MOGAN: Morphologic-Structure-Aware Generative Learning From a Single Image

Jinshu Chen, Qihui Xu, Qi Kang, Senior Member, IEEE, and MengChu Zhou, Fellow, IEEE

Abstract—In most interactive image generation tasks, given regions of interest (ROI) by users, the generated results are expected to have adequate diversities in appearance while maintaining correct and reasonable structures in original images. Such tasks become more challenging if only limited data is available. Recently proposed generative models complete training based on only one image. They pay much attention to the monolithic feature of the sample while ignoring the actual semantic information of different objects inside the sample. As a result, for ROI-based generation tasks, they may produce inappropriate samples with excessive randomicity and without maintaining the related objects’ correct structures. To address this issue, this work introduces a morphologic-structure-aware generative adversarial network named MOGAN that produces random samples with diverse appearances and reliable structures based on only one image. For training for ROI, we propose to utilize the data coming from the original image being augmented and bring in a novel module to transform such augmented data into knowledge containing both structures and appearances, thus enhancing the model’s comprehension of the sample. To learn the rest areas other than ROI, we employ binary masks to ensure the generation isolated from ROI. Finally, we set parallel and hierarchical branches of the mentioned learning process. Compared with other single image generative adversarial network schemes, our approach focuses on internal features, including the maintenance of rational structures and variation on appearance. Experiments confirm a better capacity of our model on ROI-based image generation tasks than its competitive peers.

Index Terms—Generative adversarial networks (GANs), machine learning, morphologic awareness, regions of interest (ROI)-based image generation tasks, single sample.

I. INTRODUCTION

In many interactive image generation tasks, users tend to be more interested in certain targets or objects in a given sample (i.e., regions of interest or ROI) while paying less attention to the rest areas (called background). As a kind of unsupervised model, generative adversarial networks (GANs) [1] are capable of most generation tasks, which have greatly promoted the development of many fields, such as image inpainting [2], [3], [4], image-to-image translation [5], [6], [7], [8], and image synthesis [9], [10], [11]. However, owing to their frail structures that are hard to converge, GANs heavily depend on large datasets or plenty of prior knowledge to complete their training, thus making it an obstacle for GANs to get widely utilized. In many cases, it is hard to get access to sufficient data of high quality to meet a GAN’s training needs. Under such circumstances, effectively learning robust features from a few samples (or a single sample at an extreme case [12]) has become a crucial challenge. Besides many classical tasks [13], [14], [15], [16], recently proposed models [17], [18], [19] can accomplish image generation tasks efficiently by adopting hierarchical GAN pyramid structures [20], [21], [22], [23], [24], making it possible for unconditional GANs to generate various samples based on only one image. Nevertheless, such models treat different patches of an image equally regardless of their actual semantic information. Given ROI, the models mentioned tend to generate confusing results because they neither provide any interface for users to specify objects as ROI or background nor can they guarantee that the specific objects own proper structures in their generated results. Intuitively, these models produce multiform fake samples by choosing some random patches and applying “copy-shift-paste” operations, which are quite likely to destroy the rational structures of objects inside the sample.

To overcome such deficiency, we follow the same basic precondition (i.e., rely on only one natural image) and propose a novel morphologic-structure-aware GAN named MOGAN. Our target is to acquire samples that are abundant in appearance diversities while keeping the original structures of objects inside the sample correct. Besides the image, our model takes sets of coordinates specified by users as input to distinguish between ROI and background. Inside the model, we set up two parallel branches to generate ROI and background separately, which are both organized in a hierarchical way but own different characteristics. For the ROI branch, under the premise that no extra prior data is introduced, we propose a method...
that augment the original image into different forms and such augmented data can be used to learn ample knowledge of both correct structures and morphological patterns. We design a lightweight style extraction module to learn an affine transform from such data then act on the deep features produced by the generator, thus providing a guidance for the generation process. This module can be trained end-to-end along with the whole model. For the background branch, the sample for learning is the rest of the image excluding the given ROI. We apply a binary mask to the original image to shield pixels in the position of ROI and thereby switch the task of generating a complete image to generating an image with a mask, which reduces the difficulty of generation to a certain extent.

Finally, we analyze MOGAN’s capabilities of managing different tasks, including ROI-based random image generation, image editing, and single image animation. We test and compare MOGAN with other models in terms of generated results’ quality. By analyzing results qualitatively and quantitatively, this work shows that our proposed model achieves better performance than its peers. Moreover, we investigate the effects of different components on our method’s performance by conducting several ablation experiments.

In summary, this work aims to make two novel and significant contributions to the field of interactive image generation.

1) On the condition of a single image for training, we propose a novel method for generative models to gain morphologic diversities while maintaining correct structures. It utilizes the morphologic information coming from the augmented original image and we design a lightweight style injector to inject such knowledge to the model.

