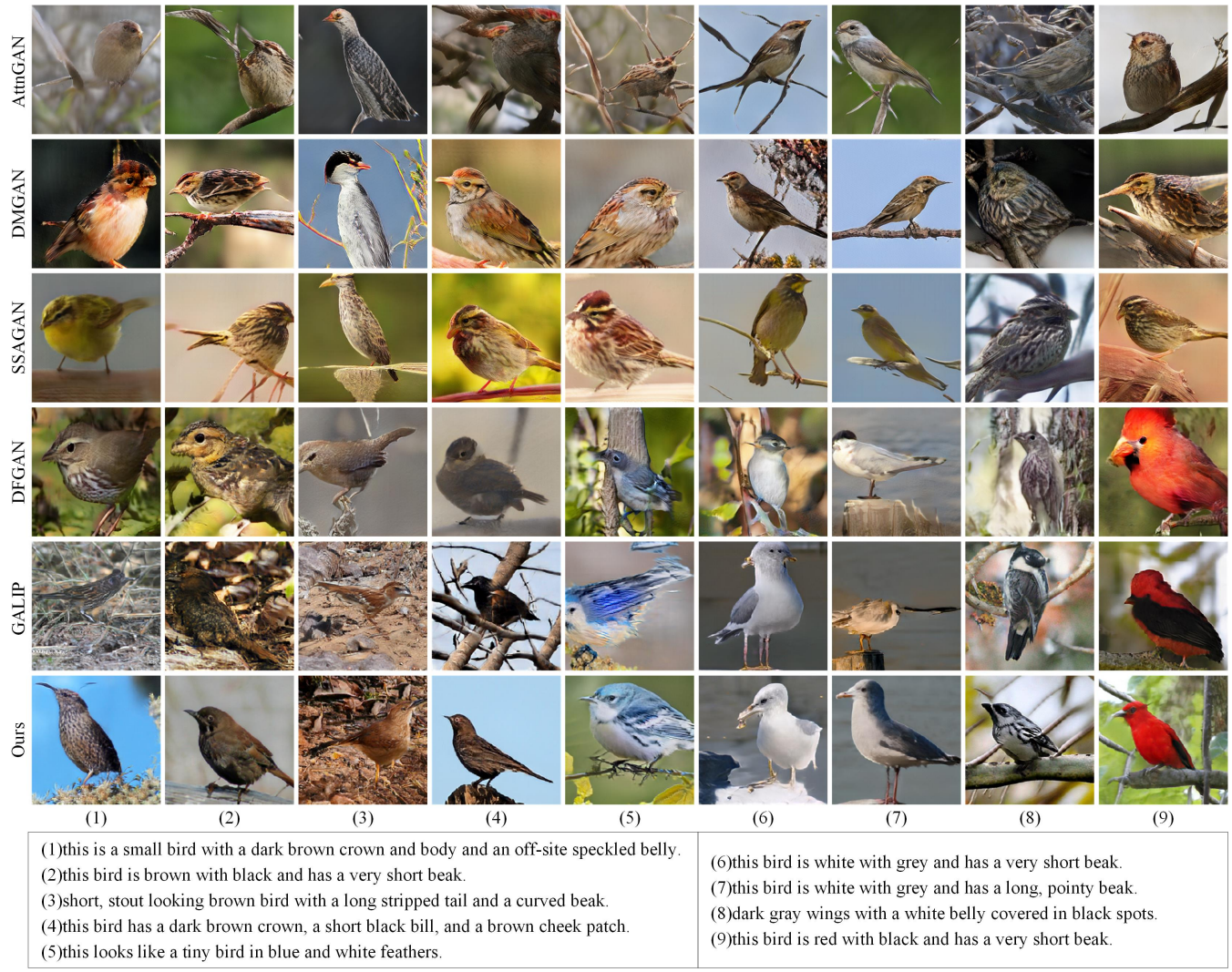

Fig. 6 Qualitative comparison between our method and AttnGAN [8], DMGAN [20], SSAGAN [27], DFGAN [22] and GALIP [29] on the test set of CUB bird dataset. The text description and the generated image correspond to the same numeric label. The text descriptions corresponding to different numeric labels are given in the boxes below, and the corresponding images generated by the same text description using different methods are shown in the same column.

