IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 17, 2020, accepted May 27, 2020, date of publication June 9, 2020, date of current version June 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3001070

Customizable GAN: A Method for Image Synthesis of Human Controllable

ZHIQIANG ZHANG[1], (Graduate Student Member, IEEE), WENXIN YU[1], (Member, IEEE), JINJIA ZHOU[2], (Member, IEEE), XUEWEN ZHANG[1], NING JIANG[1], GANG HE[3], (Member, IEEE), AND ZHUO YANG[4]

[1]School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China
[2]School of Science and Engineering, Hosei University, Tokyo 184-0002, Japan
[3]School of Communication Engineering, Xidian University, Xi'an 710071, China
[4]School of Computing, Guangdong University of Technology, Guangzhou 510006, China

Corresponding author: Wenxin Yu (yuwenxin@swust.edu.cn)

This work was supported in part by the Sichuan Science and Technology Program under Grant 2020YFS0307, Grant 2019YFS0146, and Grant 2019YFS0155, in part by the National Natural Science Foundation of China under Grant 61907009, in part by the Science and Technology Planning Project of Guangdong Province under Grant 2019B010150002, and in part by the Natural Science Foundation of Guangdong Province under Grant 2018A030313802.

ABSTRACT In the research of computer vision, artificial controllability of image synthesis is a significant and challenging task. At present, there are two available methods. One is to utilize a simple contour to determine the shape of the synthetic object. This method has a promising effect, but it can only control the shape information of the synthetic object, but not the specific content. The other is to employ the text description to synthesize the corresponding image, which effectively controls the specific content of the synthesis, but it cannot do anything for the synthesized shape. In this paper, we propose a highly flexible and human customizable image synthesis model based on simple contour and natural language description, in which the specific content of contour and text description can be determined artificially. The contour determines basic synthetic object shape, and the natural language describes specific object content. Based on these, highly authentic and customizable images can be synthesized. The experiments are executed in the Caltech-UCSD Birds (CUB) and Oxford-102 flower datasets, and the experimental results demonstrate the effectiveness and superiority of our method. The results not only maintain the contour but also conform to the natural language description. Simultaneously, the high-quality image synthesis results, based on artificial hand-drawing contour and text description, are displayed to illustrate the high flexibility and customizability of our model.

INDEX TERMS Artificial neural networks, computer vision, image generation, image processing, text analysis, text processing.

I. INTRODUCTION

Image synthesis is always the core of research in computer vision. In recent years, with the development of deep learning technology, image synthesis has made many breakthroughs. Especially after the introduction of Generative Adversarial Networks (GAN) [1]–[4], the research of image synthesis has developed rapidly and obtained many promising results. However, the original input of GAN and its related variants is centered around the Gaussian distribution or uniform distribution noise variable, which makes the whole image synthesis process difficult to control artificially.

In order to make the image generation structure more valuable, it is necessary to provide high-level control information. The current research mainly starts from two directions: one is to control the shape of synthesis, the other is to control the content of synthesis. The main form of shape control is to enter a profile, such as a simple outline of a shoe or bag. Then the input contour is used to synthesize the image. The biggest problem of this method is that only shape information can be controlled, but not the specific details. For example, if input the contour of a package, this method cannot determine the color information of the package in the synthesis result. In the

The associate editor coordinating the review of this manuscript and approving it for publication was Inês Domingues.

related model [5], [6], the specific details (such as color) are determined by the image in the training set. If the training set has the yellow packet, it is possible to synthesize the yellow packet based on the contour of the packet. However, if there is no blue package in the training set, the model cannot synthesize the blue package. This reflects that the degree of control for this approach is limited.

The method to control the content of synthesis starts with the use of text information control. At first, conditional GAN (CGAN) [7], [8] used the category attributes of images (such as flower and bird) to control the categories of image synthesis. This method can only control the category of the composite content, but not for more specific details. For example, if the category label is a bird, the model can synthesize a bird image, but the color, size, and other information of the bird cannot be controlled.

Furthermore, Reed *et al.* [9] proposed image synthesis based on text description information (like ''this bird is black with white and has a long, pointy beak''), which makes the whole synthesis process more flexible and conforms to human input habits. This approach demonstrates great flexibility and more control over the content. Since it is more conform to people's input habits, it has better application prospects because the current research of artificial intelligence is more inclined to serve people. Nevertheless, the text description controls both the object and detailed information, but for the shape, size, and position of the object, it seems to be ineffective. For the research of image synthesis based on the text description, many works have been done, and encouraging results have been achieved. However, none of these works can control the shape, size, and position of the synthesized object.

To alleviate this problem and achieve better control of the synthesis details, Reed *et al.* proposed the Generative Adversarial What-Where Network (GAWWN) [10], using the bounding box and the key points to determine the location and shape of the target, and then generated specific content based on the text description. GAWWN is more flexible and controllable. On the one hand, the input information (bounding box, key point, and text description) can be determined artificially. On the other hand, the overall control degree is higher than that of only using the text description. Although GAWWN has achieved some success, it has two obvious problems. Firstly, the authenticity of the result is comparatively poor. Secondly, the control implemented by using the bounding box or key points is relatively rough, which does not achieve the real fine-grained control effect.

To achieve better fine-grained control and generate more authentic results, we propose a customized GAN. The image is generated by combining the contour and text description, as shown in Fig. 1. The contour is used to determine the specific shape, size, and position information of the object. Then the text description is used for generating the specific content. Finally, the high-quality images based on the hand-drawing contour and artificial text description are

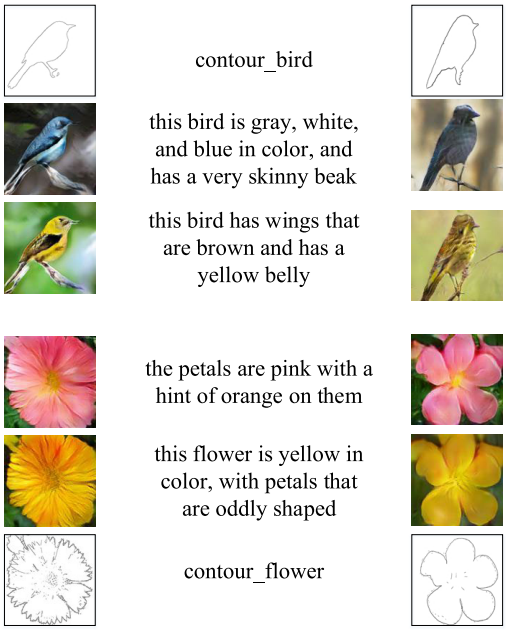

FIGURE 1. The results of the corresponding birds and flowers under different texts and contours. They are consistent with the corresponding text description while retaining the contour shape. The left contours are obtained by pre-processing the original dataset. The contours on the right are drawn by hand.

obtained by our method. It realizes the fine-grained control while also completes the generation of the realistic image.

Our contributions are as follows: (1) a new customized image generation method is proposed to achieve fine-grained control and high-quality image generation. (2) the whole process of image generation can be controlled manually (draw the shape and describe content manually), which makes our method have the best flexibility. (3) experiments on the Caltech-UCSD Birds [11] and the Oxford-102 flower [12] datasets show the effectiveness of the method. (4) due to the realization of the fine-grained and full artificial control of the input, image synthesis has taken a big step towards the direction of industrial application.

The rest of this paper is arranged as follows. Section II briefly reviews the related research works on image synthesis. The related background techniques are introduced in Section III. Our method details are discussed in Section IV and validated in Section V with promising experimental results. Section VI concludes our work.

II. RELATED WORK

The rapid development of deep learning technology in recent years has made great progress in the field of image processing, such as image segmentation [13]–[16], image inpainting [17]–[20], image enhancement [21], [22], and image clustering [23], [24].

Compared with the traditional research of image processing, image generation is more challenging. Mansimov *et al.* [25] proposed the alignDRAW model, which is an extension

of the Deep Recurrent Attention Writer (DRAW) [26] model, by learning to estimate alignment between generating results and text. Autoregressive models [27], [28] obtained arresting results by modeling the conditional distribution of pixel space using the neural network. [29], [30] realized image synthesis by using the deterministic network as function approximation. Variational Autoencoders (VAE) [31], [32] defined the generation problem as a probability graph model and achieved the final generation by maximizing the lower bound of data likelihood. Besides these, the best overall performance ability is Generative Adversarial Networks (GAN) [1]–[3], [33]–[36]. It has shown encouraging image generation results. Because of the instability of training, many improvement works have been proposed to stabilize the training process and improve the quality of synthesis.