2) To well accomplish ROI-based image generation tasks (generating images according to regions users are interested in), we introduce a novel model with parallel branches to handle the concurrent and separate generation of ROI and background. In this way we manage to generate various realistic images with only one image to learn from.

In addition, this work analyzes MOGAN’s abilities of managing different interactive image generation tasks, such as generating random samples based on a single image, image editing, and animation. Experiments are performed to validate that MOGAN achieves better performance on the quality of generated samples than its peers. Examples of MOGAN’s generated results are shown in Fig. 1.

The remainder of this article is organized as follows. Section II review related work. In Section III, we first make a short discussion about existing methods, then we introduce our method in detail. The experimental results are presented in Section IV, and the conclusion is drawn in Section V.

II. RELATED WORK

A. GANs for Image Processing

Deep learning methods have excellent performances in image generation and feature extraction [25], [26], [27]. Goodfellow et al. [1] proposed a GAN based on the idea of a zero-sum game. GAN has become an important branch in the area of deep learning. It has a wide range of applications in the field of image processing [28], [29], [30], [31], [32], [33]. Image generation is the most common one. For image generation tasks, the characteristic of GANs that generating based on noise makes the GANs’ generated results diverse. More and more techniques and tricks have also been proposed to enhance the stability of GANs’ training process [34], [35], [36], [37].

However, most GANs used for image generation tasks rely on specific datasets or pretrained models [20], [38].Getting GANs well-trained on specific datasets places restrictions on GANs’ flexibility for different tasks. Instead of extracting the common features of different images, this work focuses on the information contained in a single natural image.

B. Single-Image GANs

In recent years, researchers have used the information contained in a single image to build deep learning models, thereby solving the problem of inadequate training data in some cases. InGAN [39] represents the first such model that applies GANs to the processing of a single image. However, it is a conditional generative model. Its generation process requires specific images as input (i.e., mapping the image to an image), thus leading to its poor generalization and failure to generate images randomly. To overcome its drawbacks, Shaham et al. [17] proposed SinGAN. It realizes random generation based on a single image by using an unconditional generation model (i.e., mapping from random noise to an image). Hence, it is suitable for many different image processing tasks. Then ConSinGAN [18] was proposed with a series of improved techniques for training single-image unconditional GANs. HP-VAE-GAN [19] was designed for single-image video generation. But the last three models [17], [18], [19] share serious defects like excessive randomness and uncontrollability on ROI-based tasks. Compared with them, our proposed model can unconditionally generate random samples with correct structures and variable appearances based on a single image and users’ expectations. It broadens the application range of single image generation.

III. PROPOSED METHOD

Taking SinGAN [17] as an example, We first give a brief introduction to unconditional single-image GANs along with
Fig. 2. MOGAN contains two parallel hierarchical branches responsible for the generation of ROI and background. The ROI branch takes ROI cut from the original image as the training target while the background branch takes the original image with a binary mask standing for regions of background. Finally, the generated results produced from two branches can be fused into complete images which are of high quality.

a short discussion about limitations of existing methods. Then we introduce MOGAN in detail.

### A. SinGAN

SinGAN is a kind of unconditional GAN (generating samples from latent vectors). It is able to generate diverse samples from randomly sampled noise based on only one single natural image. Obeying the design of a hierarchical structure, it stacks several sub-GANs into a pyramid structure. All sub-GANs share exactly the same structure but not parameters. It takes a latent vector sampled from a Gaussian distribution along with the generated result from the previous one as input (by noting that the input of the first sub-GAN is the latent vector only). The training target of each sub-GAN is the original natural image downsampled to different sizes. Such a design makes the learning goal of the overall model gradually shift from small-scale samples which are rich in global structural information to large-scale samples which contain plenty of texture details, deepening the model’s comprehension of a given sample.

However, SinGAN’s generative process fails to distinguish between different areas or instances inside a given image. Qualitatively, the embodiment of the generated samples’ diversity seems like randomly choosing patches from the original image, copying, randomly shifting and then pasting. As a consequence, such a random copy-shift-paste type of generation is quite likely to break down an object’s structure, thereby resulting in possible irrational outcomes. In many interactive generation tasks, there are some regions in an image in which users are more interested (i.e., ROI). It is thus required for a method to generate random samples with as many changes as possible but no changes in the original semantic structures. For example, if a tree is ROI that interest users, the shape of its crown or bending angle of its trunk may change, while neither the crown nor the trunk should disappear, and the semantic structure that “the crown is above the trunk” should not change either. Such ROI-based tasks are hard for SinGAN to complete. Similarly, ConSinGAN [18] whose structure is almost exactly same with SinGAN also has poor ability upon ROI-based tasks mentioned above. HP-VAE-GAN [19] combines multiple VAEs [40] with SinGAN-like structures to achieve better performance on video-based tasks, which shares the same shortcomings as SinGAN.