backgrounds of the images they generate appear very rough in the virtualization, and the artistic sense is too strong to lose authenticity. In addition, the generated images also have the problem of semantic inconsistency and error. For example, the images in the second and third columns fail to generate 'orange bill', and even generate multiple beaks. The 'lake' in the fifth column image is also not well presented. In our approach, the generated images are able to contain the main elements and are more in line with the text description. For example, 'red with black' in the first column, 'orange bill' in the second and third columns, 'people' in the fourth column, 'beach' in the fourth column, and 'cattle' and 'lake' in the fifth column. In addition, the images generated by our method are also more realistic than the large model.

Table 3 Ablation studies of each component on the CUB and COCO dataset. ✓indicates that the component is used in the experiment. We compare the FID scores for each experiment as follows.

Architecture	exp0	exp1	exp2	exp3
baseline	✓	✓	✓	✓
SD Block		✓		✓
MLCA Attention			✓	✓
CUB-FID↓	11.33	11.14	11.24	**10.49**
COCO-FID↓	8.12	7.98	8.03	**7.84**

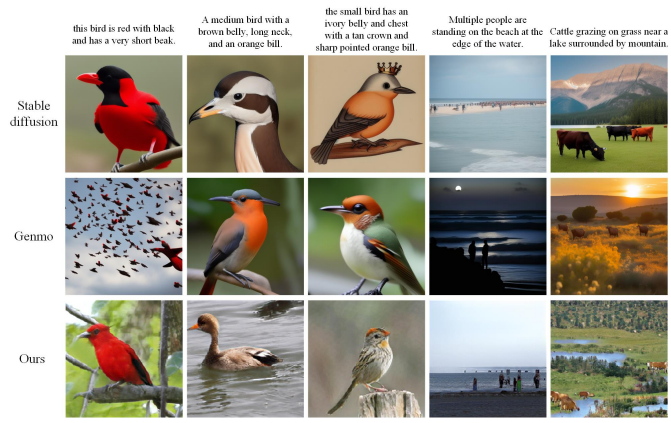

Fig. 7 Results of qualitative experimental comparisons with the large model methods Stable diffusion, and Genmo on the COCO and CUB datasets. The top side shows the input text descriptions. The left side shows the model methods. The generated experimental images are shown on the right side of the model method.

6.3 Ablation Study

We conducted ablation experiments on the CUB [36] and COCO dataset [37] to evaluate the performance of each component in our approach, and to show experimental comparisons under different component settings. The components tested include SD Block and MLCA Attention. The FID for each experiment demonstrates their effectiveness compared to the baseline.

Based on the baseline method, we added SD Block and MLCA Attention. The results of the ablation experiments are shown in Table 3. On the CUB, the FID score for the baseline is only 11.33. In exp1, with the addition of SD Block in the baseline, the FID score drops to 11.14. In exp2, when we add MLCA Attention, the FID score drops to 11.24. On the COCO, the FID score for the baseline is only 8.12. In exp1, with the addition of SD Block in the baseline, the FID score drops to 7.98. In exp2, when we add MLCA Attention, the FID score drops to 8.03. These two experiments demonstrated the effectiveness of the two components respectively. Finally, we experimented with all the components together. The FID score drops to 10.49 on the CUB and 7.84 on the COCO.

The results of the ablation experiments demonstrate the superior performance of SDeep, with each component contributing effectively to the fusion of text and image information, thereby improving image quality. Overall, we also compared the visualization results of these four sets of experiments for the visualization results on the CUB and COCO, and the experimental results are displayed in Figure 8.

To demonstrate the individual validity of the two components, we show the experimental results for both components. The SD Block is implemented using cross-stacking of three affine transform layers and ReLU layers. Their stacking allows the generator to hierarchically learn different levels of abstract features of the data for a more textually relevant image representation. Moreover, by introducing nonlinear transformations, the model can adapt more flexibly to the higher-order features and complexity of the data. When there is only one layer of affine transformation, the conditional representation space of the generator is very limited. By appropriately multiplying the layer count, the performance can be improved while avoiding the introduction of excessive complexity. We experimentally compare the effect of superimposing two, three, and four layers, and the experimental results are displayed in Table 4. According to the experimental results, both FID and CS achieved the best results when we stacked the number of layers as 3. With the consideration of too many layers leading to overfitting and computational parameters, we determined the number of stacked layers to be 3.