In order to make the generative image model useful, conditional image synthesis has been explored. The initial condition generation is based on simple image attributes or class labels [7], [8], which has achieved some better results, but it is not suitable for human basic input habits because it may require some professional knowledge. Besides, using property or category labels can not control the details. After that, there are some works of image generation conditioned on the image (pixel to pixel), including image super-resolution [37], [38], image editing [39]–[41], image style transfer [5], [42], [43]. Since the image is as the input, the overall content cannot be changed greatly, which limits the artificial control factors to a certain extent. In these works, there is a way of simple input and strong control, that is to utilize simple contour to synthesize image. This method is more practical than using labels because it fixes the basic shape of the synthetic image. Nevertheless, it can only control the shape and but not detailed information. At present, the image generation, which accords with the habit of human input, is using text description to synthesize images. Reed *et al.* [9] first implemented text to image synthesis using the end-to-end GAN architecture based on adversarial learning, which generated realistic images. Subsequently, Zhang *et al.* [44], [45] proposed StackGAN to generate more realistic results through multi-stage adjustment. Xu *et al.* [46] used the attention mechanism to make local fine-tuning to obtain better results. Based on the attention mechanism, Qiao *et al.* [47] and Zhu *et al.* [48] respectively utilized text reproduction and dynamic memory to improve the quality of results further. Zhang *et al.* [49] proposed a hierarchical nesting structure, and could generate larger and vivid images. Qiao *et al.* [50] employed prior knowledge to improve the quality of synthetic images further. Its prior knowledge is obtained from the result with the mask. Although the results of text-to-image synthesis are more real and more abundant, there is the same problem — for the same text description, the model can generate a variety of results that conform to the text description but have different shapes, sizes, and orientations, which means that the input text can only control the generated content, but not the specific shape. This problem makes the current image synthesis model based on text description less practical.

For better flexible and effective control, based on the text description, Reed *et al.* [10] proposed the GAWWN structure and realized the controllable image generation process for the first time by combining the object location and other annotations. The size and position of the object are determined by inputting the bounding box and key points information. No matter the bounding box, key points, or text description, GAWWN can be input artificially, which makes GAWWN have pretty practicability. However, their results are not satisfactory as well as the bounding box, and key points are rough information, which cannot accurately determine the specific shape of the object. Our customizable generation structure combines contour and text description to generate high-quality results, which realizes fine-grained image control generation and effectively solves the problems in GAWWN. Our method allows for the input of manually drawn contours and descriptions, which do not need to be matched one by one. It shows the whole process of fine-grained control of image generation. To the best of our knowledge, this is the first time to realize the controllable image generation process based on artificial hand-drawing contour and text description.

III. PRELIMINARIES

A. GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GAN) [3] realize high-level image synthesis via adversarial learning, which consists of two networks: a generator G and a discriminator D. G and D continue to conduct iterative adversarial training. G synthesizes images by mapping latent variable z to real data space. Its goal is to make D think the synthesized images are real. D receives the real image and the fake image generated by G, and it needs to distinguish the true and false of the received images accurately. The specific process can be defined as a minimax game, as shown in Equation 1:

$$\min_G \max_D V(D,G) = \sum_{x\sim p_{data}}[\log D(x)] + \sum_{z\sim p_z}[\log(1-D(G(z)))] \quad (1)$$

where z is noise vector sampled from a Gauss or uniform distribution p_z and x is sampled from original data distribution p_{data}. In practice, the task of G is modified to maximize $log(D(G(z)))$ rather than minimize $log(1 - D(G(z)))$ because of the problem of gradient vanishing [3].

The basic process of conditional GAN [7], [8] is the same as that of GAN. The difference is that in addition to receiving z or x, conditional GAN will add conditional information c such as $G(z, c)$ and $D(x, c)$ in generator and discriminator.

B. IMAGE-TEXT MATCHING

The core of the image-text matching task is to embed the image and text description in the same embedding space.

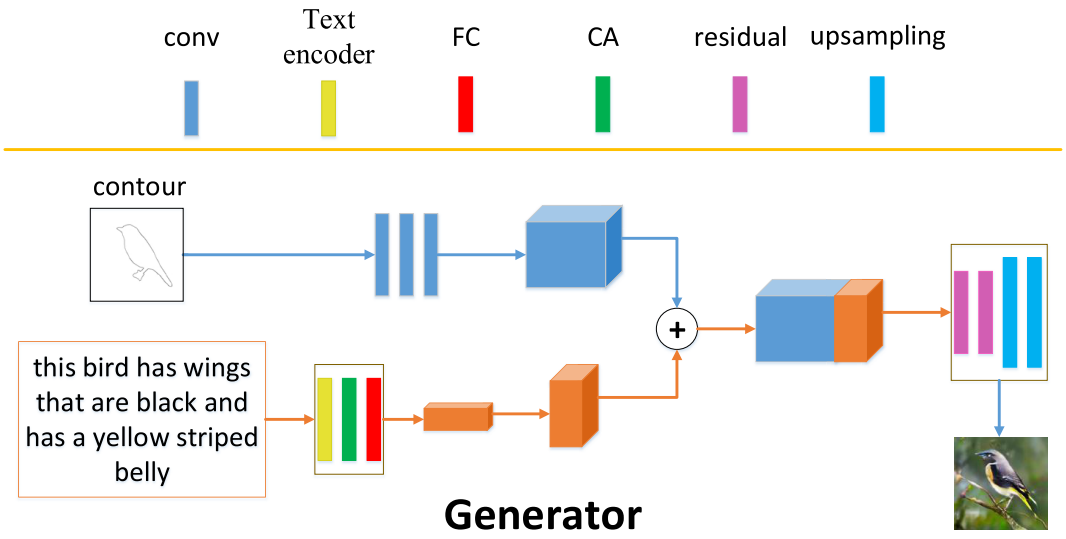

FIGURE 2. The generator structure of the model. The generator synthesizes the corresponding image based on the text description and contour. The synthesized image not only retains the contour shape but also conforms to the text description information.

Specifically, the convolutional neural network (CNN) is used to encode the image while the recurrent neural network (RNN) is used to encode the text, and then a joint embedding space is found to realize image-text matching. In this work, we adopt the method of Kiros *et al.* [51] for visual-semantic text embedding. The specific ranking loss function of image and text pairing is as follows:

$$\min_{\theta} \sum_{I}\sum_{T} max\{0, \alpha - cs(\varphi(I), \phi(T)) + cs(\varphi(I), \phi(T_{mis}))\} + \sum_{I}\sum_{T} max\{0, \alpha - cs(\varphi(I), \phi(T)) + cs(\varphi(I_{mis}), \phi(T))\} \tag{2}$$

where I represents the image and T for text. I_{mis} and T_{mis} are mismatching image and text. φ and ϕ represent image and text encoders, respectively. α is a margin value. cs denotes the cosine similarity of $\varphi(I)$ and $\phi(T)$. θ indicates all parameters in the encoder. The goal of the loss function is to minimize the cosine similarity between matched image-text and maximize the cosine similarity between mismatched image-text.

IV. CUSTOMIZABLE GAN

A. NETWORK ARCHITECTURE

The architecture of our method is shown in Fig. 2 and 3. It is built upon conditional GAN framework conditioning on both contour and text description. Fig. 2 shows the network structure of the generator. In the generator, the contour and text description in the input is encoded in different ways and combined together, and then the corresponding result is synthesized by de-convolution [52].