### B. Proposed MOGAN

Motivated by the problems mentioned above, we introduce our MOGAN with its structure shown in Fig. 2.

**Parallel-Branch Architecture:** In our problem setting, the usable information includes a single natural image \( I \) and a set of coordinates \( ((x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}})) \) provided by users to mark the region to which they pay attention. Allowing the model to deal with ROI and background areas separately leads to the problem of disentanglement. Under the circumstance that the number of learnable samples is limited to one, many existing methods of disentanglement based on extra specific prior data [41], [42], [43] are not applicable. Therefore, we use two different latent vectors marked as \( Z_a \) and \( Z_b \) to be responsible for the generation of ROI and background in turn similarly to [44] and [45]. Since our requirements for the generation of ROI and background are often different, the structures in charge of generating ROI and background should be independent of each other such that special adjustments...
A hierarchical design similar to SinGAN’s [17]. To be more specific, we organize a number of sub-GANs marked as \( I^a \) according to the coordinates mentioned above, which is the learning target of the ROI branch. The semantic structure remains correct. The results to own sufficient morphological changes under the premise that the structure remains correct. The appearances controls the style of generation through affine transforms. (b) Background branch. Generators are built based on residual blocks mainly containing two gated convolution layers. Discriminators of both branches are Markovian discriminators. Details of residual blocks in ROI branch and background branch are shown in (c) and (d), respectively.

**ROI Branch**: For this branch, we expect the generated ROI to satisfy the semantic structure inside the augmented original image along with rational semantic structures inside the augmented original images. As a guidance, such augmented data not only shows feasible changing directions about appearance but also emphasizes the structure information for the generator. Thus, during the training of each GAN, we transform \( I^a \) into different forms through regular data-augmentation methods, including random flipping (vertical and horizontal), random rotation (clockwise and counterclockwise), random scaling, and random perspective transform, which are marked as \( \psi \) in a general way. For these methods, we use the official implementations of PyTorch [48]. We find that data augmentation methods, such as [49], which may be widely used in other coefficient of variation (CV) tasks do not work in our experiments. In order to extract useful information from \( \psi(I_n^a) \), we build an extra lightweight module named a style injector which takes \( \psi(I_n^a) \) as input and outputs weight \( w_n \) and bias \( b_n \). Marking the style injector for GAN as \( \Omega_n^a \), the process mentioned can be described as the following equation.

\[
[w_n, b_n] = \Omega_n^a(\psi(I_n^a)).
\]  

(1)

Next, we apply the learned affine transform on \( G_n^a \)'s original dataflow, guiding the generator toward an expected generation direction provided by \( \psi(I_n^a) \). The learned \( w_n \) will be multiplied...
TABLE I
MAIN TRAINING PROCESSES OF THE ROI BRANCH AND THE BACKGROUND BRANCH

Algorithm 1 The main training processes of a ROI branch. Hyperparameters includes: The scales of the pyramids, N.

Initialize:
- Crop and scale the image pyramids \( \{I_0, \ldots, I_N\} \).
- Sample N noises \( \{Z_0, \ldots, Z_N\} \).
- Organize the GAN pyramids \( \{GAN_0, \ldots, GAN_N\} \).
- Initialize the original image input \( I_N \) as 0.

Training:
for n in N scales do
- Use the randomly selected data augmentation method \( \varphi \) to augment \( I_n \).
- Train the sub-GAN \( GAN_n \) with \( Z_n, I_n, \varphi(I_n) \) as the inputs.
- \( \tilde{I}_n \leftarrow \text{The output of the trained generator inside the sub-GAN } GAN_n \).
end for

Algorithm 2 The main training processes of a background branch. Hyperparameters includes: The scales of the pyramids, N.

Initialize:
- Crop and scale the image pyramids \( \{I_0, \ldots, I_N\} \).
- Sample N noises \( \{Z_0, \ldots, Z_N\} \).
- Organize the GAN pyramids \( \{GAN_0, \ldots, GAN_N\} \).
- Initialize the original image input \( I_N \) as 0.

Training:
for n in N scales do
- Train the sub-GAN \( GAN_n \) with \( Z_n, \tilde{I}_n \) as the inputs.
- \( \tilde{I}_n \leftarrow \text{The output of the trained generator inside the sub-GAN } GAN_n \).
end for

Fig. 4. Details of a style injector. It is a lightweight encoder essentially, which contains two bypasses producing a weight and a bias, respectively. Taking the augmented original image as the input, it controls the changing direction of the generator’s dataflow through affine transforms.