Table 4 Ablation study of the number of layers stacked inside the SD Block. Comparison of FID and CS scores for stacking two, three, and four layers, respectively.

Architecture	2 Layers	3 Layers	4 Layers
CUB-FID↓	11.75	**10.49**	13.47
CUB-CS↑	0.3127	**0.3188**	0.3117

For the component MLCA Attention, we added visualization results for comparison. As shown in Figure 9, the quality of the first row of images is poor without MLCA Attention. All seven images appear to overemphasize the background and ignore the details of the subject. And the relevant parts of the bird are seriously missing or distorted. The second row is the result of the experiment with the addition of MLCA Attention. In comparison, all seven images were able to maintain good visual effects. The main body and specific parts of the bird are relatively intact, especially the beak, eyes, and tail. In addition, the image as a whole is also very natural, and the generation of backgrounds like tree branches and rocks is also coordinated.

6.4 Limitations

Although our experimental results are competitive, there are still limitations in some cases. Firstly, our visual results are still not comparable to large models such as stable diffusion [13]. Our method can also only generate realistic images in the two current datasets and

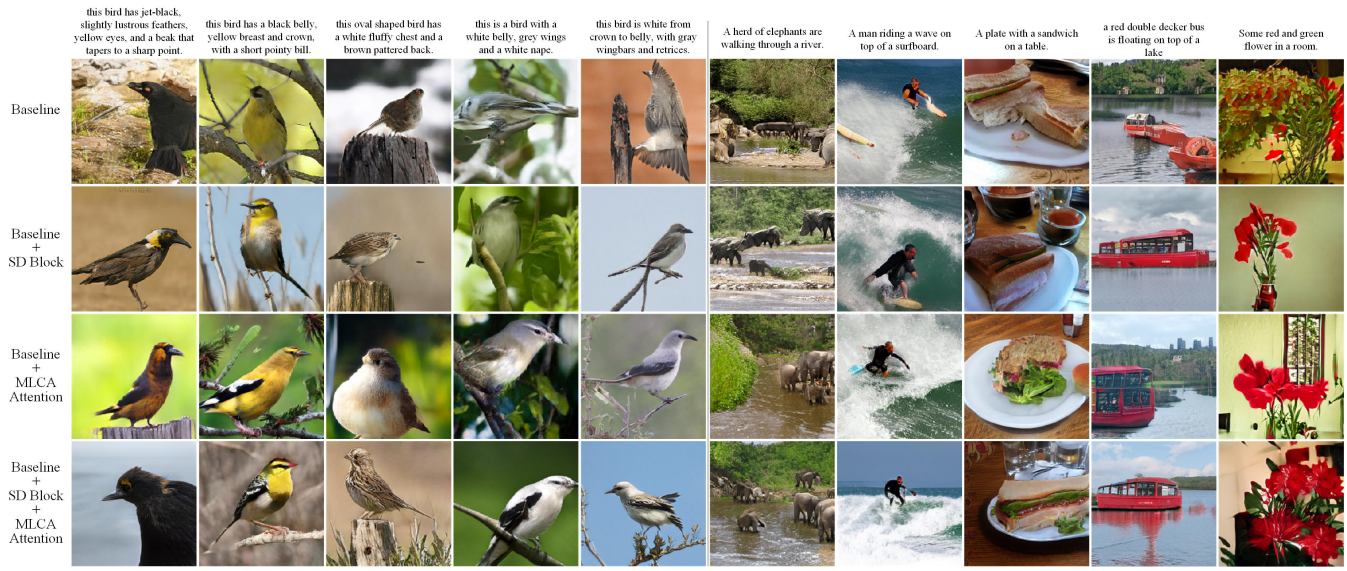

Fig. 8 Comparison of visualization results for ablation experiments on the CUB and COCO datasets. The top shows the input text descriptions. The left side shows the component composition of each experiment. Each column shows the experiment results generated by different models for the same text description.