Specifically, in the generator, the contour is encoded as features by a convolutional neural network (CNN). There are three layers of convolution, among which the ReLU activation function is used after convolution. In addition to the first layer, each ReLU has a Batch Normalization (BN) [53] before it. The text description is encoded as a text vector by the pre-trained text encoder, and then its dimension is changed to 128 through a fully connection (FC). Referring to the work of Zhang *et al.* [44], conditional augmentation (CA) has also been added to increase the number of text embeddings. The conditional augmentation technology is designed to produce more conditioning variables for the generator. It can make the latent data manifold more continuous, which is beneficial to the whole training process. The specific implementation equation is as follows:

$$D_{KL}(\mathcal{N}(\mu(\varphi_t), \sum(\varphi_t)) \,||\, \mathcal{N}(0, I)) \tag{3}$$

where $\mathcal{N}$ represents a Gaussian distribution, φ_t represents the encoded text vector, μ and $\sum$ represent the operation of the mean and diagonal covariance matrix, respectively. KL represents the Kullback-Leibler divergence, $\mathcal{N}(0, I)$ is a regularization term to prevent over-fitting.

In order to combine text embeddings with features extraction from the contour, spatial replication is performed to expand the dimension of text embeddings. Finally, the dimension of the contour extraction feature is $16 \times 16 \times 512$, and the dimension of text embeddings is $16 \times 16 \times 128$. After the connection, it will pass through two residual transformation units, which are composed of residual blocks [54]. Accordingly, the employment of residual blocks is to make the connection features more effective through the deeper layer processing. On the other hand, it can better learn the feature representations to ensure the contour of the generated image is consistent with the input contour. Finally, the generator synthesizes the corresponding result by up-sampling.

The discrimination process consists of two parts: the discrimination of true or fake image and of the consistency of image and text description. Fig. 3 shows the network structure

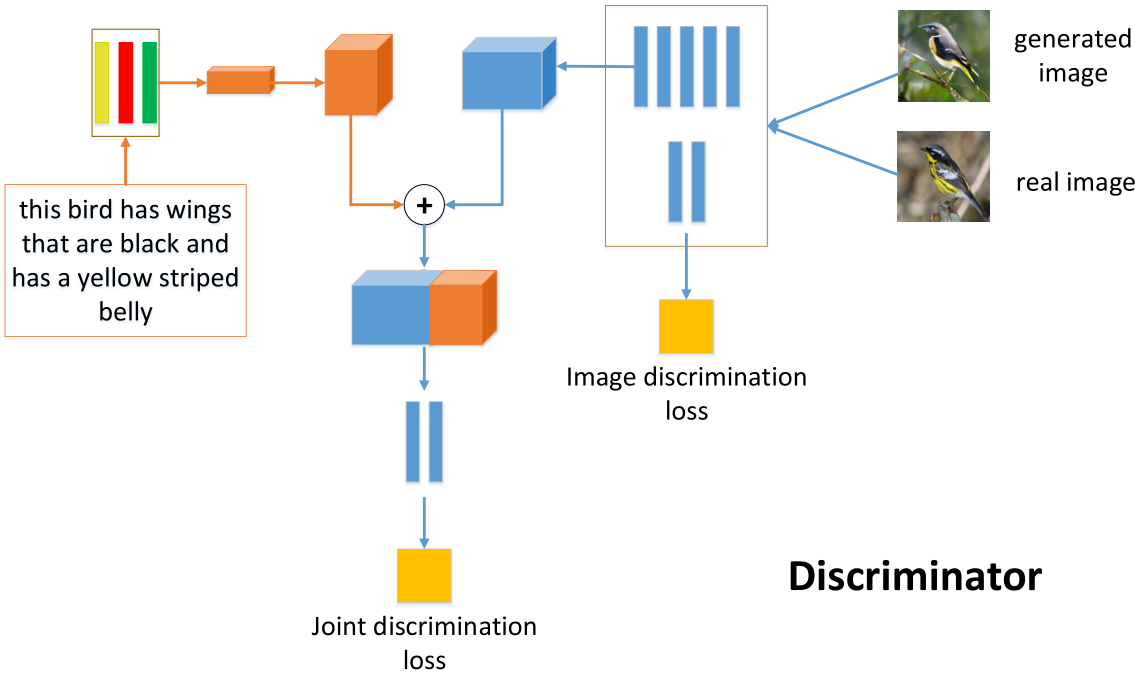

FIGURE 3. The discriminator structure of the model. The discriminator judges whether the received image itself is true or fake and the matching degree between the image and text.

of the discriminator. In the discriminator, there is feature extraction of the input image through down-sampling. There are two kinds of down-sampling, one is used to distinguish the real or fake image. The other is to distinguish the consistency of the image and text. The down-sampling for the real or fake image discrimination consists of two convolution layers: the first layer is followed by BN [53] and leaky-ReLU [55], the second layer is directly followed by the sigmoid function. For the discrimination of the consistency of the image and text, the image features first are extracted through five convolution layers, then combined with the text vector of extended dimension, and finally identified by two convolutions layers. Each convolution layer is followed by BN and leaky-ReLU, except for the last layer for discrimination. Unlike the generator, the features extraction dimension in the discriminator is $4 \times 4 \times 512$, and the text embeddings dimension is $4 \times 4 \times 128$.

The characteristic of the GAN is to achieve the goal of mutual promotion through adversarial learning. Therefore, both the generator and the discriminator can improve the performance of each other. In the discriminator, the image itself and the consistency matching of the image and text are used to distinguish, which can make the discriminator have better discrimination ability. As just mentioned, the performance improvement of the discriminator can promote the performance of the generator, so that the final synthesis effect is excellent.

B. ADVERSARIAL LEARNING PROCESS

Customizable image synthesis determines the shape through the contour and defines the specific content through the text description. This indicates that the result of the synthesis should match the basic shape of the input contour as well as the text description. We utilize the method of adversarial learning to train the whole process, as shown in algorithm 1.

Algorithm 1 CustomizedGAN Training Algrotihm

Input: matching text T, mismatching text T_{mis},
relevant text T_{rel}, real image I_{real},
contour con, number of epochs N
for $n = 1$ **to** N **do**
  $s = \phi(T)$
  $s_{mis} = \phi(T_{mis})$
  $s_{rel} = \phi(T_{rel})$
  $I_{fake} = G(con, s)$
  $d, d_{ucond} = D(I_{real}, s)$
  $d_f, d_{f\_ucond} = D(I_{fake}, s_{rel})$
  $d_{mis}, d_{mis\_ucond} = D(I_{real}, s_{mis})$
  $L_{D\_real} = \log(d) + \log(d_{ucond})$
  $L_{D\_mis} = (\log(1 - d_{mis}) + \log(d_{mis\_ucond}))/2$
  $L_{D\_fake} = (\log(1 - d_f) + \log(1 - d_{f\_ucond}))/2$
  $L_D = L_{D\_real} + L_{D\_mis} + L_{D\_fake}$
  $D = D - ss * \partial L_D / \partial D$
  $L_G = \log(d_f) + \log(d_{f\_ucond})$
  $G = G - ss * \partial L_G / \partial G$
end

There are three types of text input in the training process, that is, the matching text T, the mismatching text T_{mis}, and the relevant text T_{rel}. The relevant text represents the text related to the current training set. The purpose of this is to achieve a better result without being restricted to the other side, either in contour or text. This makes the final trained model robust.

For any contour and any text description in the same dataset, it can produce high-quality results. In the algorithm, ϕ is a pre-trained text encoder used to encode text into vectors. The generator synthesizes fake images based on the input contour and text. The discriminator distinguishes three situations: the real image with the matched text, the fake image with the relevant text, the real image with the mismatched text. The purpose of introducing text that does not match the real image is to make the whole network learn the situation of mismatch so that it can improve the final matching degree. Unlike the general GAN, the discriminator in our algorithm returns two outputs: first is the degree of the image-text matching, and second is the judgment of the authenticity of the image. The advantage of this method is to distinguish the results from many aspects to improve the discrimination ability of the discriminator. Because of the antagonism characteristic of GAN, the improvement of discrimination ability will promote the generation ability of generator so that high-quality results can finally be obtained. The specific loss functions are as follows:

$$\begin{aligned} L_D = & \sum_{(I,T)\sim p_{data}} \{\log D_0(I_{real}, T) + [\log(1 - D_0 \\ & (I_{real}, T_{mis})) + \log(1 - D_0(I_{fake}, T_{rel}))]/2\} \\ & + \{\log D_1(I_{real}, T) + \log D_1(I_{real}, T_{mis}) \\ & + \log(1 - D_1(I_{fake}, T_{rel}))/2\} \end{aligned} \tag{4}$$

$$L_G = \sum_{(I,T)\sim p_{data}} \log D_0(I_{fake}, T_{rel}) + \log D_1(I_{fake}, T_{rel}) \tag{5}$$

where D_0 represents the first output of the discriminator and D_1 represents the second. In L_D, the content of the first brace represents the conditional loss (image-text matching), and the second brace content represents the unconditional loss (image). G and D are updated by the SGD method, where ss is the step size.