The entire generation process can be expressed as

\[
\tilde{I}_n = \begin{cases} 
G_n^\omega(Z_n, \varphi(I_n^\omega)), & n = N \\
G_n^\omega(Z_n^\varphi, \varphi(I_n^\varphi), \{I_{n+1}^{\text{upsample}}\}), & n < N.
\end{cases}
\]  

(2)

Now that we have brought in more information which may not only lead to better training results but confuse the generator as well. We then consider a more robust design for \( G_n^\omega \).

Moreover, we have noticed relative methods, such as [57] and [58], which augment the input data of the discriminator. We next explain the major difference between our method and theirs. In both [57] and [58], the main obstacle to conquer is the lack of training data, which leads to the severe overfitting of the discriminator and the bad quality of generated samples (e.g., artifacts) in the end. So these methods apply data-augmentation to the discriminator’s inputs in order to emphasize the same label information (real or fake) over different patterns of samples, just like other CV tasks, such as image classification. However, in our case, we aim to fetch the rich morphological information from the augmented data instead of the labels. According to methods like AdaIN in [20] and [59], such information can be properly expressed in the form of affine transforms. Hence, we believe that it is reasonable and essential to set a model to extract such affine transforms.
of AdaIN-style methods is their better interpretability and controllability, i.e., we can change the values of the learned affine transforms to change the generated samples directly, while diff-augmentation like [57] and [58] cannot.

As for other structure and training details of the ROI branch, $D^n_r$ is a Markovian discriminator [8], [60] with a fully convolution structure. For the discriminators, besides the traditional adversarial loss marked as $L_{GAN}$, we use WGAN-GP loss [61] marked as $L_{WGAN-GP}$ to stabilize $D^n_r$’s training process. The traditional GAN loss can be expressed as

$$L_{GAN}(G, D) = \log[D(x)] + \log[1 - D(G(z))].$$  \hspace{1cm} (3)

WGAN-GP loss can be expressed as

$$L_{WGAN-GP} = \left( \left\| \nabla_{\tilde{I}} D(\tilde{I}) \right\| - 1 \right)^2.$$  \hspace{1cm} (4)

The final loss function for the discriminators can be expressed as

$$L_D = L_{GAN}(G^n_r, D^n_r) + \lambda L_{WGAN-GP}(D^n_r).$$  \hspace{1cm} (5)

For $G^n_b$, besides traditional adversarial loss, we choose mean squared error (MSE) and cosine distance together as a loss function. They measure the similarity between $\tilde{I}_a$ and $I^n_b$ in different aspects: cosine distance marked as $L_1$ emphasizes the coherence of global direction and may tolerate local structure difference to some extent. It can be described as

$$L_1 = 1 - \cos(\tilde{I}_a, I^n_b).$$  \hspace{1cm} (6)

MSE marked as $L_2$ requires pixel-level consistency to restrain $G^n_b$ from generating bad texture futures. It can be described as

$$L_2 = \left\| \tilde{I}_a - I^n_b \right\|^2.$$  \hspace{1cm} (7)

The final loss function for the generators can be expressed as

$$L_G = L_{GAN}(G^n_r, D^n_r) + \alpha L_1(G^n_b) + \beta L_2(G^n_b).$$  \hspace{1cm} (8)

Background Branch: Actually, background regions of different samples may vary considerably in complexity. In other words, some samples may own extremely simple background even pure color while others may contain various disparate instances. Since users are less interested in the background, we expect background generation to maintain the global uniformity primarily while making changes in a few local patches, thus no need to introduce extra modules like a style injector into background generation. It does not mean it is unnecessary for the generated background to change. For some tasks like image-manipulating task, we believe that only changes in ROI are needed. However, ROI-based tasks also include certain tasks like the data augmentation task where changes of the background are expected: samples with diversity in both ROI and background are more beneficial for the following CV tasks than those with diversity in only the ROI part. We aim to make our method more general for ROI-based tasks and we believe our “two-branch” idea gives a more proper way.