Fig. 9 Comparison of visualization results for MLCA Attention ablation experiments. The first row shows the experimental results generated by the model without the MLCA Attention component. The second row is the experimental results generated by the model with the MLCA Attention component.

cannot creatively generate stylized high-quality images. Secondly, we also need to address the issue of utilizing compressed models to improve model performance. Although it is possible to extend the model by adding additional conditions, the model will become larger and require more computational power than the ordinary user can afford. Thirdly, we also need to continue to investigate the effective fusion of cross-modal information. Although the fusion of the current model works well, it is not yet able to be applied to more modal information. In our subsequent work, we will extend the model to dictionary learning [52] and try efficient feature representation to achieve better feature fusion and representation between textual descriptions and image information.

7 Conclusion and Future Work

In this paper, we introduce a GAN-based network architecture named SDeep. Our primary objective is to generate more detailed and realistic images while ensuring a strong semantic alignment between text and images. To achieve this, we address the challenge of efficiently integrating textual information into the image synthesis pipeline. First, we propose MLCA Attention, a channel attention mechanism designed to enhance feature aggregation, emphasize important feature channels, and facilitate the integration of text features with image data. Furthermore, we present the SD Block, which strengthens the fusion of text and image features, enabling a deeper integration of textual information. By

combining these strategies, we thoroughly evaluate our method on the challenging CUB and COCO datasets, comparing it with existing techniques. Experimental results demonstrate that our approach delivers substantial performance improvements.

Our proposed method builds upon the application of GANs in text-to-image generation tasks, further advancing their use by extending the capabilities to support more efficient and effective text-to-image synthesis. The experimental results we present confirm both the feasibility and the practicality of our approach. In future work, we hope that our models can be improved much more in terms of performance. As more and more powerful models become available, the generation results have been drastically improved and extended. However, the larger the model, the higher the computational requirements are required. Generating images faster and better under limited resource conditions is a challenging problem. At the same time, converting textual descriptive content into images more aptly needs further research. Effective fusion of image information and text information is also one of the key issues that constantly needs to be overcome. In addition, we are also concerned with applying attention to 3D scenes [53], which our approach is not yet able to extend to.

Declaration of competing 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.

Data availability

The datasets used in this paper are all public datasets, which can be found in the links we provide.

Author information

All authors had equal contributions and all authors reviewed the manuscript.

References

1. Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.
2. Jianping Gou, Xiabin Zhou, Lan Du, Yibing Zhan, Wu Chen, and Zhang Yi. Difference-aware distillation for semantic segmentation. *IEEE Transactions on Multimedia*, 2024.
3. Haojie Zhang, Yongyi Su, Xun Xu, and Kui Jia. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23385–23395, 2024.
4. Mei Zhang, Lingling Liu, Yongtao Pei, Guojing Xie, and Jinghua Wen. Semantic segmentation of multi-scale remote sensing images with contextual feature enhancement. *The Visual Computer*, pages 1–15, 2024.
5. Jianping Gou, Liyuan Sun, Baosheng Yu, Shaohua Wan, Weihua Ou, and Zhang Yi. Multilevel attention-based sample correlations for knowledge distillation. *IEEE Transactions on Industrial Informatics*, 19(5):7099–7109, 2022.
6. Jianping Gou, Liyuan Sun, Baosheng Yu, Shaohua Wan, and Dacheng Tao. Hierarchical multi-attention transfer for knowledge distillation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(2):1–20, 2023.
7. Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.
8. Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.
9. Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. *Advances in neural information processing systems*, 30, 2017.
10. Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.
11. Shouming Hou, Ziying Li, Kuikui Wu, Yinggang Zhao, and Hui Li. Masked cross-attention and multi-head channel attention guiding single-stage generative adversarial networks for text-to-image generation. *The Visual Computer*, pages 1–13, 2024.
12. Ming Tao, Songsong Wu, Xiaofeng Zhang, and Cailing Wang. DCFGAN: Dynamic convolutional fusion generative adversarial network for text-to-image synthesis. In *2020 IEEE International Conference on Information Technology,Big Data and Artificial Intelligence (ICIBA)*, volume 1, pages 1250–1254, 2020.
13. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
14. Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022.
15. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural lan-