C. TRAINING DETAILS AND FURTHER EXPLORATION

In the training process, the initial learning rate is set to 0.0002, and it decays to half of the original every 100 epochs. Adam optimization [56] with a momentum of 0.5 is used to optimize and update parameters. A total of 600 epochs are trained iteratively in the network, of which the batch size is 64. In leaky-ReLU, the leaky value is 0.2.

In the task of feature extraction, some current works [54], [57] have shown that deeper the layer is, better the feature representations can be extracted. In our work, the contour is a simple shape, and the images of birds and flowers are not complicated, so we choose the pre-trained VGG [58] trained on ImageNet [59] as the feature extraction model to improve the quality of the results further. Specifically, the output of the fourth convolution layer (conv4) of VGG is the result of feature extraction. Moreover, VGG16 and VGG19 are both used to explore better results. The specific experimental results are shown in Section V.

V. EXPERIMENTS

A. DATASET AND DATA PREPROCESSING

We validated our method on the Caltech-UCSD Birds [11] dataset and the Oxford-102 flower [12] dataset. 10 text descriptions are collected [60] for each image. The CUB dataset contains 11,788 images with 200 classes. The Oxford-102 dataset contains 8,189 images with 102 classes. Following Reed *et al.* [9], we split CUB dataset to 150 train classes and 50 test classes as well as Oxford-102 to 82 train classes and 20 test classes.

In order to experiment with customized synthesis, it is necessary to pre-process the contour map. For the processing of the bird dataset, we first download the corresponding binary image on its official website, then turn the black part of the background into white and retain the outermost contour lines. For the contour map of the flower dataset, the outermost peripheral contour cannot show the overall structure of the flower well. Therefore, for flower image processing, not only the outermost peripheral contour but also the hierarchical information is essential. For this goal, we use the Canny operator to process the flower foreground map, the official website provides the foreground map of the blue background, and pure foreground map can be obtained by turning the blue to white, to obtain the results required. The relevant processing results are shown in Fig. 4.

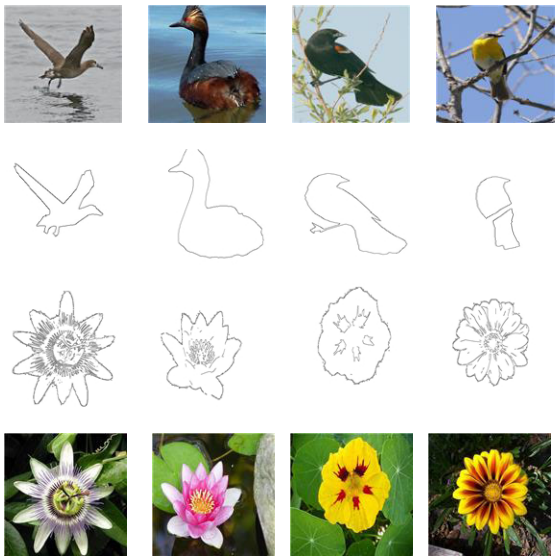

FIGURE 4. Some processed contour results. As a result, the foreground in the original image is well drawn in the form of curves. The results of flowers not only contain the outline information of the outermost part but also include the interior information.

B. QUALITATIVE RESULTS

Firstly, we compare the existing text-to-image synthesis model. The existing T2I model has two main directions. One is based on the multi-stage synthesis, and the other is based on the attention mechanism. AttnGAN [46] not only uses multi-stage synthesis but also is based on the attention mechanism, so we choose AttnGAN as the representative model for comparison. The specific results are shown in Fig. 5. From the comparison results, it can be seen that for the same text description, our model not only conforms to the text

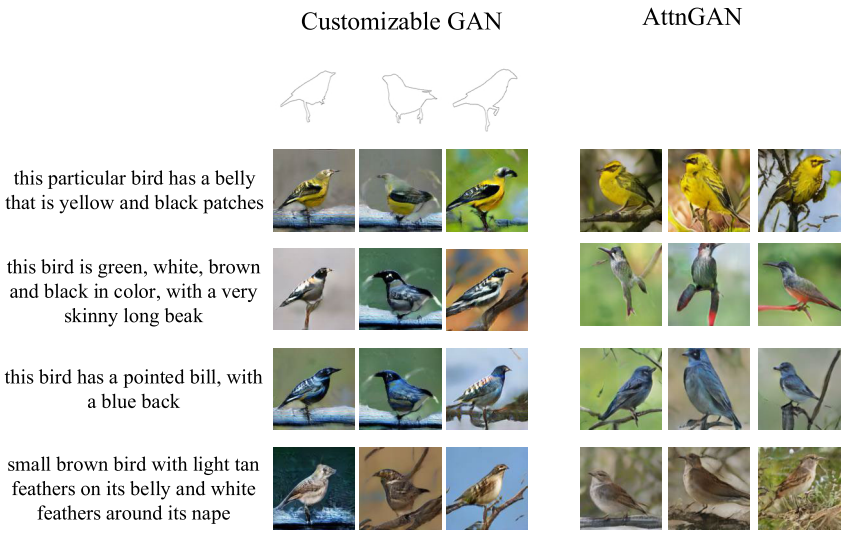

FIGURE 5. The comparison between our method and the existing text synthesis image model. The text-to-image synthesis model can not control the contour information of the synthesized object, and we can control the specific contour of the object while conforming to the basic text description information.

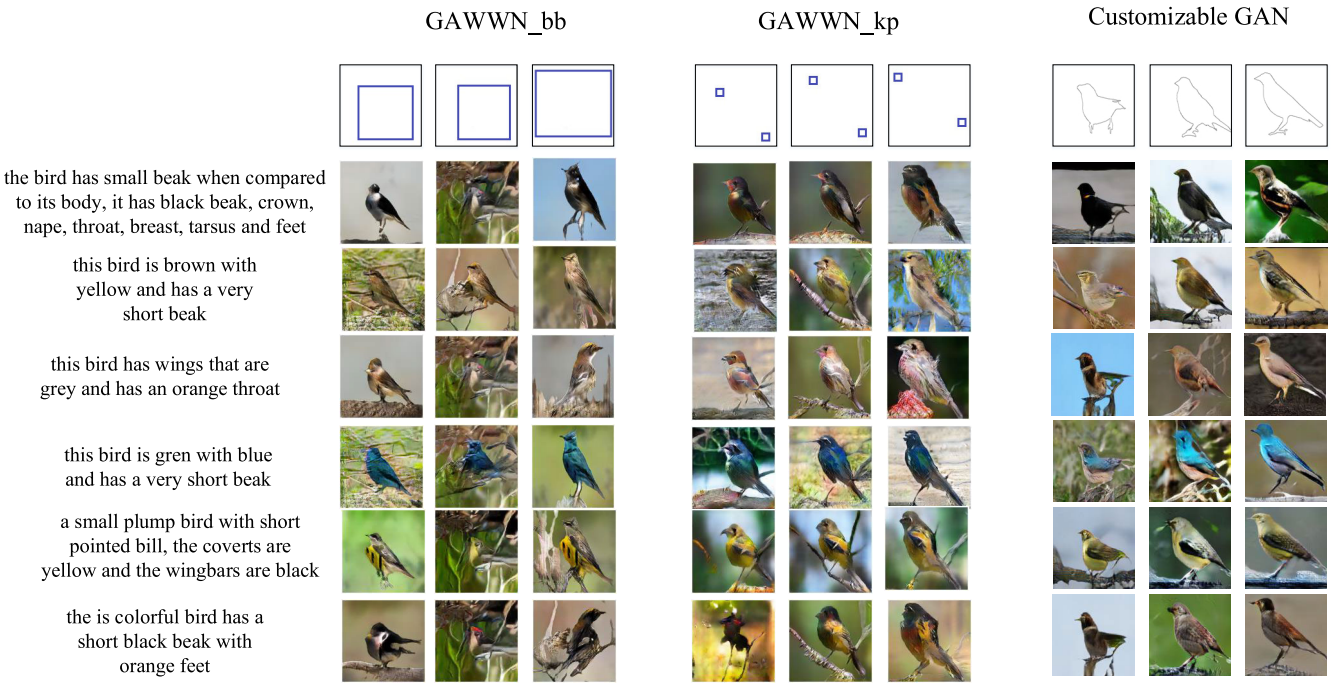

FIGURE 6. The comparison between our method and GAWWN (including two results based on bounding box and key points). It can be seen from the comparison results that our results are obviously superior to GAWWN and have a better degree of control than GAWWN.