To isolate the influence of ROI, we apply a binary mask to $I$ and mark the result as $I_b$. Similar to methods of ROI, we organize another GAN pyramid [GAN$^b_0$, ... , GAN$^b_N$] along with an image pyramid [$I^b_0$, ..., $I^b_N$], where $I^b_n$ becomes the training target of GAN$^b_n$. Note that because areas inside the mask will eventually replaced by the generated results of the ROI branch, the generated results of the background branch just need to hold the origin mask and make changes in the rest part of the sample. The general training process of the background branch is shown in Table I. Similar to the ROI branch, the generation process can be described as

$$\tilde{I}_n^b = \left\{ \begin{array}{ll}
G^n_b(Z_n), \\
G^n_b(Z_n, (I^n_b \uparrow \text{upsample})), & n < N
\end{array} \right\} \quad (n = N). \quad (9)
$$

For the generator in GAN$^b_n$ marked as $G^n_b$, we adopt a similar residual-type of design as mentioned in the ROI’s branch but replace all the convolution layers and deformable convolution layers with gated convolution ones [62]. This enables the model to learn the soft mask while learning the pixels outside the mask. Similarly, we still use combinations of InstanceNorm-LeakyReLU layers. Other training details, including loss functions and structures of discriminators, are as same as the ROI branch’s. Details of GAN$^b_n$ are shown in Fig. 3(b) and (d). Note that “GConv” means a gated convolution layer, “CA” means a channelwise attention layer, “IN” means an instance normalization layer and “LR” means a leaky-ReLU layer.

Training Details: For both ROI branch and background branches, we use the Adam optimizer [63] with $\beta_1 = 0$ and $\beta_2 = 0.99$. The learning rate of every generator and discriminator in both branches is set to 0.0003. For the loss function of all generators in the ROI branch, $\alpha$ is set to 50 but $\beta$ varies according to different scales: for coarse scales (such as scale $N$ and $N-1$), $\beta$ is set to 10; for other scales, $\beta$ is set to 5. As for generators in the background branch, $\alpha$ is set to 50 and $\beta$ is set to 10 for all scales. In both branches, $\lambda$ is set to 1 for all discriminators. We stack three ResBlocks in all generators and five convolution layers in all discriminators. We use the traditional padding way, which pads the feature map along with each convolution layer.

IV. EXPERIMENTAL RESULTS

We first explain MOGAN’s abilities of managing different ROI-based generation tasks along with revealing some of the results. Next, we compare the performance of our model with its peers’ qualitatively and quantitatively. Finally, we validate the effectiveness of our model’s components through an ablation study.

A. Applications

Generating Random Samples: To generate random samples from noise through training against a single natural image is one of the basic capacities of our model. Samples randomly generated by our model are listed in Figs. 1 and 5 which contain diverse kinds of images that our model has enough robustness against samples of different styles and topics. Qualitatively, the generated outcome of ROI has kept a reasonable structure by comparing it with the original image’s. In the meantime, ROI generated results get visible diversification on appearance, such as shape and posture. As for backgrounds, the generated results exhibit smooth changes in
local parts and retain the global layout similar to the original sample’s. Note that the model shows the prominent effects on samples that have more freedom degrees to change on appearance. As for samples with a solid structure which is hard to change, the effects seem to reveal in the aspects of postures.

It is worth noting that, we use the raw augmented data (not well pretrained models or any other specific prior knowledge
as required by some existing methods [9]) to enhance structure awareness. However, the raw data contains strong noise as described in Section III-B. Injecting such noise into the model may disturb the training and cause blur results. As a result, intuitively, we may find artifacts in the generated results. It is unavoidable to an extent and a kind of tradeoff in our opinions. To address this issue, we reason that it would help if limiting the effects of the style injecting, e.g., multiplying factors on \( w \) and \( b \) produced by the style injector before they are applied to the model. Specifically, \( w \) and \( b \) are a set of coefficients for affine transformations. Given a set of affine transformation coefficients \( (w, b) \), given an input \( x \), the output \( y \) can be obtained by \( y = w \ast x + b \). In our method, \( w \) and \( b \) are learned by the style extractor, a neural network module, in the end-to-end training process of the whole network. Specifically in Fig. 3, the affine transform block has the following changes: the tensor \( x \) from the previous block is transformed \( (w, b) \) from the output of the upper style extractor (orange block) in the affine transform block to output \( y \), which is passed to the next block. Therefore, it is not difficult to think that if \( w \) is limited near 1 and \( b \) is limited near 0, the value of the output \( y \) will be similar to the value of the input \( x \), thus limiting the effect of affine transformation coefficient \( (w, b) \) to a certain extent. Another helpful way to limit the effects of the style injecting is setting more ResBlocks between every two affine transforms.

Moreover, Samples which our method may not handle well are those samples whose ROIs’ shapes are too regular to make morphological changes. On such samples, some of the augmentation methods can only produce few changes. For these samples, the discriminator is very likely to overfit on the real sample and cannot produce efficient grads for the generator to update, thus resulting in artifacts and blur. On the contrary, when the semantic structures of the samples come to be more complex, then this problem can be mitigated.