guage supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
16. Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. *Advances in neural information processing systems*, 29, 2016.
17. Anh Nguyen, Jason Yosinski, Yoshua Bengio, Alexey Dosovitskiy, and Jeff Clune. Plug and play generative networks: Conditional iterative generation of images in latent space. *Retina-Vitreus*, 21(3):166–177, 2016.
18. Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.
19. Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. MirrorGAN: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019.
20. Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019.
21. Bin Zhu and Chong-Wah Ngo. CookGAN: Causality based text-to-image synthesis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5518–5526, 2020.
22. Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16515–16525, 2022.
23. Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.
24. Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li. Dr-gan: Distribution regularization for text-to-image generation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10309–10323, 2022.
25. Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2327–2336, 2019.
26. Zhenxing Zhang and Lambert Schomaker. DTGAN: Dual attention generative adversarial networks for text-to-image generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.
27. Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18187–18196, 2022.
28. Senmao Ye, Huan Wang, Mingkui Tan, and Fei Liu. Recurrent affine transformation for text-to-image synthesis. *IEEE Transactions on Multimedia*, 26:462–473, 2023.
29. Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14214–14223, 2023.
30. Qing Tian, Yanan Zhu, Heyang Sun, Songcan Chen, and Hujun Yin. Unsupervised domain adaptation through dynamically aligning both the feature and label spaces. *IEEE Transactions on circuits and systems for video technology*, 32(12):8562–8573, 2022.
31. Qing Tian, Heyang Sun, Chuang Ma, Meng Cao, Yi Chu, and Songcan Chen. Heterogeneous domain adaptation with structure and classification space alignment. *IEEE Transactions on Cybernetics*, 52(10):10328–10338, 2021.
32. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
33. Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.
34. Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
35. Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.
36. Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
37. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
38. Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. DAE-GAN: Dynamic aspect-aware GAN for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13960–13969, 2021.
39. Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021.
40. Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022.
41. Hongchen Tan, Baocai Yin, Kun Wei, Xiuping Liu, and Xin Li. Alr-gan: Adaptive layout refinement for text-to-image synthesis. *IEEE Transactions on Multimedia*, 25:8620–8631, 2023.
42. Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023.
43. Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash

equilibrium. *Advances in neural information processing systems*, 30, 2017.
44. Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer, 2022.
45. Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.
46. Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.
47. Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022.
48. Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.
49. Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
50. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
51. Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
52. Qing Tian, Chuang Ma, Meng Cao, Jun Wan, Zhen Lei, and Songcan Chen. Unsupervised multitarget domain adaptation with dictionary-bridged knowledge exploitation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):3464–3477, 2022.
53. Bin Shen, Li Li, Xinrong Hu, Shengyi Guo, Jin Huang, and Zhiyao Liang. Point cloud upsampling generative adversarial network based on residual multi-scale off-set attention. *Virtual Reality & Intelligent Hardware*, 5(1):81–91, 2023.

Wenli Chen is now a postgraduate student at the School of College of Computer Science and Technology, Hengyang Normal University. Her main research interests include deep learning and image processing, and she is learning about Artificial Intelligence Generated Content.

Yaqi Sun is currently pursuing a doctoral degree in Software Engineering at Guangxi Normal University. She received her M.S. degree in Control Engineering from Guilin University of Technology, China, in 2013. Her current research interests include machine learning and image processing.

Paul L. Rosin is currently a professor with the School of Computer Science and Informatics, Cardiff University, U.K. Previous posts include lecturer with the Department of Information Systems and Computing, Brunel University London, U.K., research scientist with the Institute for Remote Sensing Applications, Joint Research Centre, Ispra, Italy, and lecturer with the Curtin University of Technology, Perth, Australia. His research interests include low level image processing, performance evaluation, shape analysis, facial analysis, cellular automata, non-photorealistic rendering, and cultural heritage. For more information, please visit `http://users.cs.cf.ac.uk/Paul.Rosin/`

Yu-Kun Lai received his bachelor's degree and PhD degree in computer science from Tsinghua University in 2003 and 2008, respectively. He is currently a Professor in the School of Computer Science & Informatics, Cardiff University. His research interests include computer graphics, geometry processing, image processing and computer vision. He is on the editorial boards of *IEEE Transactions on Visualization and Computer Graphics* and *The Visual Computer*. For more information, please visit `https://users.cs.cf.ac.uk/Yukun.Lai/`.