description but also can control the shape of the final synthesized object through simple contour. For AttnGAN, multiple results can be synthesized, but the shape, size, and position of the synthesized object are different, which indicates that the existing T2I model can not control the specific style. This reflects the low practicability of the existing T2I model. Compared with the existing T2I model, GAWWN [10] and our method are all studying in the direction of more effective image synthesis control. Meanwhile, the input of GAWWN can also be artificially controllable. Therefore, we choose to compare our method with GAWWN.

Compare our method with the existing controllable image synthesis based on text and annotations (GAWWN), as shown in Fig. 6 and 7. There are two kinds of comments in GAWWN: the bounding box, and the key point information. In the figure, GAWWN_bb represents the GAWWN

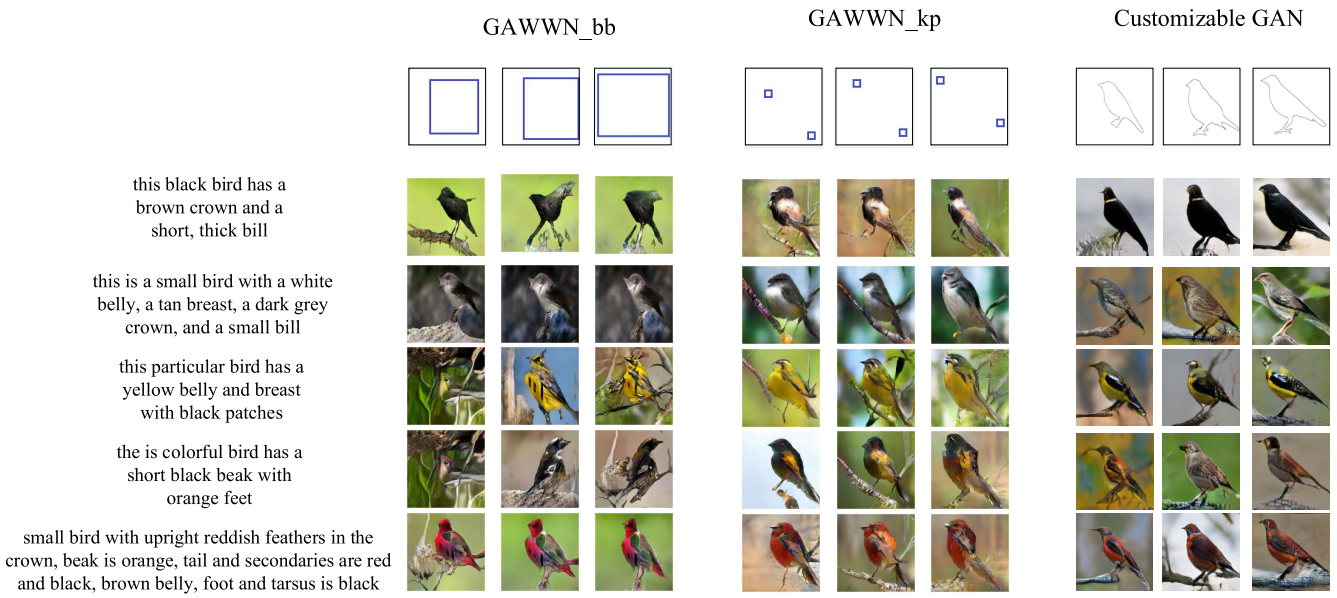

FIGURE 7. The comparison results of second group between our method and GAWWN. These results also reflect the roughness control and poor authenticity of GAWWN. In contrast, our results are more realistic, the degree of control is also more refined.

result based on the bounding box. The input bounding box can only control the generated area, which is powerless for the specific shape and orientation. GAWWN_kp represents the corresponding result based on the key points. Key points control the basic position and orientation of the generation, but the specific shape cannot be determined. Simultaneously, the synthesis results based on the bounding box and key points generally have poor authenticity. All these shows that indicate although GAWWN has high flexibility in image synthesis, its overall control is relatively poor and rough, and the results of synthesis are not satisfactory.

Compared with GAWWN, our method has higher control ability, evidenced by the specific shape, position, and orientation of the synthesized results. This shows more fine-grained control than GAWWN's rough control and realizes the genuinely customized image synthesis. For the generated results, on the one hand, our method maintains the consistency with the input contour and text description. On the other hand, it is better than GAWWN in authenticity. This demonstrates the superiority of our method in controlling the generation of authentic results.

In addition to the comparison with GAWWN, we also made an internal comparison. In this paper, we compared the three methods of contour feature extraction without VGG, with VGG16, and with VGG19, as shown in Fig. 8. The results of the three methods have a high degree of authenticity. They maintain both the shape of the input contour and match the content of the text description. From a more detailed level (eye, pecking, texture) of comparison, the results obtained by using VGG are better than those not applicable to VGG, which makes the results of using VGG have pretty authenticity. Compared to VGG16, VGG19 handles the details of the texture better to make the results more realistic.

We extended our method on the flower dataset and made the internal comparison, as shown in Fig. 9. The results of the three methods are authentic. Overall, all results maintain the shape of the contour and conform to the text description. In comparison, VGG16 has a higher degree of agreement with the contour because it reflects better the overall details of the contour, which makes its results have higher authenticity.

C. QUANTITATIVE RESULTS

For the evaluation of the generation model, Human Rank (HR) is used to quantify the comparison models. HR can be used to evaluate whether the synthesized image conforms to subjective effects (such as authenticity, matching degree with text, etc.), and it is widely used in various image synthesis evaluation, such as [40], [44], [45].

In this work, we employed 10 subjects to rank the quality of synthetic images by different methods. The text descriptions and contours corresponding to these results are all from the test set and are divided into 10 groups for use by 10 subjects. For the bird datasets, we established two ways for quantitative comparison. One of them contains three results: 1) GAWWN_bb, 2) GAWWN_kp, 3) ours without VGG. The other includes five synthetic results: 1) GAWWN_bb, 2) GAWWN_kp, 3) ours without VGG, 4) ours with VGG16, 5) ours with VGG19. In this way, the comparison of the bird is three tuples (bird_1) and five tuples (bird_2), respectively. The employers were not informed of the method corresponding to the result, but only knew the text description and contour, bounding box, and key points corresponding to the current result. The subjects were asked to rank the results