**Editing:** To select some of the patches and paste them onto the other location of the original image, then output a harmonious result. We notice the target of image editing tasks exactly fits the capacity of our model. To perform editing, we take the image pasted with edited patches as the input of a style injector. In this way, the edited information works similarly to the augmented data. For more details, we first train an MOGAN’s ROI branch against the original image without the edited patches. Then we freeze all trainable parameters of the model, input the image with edited patches to the style injector and start a forward process of the model. In this way, we expect the generated samples to retain reasonable structures and smooth texture in the blending area of the raw edited sample.

**Single Image Animation:** To generate a short video based on a single image, which is an extension of the ability of random image generation. For MOGAN, after being well-trained on a certain sample, we fix an input noise and gradually modify the effect of data-augmentation. For example, we adjust the angle of rotation with the stride of 0.5° from the original pose to rotating 30° clockwise, and then we get 60 generated samples which are produced by a gradually changing data-augmentation method. The samples themselves are changing gradually, too. In this way, we can obtain a number of generated results that change smoothly and organize them into the form of video.

**Data Augmentation:** To generate more images based on a given image, requiring newly generated images share the same label with the given image but have lots of diversities. The proposed method is fully competent for data enhancement of other computer downstream visual tasks. In principle, the new images produced by our method can retain exactly the same semantic information as the provided samples, which means that the countless new samples generated have the same label information as the provided single samples; The new samples produced at the same time will be sufficiently diverse. This is fully in line with the requirement of data enhancement task: data expansion based on given data, so that the expanded new data has constant label information and richer representation.

**B. Comparison**

We make a comparison among the state-of-the-art models based on a single image. Figs. 6 and 7 show the
the 100 generated samples, we first calculate the CV of the deviation value by the mean of the samples. Specifically, over the CV, which is calculated via dividing the standard deviation value by the mean of the samples. Moreover, we calculate the diversity following the method of MOGAN are more vivid on the edges of the edited patch than others. The texture of MOGAN's generated results is finer than its competitive peers'.

Comparison results of randomly generating tasks and editing tasks, respectively.

Qualitatively, for the randomly generating task, SinGAN and other related models have the similarity in treating samples as a whole. They tend to equally deal with all objects inside as analogical patches regardless of their different importance and semantic information, which results in blurry and ambiguous results. But our model manages the problem very well by setting two parallel branches and generating ROI and the background separately. Next, we take samples that only contain ROI as the training targets and conduct experiments again. From the results, we can find that outcomes of the other models always get stuck in two situations: 1) overfitting (excessively similar to the training target) or 2) meaningless (structure of objects being destroyed). The basic reason of such results is that such models lack a proper guidance for added noises. Restrained by MSE, the noise in the end either affect the texture slightly, which makes the result seem like overfitting, or affect the structure which deforms the related objects too much. On the contrary, our model preserves the original structure to a large extent while making various changes emerging on the object owing to the style injector. Qualitative experiments in this article are mostly conducted on the Unsplash dataset due to its high quality. Note that we have not compare the generated results of background because each sample in both datasets has a clear topic and is open-source. They are more suitable for ROI-based tasks than the Places365 dataset and Berkeley Segmentation dataset which are claimed and used in the SinGAN paper, because each sample in both datasets has a clear topic and is easy to assign ROIs. Scores of different models are recorded in Table II. Coinciding with the qualitative analysis, our model has achieved better performance than its peers given the same samples. When trained against the whole image, the other three models get higher-diversity scores than ours while their SIFID and GQI are lower due to their chaotically generated results. When trained against ROI only, SinGAN produces overfitting results, thus increasing its SIFID and GQI while reducing the diversity to a large extent. ConSinGAN and HP-VAE-GAN continue to generate meaningless images with high diversity but low quality. Our MOGAN’s results are both diverse and realistic. Besides, GQI goes up when training on ROI only because of the lower difficulty. For our model, the diversity is lower for the whole image owing to the globally similar backgrounds. Another quantitative experiment is performed by using AMT realness score as the metrics. We have conducted the AMT metric on two settings: 1) paired (real-versus-fake image pairs are shown) and 2) unpaired (either fake or real image is shown). 1) Paired (Real-Versus-Fake): presenting the test workers with 50 pairs of images, in each of which a fake image (generated by MOGAN) is shown against its real background. Another quantitative experiment is performed by using AMT realness score as the metrics. We have conducted the AMT metric on two settings: 1) paired (real-versus-fake image pairs are shown) and 2) unpaired (either fake or real image is shown). 1) Paired (Real-Versus-Fake): presenting the test workers with 50 pairs of images, in each of which a fake image (generated by MOGAN) is shown against its real training image for one second and 2) Unpaired (Either Real or Fake): presenting the test workers with a single image for one second, and asking them if it is fake. The total number

| Metrics            | SinGAN [17] | ConSinGAN [18] | HP-VAE-GAN [19] | MOGAN (Ours) |
|--------------------|-------------|----------------|-----------------|--------------|
| SIFID (whole)      | 0.72        | 0.63           | 0.61            | 0.22         |
| Diversity (whole)  | 0.42        | 0.49           | 0.51            | 0.20         |
| GQI (whole)        | 0.58        | 0.78           | 0.84            | 0.91         |
| SIFID (ROI-only)   | 0.19        | 0.59           | 0.56            | 0.11         |
| Diversity (ROI-only) | 0.21       | 0.51           | 0.50            | 0.39         |
| GQI (ROI-only)     | 1.11        | 0.86           | 0.89            | 3.55         |
TABLE III
SCORES OF AMT REALNESS TESTS ON SAMPLES GENERATED BY SinGAN [17], ConSinGAN [18], HP-VAE-GAN [19], AND Mogan

| Metrics               | SinGAN [17] | ConSinGAN [18] | HP-VAE-GAN [19] | Mogan (Ours) |
|-----------------------|-------------|----------------|----------------|-------------|
| AMT (whole, paired)   | 21.92±1.4%  | 23.07±1.4%     | 24.84±1.3%     | 31.18±1.2%  |
| AMT (whole, unpaired) | 32.33±1.1%  | 37.25±1.4%     | 35.70±1.1%     | 44.66±1.1%  |
| AMT (ROI, paired)     | 25.49±0.4%  | 30.62±0.5%     | 29.11±0.6%     | 35.90±0.7%  |
| AMT (ROI, unpaired)   | 38.35±0.4%  | 40.84±0.2%     | 41.32±0.3%     | 47.26±0.5%  |

Fig. 8. Results of an ablation study. It is clear that a style injector plays an important part in generating diversely. Deformable convolution and channel attention both make the results refiner. Gated convolution enables the model to handle data with masks.

of test images is 100. True and fake samples each account for half of them. For AMT scores, 50% means the best-generation qualities: fake samples are so realistic that the scores are close to random guessing. Obviously, the AMT scores conform to the qualitative results. The results of AMT scores can be seen in Table III.

Particularly, for SinGAN, users may start generating from a finer scale to get samples with less diversities but far more stable structures, which can be expressed as SinGAN’s low-diversity mode. However, for SinGAN, generated samples that come from a finer scale ONLY have diversity in the field of texture. We think that for SinGAN, there is a severe tradeoff between the diversity and stability of structures. For SinGAN, we can either get diverse but meaningless samples or get stable samples with changes sometimes quite hard to find. Compared with SinGAN’s low-diversity mode, our method succeeds in generating samples with both diverse shapes and different details of texture. More importantly, our results maintain stable and correct semantic structures.

The diffusion model [69], [70] is a paradigm that has received a lot of attention recently in the field of image generation. The diffusion model adopts the method of gradually adding noise and de-noising, and relies on a robust reconstruction loss to learn the representation contained in image samples, so it can generate fairly high-quality samples. There is also some recent work based on diffusion models, which can only rely on a single sample or even zero sample to complete image generation [71], [72]. But there are two essential differences between these approaches and ours: first, the above method adopts the adding noise and de-noising strategy of diffusion model to complete the training, and adopts the iterative method to gradually restore a clean sample from a random sampled noise or a mixed result of a sample and noise. This means that the above methods are multistep and uncontrollable in generating the image. Multistep process means that if the same sample is generated, the complexity of diffusion model method is several times that of single-step generation (such as unconditional GAN adopted in our method), which greatly reduces the efficiency of the whole pipeline when the sample size is large; Uncontrollability means that random noise is introduced in the forward process of diffusion model to provide image details, which means that these methods cannot guarantee the semantic information in the generated samples is completely consistent with the provided samples, and the uncontrollability problems caused by such randomness are completely solved in our method. Second, in the above methods, the methods to complete image generation based on single sample or zero sample completely rely on prior knowledge pretrained in a specific field, while our method really only relies on a single sample and does not require any other prior knowledge. Therefore, in summary, although the diffusion model has replaced GAN and become the SOTA in the current field of image generation, we believe that the method proposed by us is optimal in the task of rapid and efficient diverse-image generation without relying on other priors and only relying on single samples.

C. Ablation Study

To analyze our design’s impact on the generation process, we take SinGAN as the baseline and conduct ablation experiments on two branches separately. The qualitative results are described in Fig. 8 and quantitative ones are in Table IV. Quantitative results shown in the table indicate the same conclusion as Fig. 8.
For the ROI branch, improvements we make onto a structure include i) deformable convolution layer; ii) channel attention layer; and iii) style injector. The effects on the generated results after removing certain design methods can be seen in Fig. 8. Note that “-” means that the method is disabled and “removing deformable convolution layer” means replacing it with a full convolution layer. Obviously, the style injector plays an important role in generating diversely as models without it induce overfitting. Deformable convolution and channel attention layers both prevent the results from strong noise and stripes to a large extent, while the former plays a more effective role.