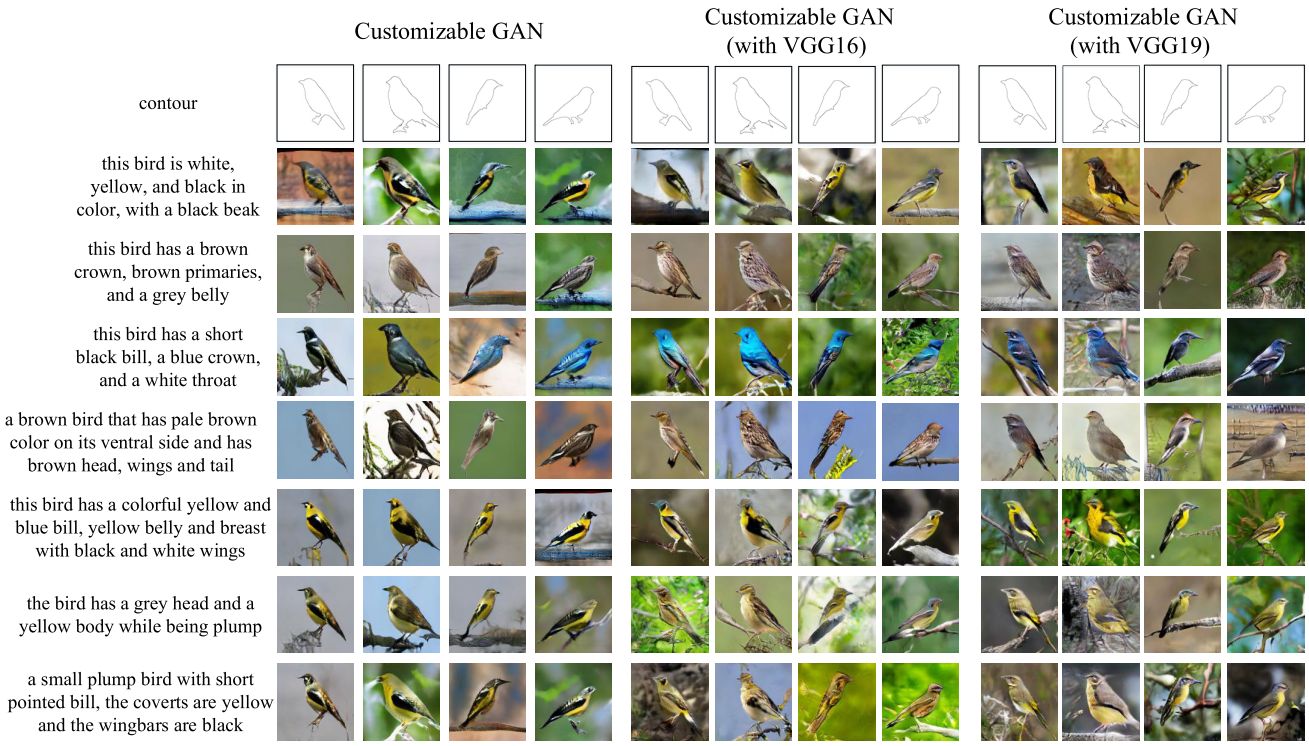

FIGURE 8. The comparison bird results of our method without VGG and with VGG16, with VGG19. It can be seen that the results of using VGG are better in details (such as eyes, pecking).

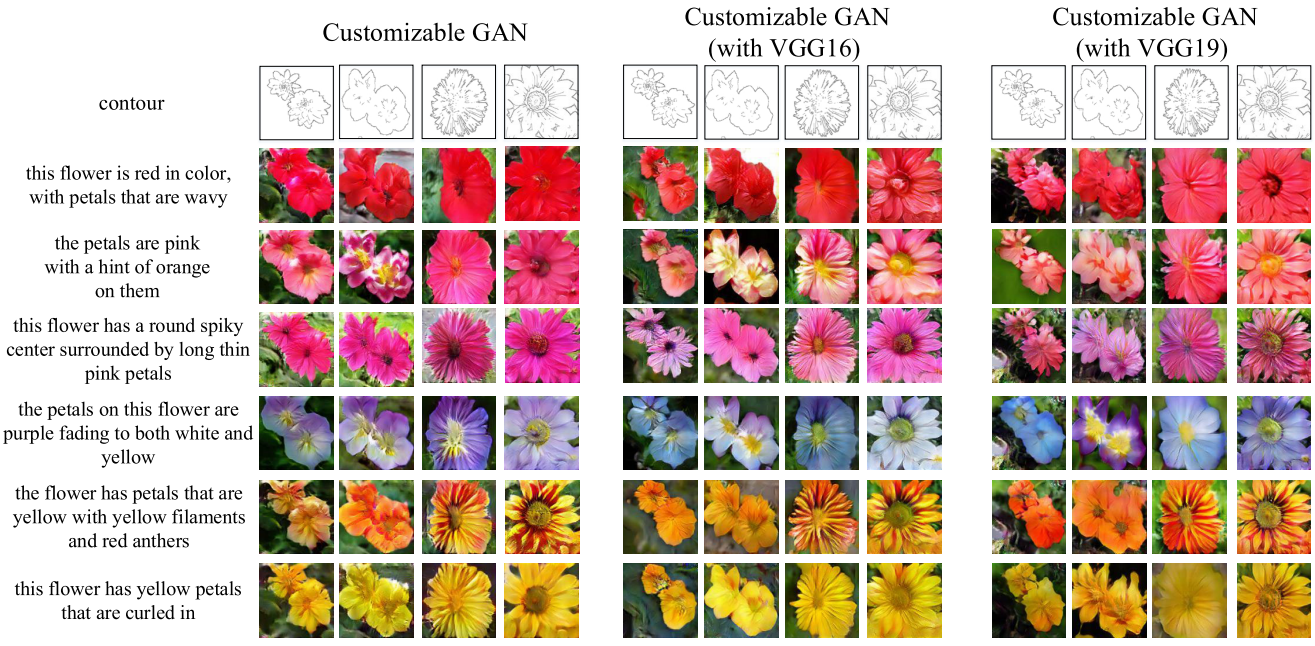

FIGURE 9. The comparison flower results of our method without VGG and with VGG16, with VGG19.

(bird_1: 1 is best, 3 is worst; bird_2: 1 is best, 5 is worst) in the following ways:

- Whether the result is highly consistent with control information (the contour or bounding box or key points)
- Whether the result matches the text description
- The level of the authenticity of all results

The average score will be calculated for the ranking results of all subjects, as shown in Tables 1 and 2. The comparative results show the following points:

TABLE 1. The results of quantitative comparison between our three methods and GAWWN. It includes three aspects of comparison: one is the consistency with the control information (consistency), the other is the matching with the text content (text), and the third is the authenticity of the results (authenticity).

	GAWWN_bb	GAWWN_kp	ours	ours+VGG16	ours+VGG19
consistency	4.64	4.184	2.116	2.004	2.02
text	4.096	3.796	2.4	2.336	2.304
authenticity	4.456	3.904	2.368	2.2	2.072

TABLE 2. The quantitative comparison results between our method with GAWWN in CUB dataset.

	GAWWN_bb	GAWWN_kp	ours
consistency	2.784	2.510	1.269
text	2.457	2.277	1.440
authenticity	2.674	2.342	1.421

1) MORE AUTHENTICITY AND BETTER TEXT MATCHING

Compared with GAWWN in Tables 1 and 2, it is obvious that our method has higher authenticity and degree of text matching. In comparison, the results of using key points (kp) in GAWWN are better than those of using the bounding box (bb). However, compared with our results, the overall authenticity and matching of GAWWN_kp are still worse than us.

2) MORE EFFECTIVE CONTROL

In the process of image synthesis, the control of our method is more effective since it shows better consistency with the control information. The control degree of GAWWN_kp is better than that of GAWWN_bb. This is consistent with the subjective comparison. In subjective results, GAWWN_kp can control the basic direction of synthesis, but GAWWN_bb cannot. Compared with GAWWN_kp, our method has more excellent control. The reason for this circumstance is that our results can not only control the synthesis direction but also control the specific shape, while GAWWN_kp can not control the shape.

3) BETTER PERFORMANCE WHEN USING VGG

Table 1 shows that the results obtained by our three methods (without VGG, wit VGG16, with VGG19) are not significantly different. In close comparison, the results of using the VGG model are better than those of not using VGG. This reflects that VGG can extract better contour features and promote the synthesis of final results.

D. ABLATION STUDY

It can be found in Table 1 that among the results of birds, VGG19 is better than VGG16, and VGG16 is better than not using VGG. Does this phenomenon also apply to flower results? What are the differences between not using VGG and using VGG 16 and VGG 19 and the reasons behind the differences? To solve these problems, we conducted an ablation study.

TABLE 3. The internal quantitative comparison results of our methods in CUB dataset.

	ours	ours+VGG16	ours+VGG19
consistency	1.269	1.202	1.212
text	1.440	1.402	1.382
authenticity	1.421	1.320	1.243

TABLE 4. The internal quantitative comparison results of our methods in Oxford-102 flower dataset.

	ours	ours+VGG16	ours+VGG19
consistency	1.245	1.121	1.229
text	1.229	1.154	1.225
authenticity	1.282	1.153	1.169

For the internal comparison of our three methods, it can be seen from Table 1 that there is no obvious difference. In the separate comparison, the result of using VGG is better than that of not using VGG. In Tables 3 and 4, among the results of birds, the overall authenticity of VGG19 is better than that of VGG16, while that of flowers is the opposite. The reason for this is that the proportion of birds in the image is relatively small (generally less than 50%), so the judgment of the authenticity of bird image is more dependent on the generation of bird details. VGG19 performs the best authenticity in generating bird results, which shows that it does best in detail generation. Compared with bird images, the proportion of flowers in the image is generally more than 80%, so its authenticity depends on the overall structure. In the authenticity of flower results, VGG16 is better than VGG 19, which indicates that VGG 16 does the best performance in structural consistency. Although VGG19 can obtain pretty detailed information in flower results, the authenticity of VGG16 results is better because flowers pay more attention to integrity. VGG16 also showed the best structural consistency in birds results, indicating that VGG16 is indeed better than VGG19 in terms of structural consistency.

On the whole, VGG19 is better than VGG16 in detail synthesis, and VGG16 is better than VGG19 in overall structure synthesis. This is reasonable because VGG19 is deeper

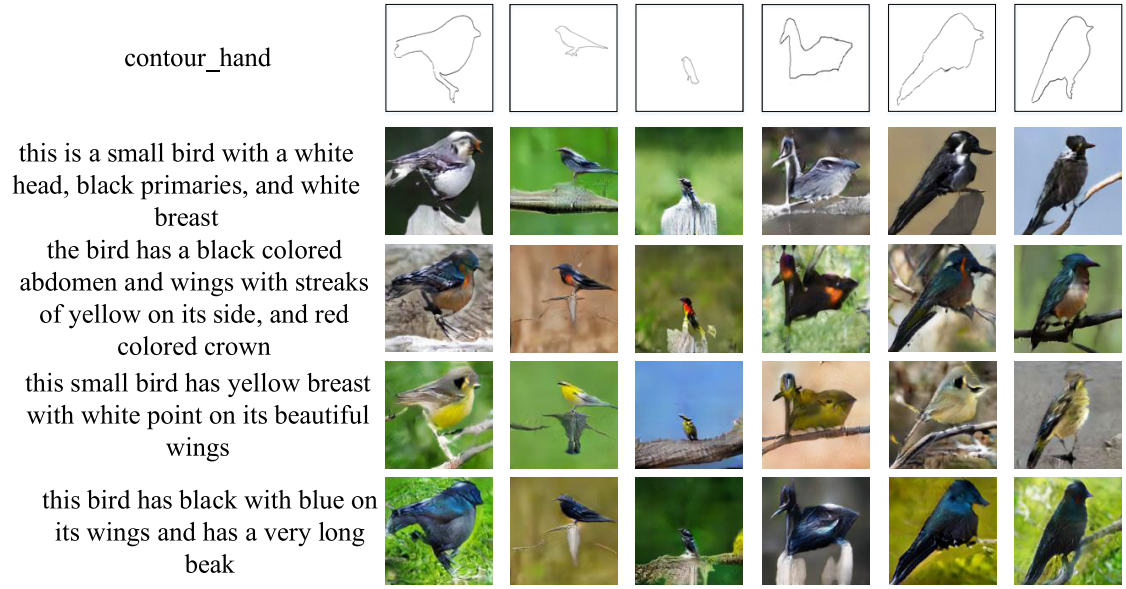

FIGURE 10. The text descriptions on the left are all artificial descriptions that do not exist in the dataset. The contours are also drawn manually. The results show the effectiveness of our method in generating high-quality results and the high flexibility image control generation.

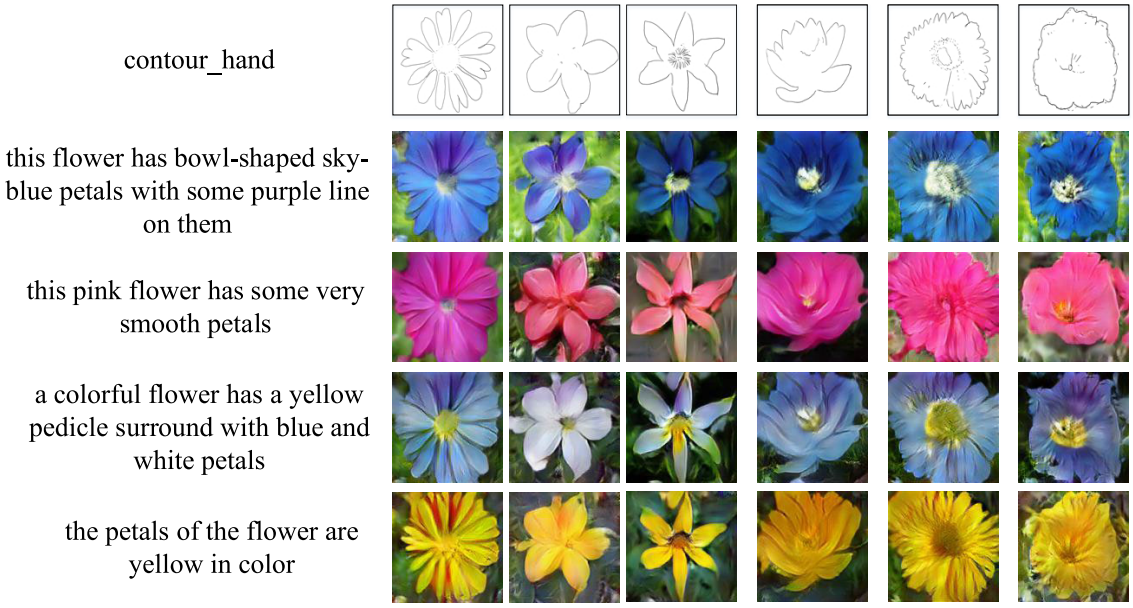

FIGURE 11. The customized results of flowers. Equivalent to the condition of birds, the text descriptions also do not exist in the dataset of artificial description, and the contours are also drawn by artificial.

than VGG16, so it can extract more detail-oriented feature information. The number of layers of VGG16 is relatively small, so it pays more attention to the overall feature information. VGG is a network structure specially designed for feature extraction, which performs well in classification, segmentation, and other tasks. Therefore, the use of VGG is better than the simple use of convolution operation (without VGG) to extract features, so the final performance is better.

E. CONTROLLABLE IMAGE SYNTHESIS

The most important feature of our work is to realize fine-grained controllable image synthesis based on artificial hand drawing and manual description. The relevant results are shown in Fig. 10 and 11. Both the contour and the text description in the figure are artificial and do not exist in the dataset. Besides, it can also be seen from the results that our model can generate corresponding high-quality results for the different contour of shapes, sizes, positions, and orientations. Such as shown in the bird results in Figure 10, the first, second, and third columns well show that the model can synthesize high-quality results based on different contour sizes and positions. At the same time, the fourth, fifth, and sixth columns also show that the model can adapt to different contour orientations and generate high-quality results.

The flower results in Figure 11 also reflect that the model can adapt to different contour shapes, sizes, and orientations and generate high-quality flower results. These results not only reflect well the hand-drawn contour and artificial text description content but also have a high degree of authenticity. This demonstrates the effectiveness of our method in synthesizing high-quality authentic images and shows the high flexibility and controllability of our method because all inputs can be controlled artificially.

VI. CONCLUSION

In this work, we propose a customizable image synthesis based on contour and text descriptions. The high-quality image synthesis is achieved through adversarial learning. The synthesis results indicate that our method maintains the basic shape of the contour, while also conforms to the text description. Furthermore, we have evaluated the model on the Caltech-UCSD Birds dataset and the Oxford-102 flower dataset. The experimental results demonstrate the effectiveness and robustness of our method. Besides, the high-quality image synthesis results based on hand-drawn contour and artificial descriptions are also illustrated to prove that our method is highly controllable and flexible.

From the synthetic results, the foreground content generated by the model is basically realistic, but the background information in both bird and flower results is relatively general, even there are some subjective unrealistic situations. To solve this problem, we will separate the synthesis of foreground content and background information in the future research, and then merge the synthesized foreground and background results to generate the final results.

REFERENCES

[1] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," unpublished.

[2] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. NIPS*, Montreal, QC, Canada, 2015, pp. 1486–1494.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Montreal, QC, Canada, 2015, pp. 2672–2680.

[4] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR*, San Juan, PUR, 2016, pp. 1–16.

[5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5967–5976.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2242–2251.

[7] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 2172–2180.

[8] M. Mirza and S. Osindero, "Conditional generative adversarial nets," unpublished.

[9] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. ICML*, New York, NY, USA, 2016, pp. 1060–1069.

[10] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 217–225.

[11] W. Catherine, B. Steve, W. Peter, P. Pietro, and B. Serge, "The Caltech-UCSD birds-200-2011 dataset," Caltech, Univ. Pasadena, CA, USA, Tech. Rep. CNS-TR-2010-001, Dec. 2011.

[12] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Bhubaneswar, India, Dec. 2008, pp. 722–729.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[14] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 91–99.

[16] K. He, G. Gkioxari, P. Dollar, and R. B. Girshick, "Mask R-CNN," in *Proc. ICCV*, Venice, Italy, 2017, pp. 2980–2988.

[17] F. Liu, F. A. Reda, K. J. Shih, T. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. ECCV*, Munich, Germany, 2018, pp. 89–105.

[18] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Image inpainting via deep feature rearrangement," in *Proc. ECCV*, Munich, Germany, 2018, pp. 3–19.

[19] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 4470–4479.

[20] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4076–4084.

[21] S. Park, S. Yu, M. Kim, K. Park, and J. Paik, "Dual autoencoder network for retinex-based low-light image enhancement," *IEEE Access*, vol. 6, pp. 22084–22093, Jul. 2018.

[22] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.

[23] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang, "Structured AutoEncoders for subspace clustering," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5076–5086, Oct. 2018.

[24] X. Peng, H. Zhu, J. Feng, C. Shen, H. Zhang, and J. T. Zhou, "Deep clustering with sample-assignment invariance prior," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 31, 2019, doi: 10.1109/TNNLS.2019.2958324.

[25] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," in *Proc. ICLR*, San Juan, Puerto Rico, 2016, pp. 1–12.

[26] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," in *Proc. ICML*, Lille, France, 2015, pp. 1462–1471.

[27] A. V. D. Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves, "Conditional image generation with PixelCNN decoders," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 4790–4798.

[28] A. V. D. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. ICML*, New York, NY, USA, 2016, pp. 1747–1756.

[29] A. Dosovitskiy, J. T. Springenberg, and T. Brox, "Learning to generate chairs with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1538–1546.

[30] S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee, "Deep visual analogy-making," in *Proc. NIPS*, Montreal, QC, Canada, 2015, pp. 1252–1260.

[31] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, Banff, AB, Canada, 2014.

[32] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. ICML*, Beijing, China, 2014, pp. 1278–1286.

[33] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," in *Proc. ICLR*, Toulon, France, 2017, pp. 1–13.

[34] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," in *Proc. ICLR*, Toulon, France, 2017, pp. 1–25.

[35] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: Conditional iterative generation of images in latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3510–3520.
[36] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 2226–2234.
[37] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 105–114.
[38] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised MAP inference for image super-resolution," in *Proc. ICLR*, Toulon, France, 2017, pp. 1–17.
[39] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," in *Proc. ICLR*, Toulon, France, 2017, pp. 1–15.
[40] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic image synthesis via adversarial learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5707–5715.
[41] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proc. ECCV*, Amsterdam, The Netherlands, 2017, pp. 597–613.
[42] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. NIPS*, Long Beach, CA, USA, 2017, pp. 700–708.
[43] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. ICLR*, Toulon, France, 2017, pp. 1–14.
[44] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5908–5916.
[45] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019.
[46] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1316–1324.
[47] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1505–1514.
[48] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5802–5810.
[49] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6199–6208.
[50] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Learn, imagine and create: Text-to-image generation from prior knowledge," in *Proc. NIPS*, Vancouver, BC, Canada, 2019, pp. 885–895.
[51] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," unpublished.
[52] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, Zürich, Switzerland, 2014, pp. 818–833.
[53] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, Lille, France, 2015, pp. 448–456.
[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA Jun. 2016, pp. 770–778.
[55] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," unpublished.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–15.
[57] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826.
[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–14.
[59] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[60] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 49–58.

ZHIQIANG ZHANG (Graduate Student Member, IEEE) was born in Lu'an, Anhui, China, in 1995. He received the B.S. degree from the Southwest University of Science and Technology, Mianyang, China, in 2017, where he is currently pursuing the M.S. degree. His research interests include image synthesis, multimodal information transformation and fusion, game theory, computer vision, and deep learning.

WENXIN YU (Member, IEEE) was born in Mianyang, Sichuan, China, in 1984. He received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 2006, and the M.S. and Ph.D. degrees from Waseda University, Tokyo, Japan, in 2010 and 2013, respectively. From 2015 to 2017, he was an Associate Research Fellow with the School of Computer Science and Technology, Southwest University of Science and Technology. Since 2018, he has been the Vice President of the School of Computer Science and Technology. He has been exploring the cutting-edge direction of video and image processing and focused on 3-D image synthesis, image stereo matching, and other issues. His main research interests include 3-D multiview synthesis filling technology, image stereo matching technology, multiview compatible fast coding algorithm, neural networks, pattern recognition, low-power consumption video decoding algorithm, and image error concealment technology.

JINJIA ZHOU (Member, IEEE) received the B.E. degree from Shanghai Jiao Tong University, China, in 2007, and the M.E. and Ph.D. degrees from Waseda University, Fukuoka, Japan, in 2010 and 2013, respectively. From 2013 to 2016, she was a Junior Researcher with Waseda University. She was selected as a JST PRESTO Researcher for the period of 2017–2021. She is currently an Associate Professor and a Co-Director of the English-Based Graduate Program, Hosei University. She is also a Senior Visiting Scholar with the State Key Laboratory of ASIC and System, Fudan University, China. Her research interests include algorithms and VLSI architectures for multimedia signal processing.

Dr. Zhou received the Research Fellowship of the Japan Society for the Promotion of Science, from 2010 to 2013, and the Hibikino Best Thesis Award, in 2011. She was a recipient of the Chinese Government Award for Outstanding Students Abroad of 2012. She was a co-recipient of the ISSCC 2016 Takuo Sugano Award for Outstanding Far-East Paper, the Best Student Paper Award of VLSI Circuits Symposium 2010, and the Design Contest Award of ACM ISLPED 2010. She participated in the design of the world first 8K UHDTV video decoder chip, which was granted the 2012 Semiconductor of the Year Award of Japan.

XUEWEN ZHANG was born in Chengdu, China, in 1996. He received the B.S. degree from the Southwest University of Science and Technology, Mianyang, China, in 2019, where he is currently pursuing the M.S. degree in computer science and technology. His research interests include image processing, machine learning, and deep learning.

NING JIANG received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 2006, and the M.S. and Ph.D. degrees from Waseda University, Tokyo, Japan, in 2010 and 2013, respectively. From 2016 to 2018, he was a Lecturer with the School of Computer Science and Technology, Nantong University. Since 2018, he has been an Associate Professor with the School of Computer Science and Technology, Southwest University of Science and Technology. His research interests include computer vision and pattern recognition, deep learning, and artificial neural networks.

GANG HE (Member, IEEE) received the B.S. degree in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 2008, and the M.S. and Ph.D. degrees in system LSI from Waseda University, Kitakyushu, Japan, in 2011 and 2014, respectively.

He currently works with the State Key Laboratory of Integrated Services Networks, Xidian University. He has authored or coauthored over 20 articles in international journals and conferences. His current research interests include video coding algorithm and its VLSI architecture, image processing, and machine learning.

ZHUO YANG received the B.S. and M.S. degrees from the Beijing Institute of Technology, Beijing, China, in 2005 and 2008, respectively, and the Ph.D. degree from Waseda University, Tokyo, Japan, in 2012. He is currently an Assistant Professor with the School of Computers, Guangdong University of Technology. His main research interests include image processing, computer vision, and VR/AR.