For the background branch, we introduced 1) gated convolution layers and 2) channel attention layers. We take away modules or methods above in turn and record the effects on the generated results in Fig. 8. Similarly, “removing gated convolution layer” means replacing it with a full convolution layer. The results suggest that without gated convolution layers, the model will treat the mask as an object of certain semantic information. Thus, gated convolution layers enable the model to handle masks which isolates background from ROI. Channel attention layers make the training more stable.

V. CONCLUSION

We have introduced MOGAN, an unconditional generative model to generate random samples based on only one natural image. The generation results from our model can maintain correct structures and exhibit plenty of diversity in appearance, which is the main improvement over other recent models. As demonstrated by our experiments, MOGAN can produce samples of high quality over different kinds of images. Our future work intends to handle samples with multiple ROIs, and to guide the generated results of ROI and background to cohere with each other. Another one is to embed some recent optimization methods [73], [74] into MOGAN to make it more powerful.

REFERENCES

[1] I. J. Goodfellow et al., “Generative adversarial nets,” in Proc. Conf. Workshop Neural Inf. Process. Syst., 2014, pp. 1–27.
[2] C. Zheng, T.-J. Cham, and J. Cai, “Pluralistic image completion,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 1438–1447.
[3] L. Zhao et al., “UCTGAN: Diverse image inpainting based on unsupervised cross-space translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5741–5750.
[4] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, “Contextual residual aggregation for ultra high-resolution image inpainting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 7508–7517.
[5] I. Anokhin et al., “High-resolution daytime translation without domain labels,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 7488–7497.
[6] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, “Cross-domain correspondence learning for exemplar-based image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5143–5153.
[7] T. Bhattacharjee, S. Kim, G. Vizier, and M. Salzmann, “DUNIT: Detection-based unsupervised image-to-image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 4787–4796.
[8] 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), 2017, pp. 5967–5976.
[9] A. Shocher et al., “Semantic pyramid for image generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 7455–7464.
[10] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, “StarGAN V2: Diverse image synthesis for multiple domains,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 8188–8197.
[11] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “MaskGAN: Towards diverse and interactive facial image manipulation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5549–5558.
[12] M. Zontak and M. Irani, “Internal statistics of a single natural image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2011, pp. 977–984.
[13] M. Zontak, I. Mosseri, and M. Irani, “Separating signal from noise using patch recurrence across scales,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp. 1195–1202.
[14] T. Michaeli and M. Irani, “Blind deblurring using internal patch recurrence,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 783–798.
[15] Y. Bahat and M. Irani, “Blind dehazing using internal patch recurrence,” in Proc. IEEE Int. Conf. Comput. Photography (ICCP), 2016, pp. 1–9.
[16] G. Freedman and R. Fattal, “Image and video upsampling from local self-examples,” ACM Trans. Graph., vol. 30, no. 2, pp. 1–12, 2011.
[17] T. R. Shaham, T. Dekel, and T. Michaeli, “SinGAN: Learning a generative model from a single natural image,” in Proc. IEEE/CVF Int. Conf. Conf. Comput. Vis. (ICCV), 2019, pp. 4569–4578.
[18] H. Hirn, M. Fisher, O. Wang, and S. Wermter, “Improved techniques for training single-image GANs,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., 2021, pp. 1300–1309.
[19] S. Gur, S. Benaim, and L. Wolf, “Hierarchical patch VAE-GAN: Generating diverse videos from a single sample,” in Proc. Conf. Workshop Neural Inf. Process. Syst., 2020, pp. 1–8.
[20] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4401–4410.
[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in Proc. Int. Conf. Learn. Rep., 2018, pp. 1–8.
[22] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional GANs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8798–8807.
Jinshu Chen received the B.S. degree in automation from Tongji University, Shanghai, China, in 2019, where he is currently pursuing the M.S. degree in control science and engineering. His research interests include computer vision, deep learning, and image processing.

Qi Kang (Senior Member, IEEE) received the B.S. degree in automatic control and the M.S. and Ph.D. degrees in control theory and control engineering from Tongji University, Shanghai, China, in 2002, 2005, and 2009, respectively. From 2007 to 2008, he was a Research Associate with the University of Illinois at Chicago, Chicago, IL, USA. From 2014 to 2015, he was a Visiting Scholar with the New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Professor with the Department of Control Science and Engineering and the Shanghai Institute of Intelligent Science and Technology, Tongji University. His interests are in swarm intelligence, evolutionary computation, machine learning, and intelligent control and optimization.

Dr. Kang served as the General Chair of the 19th IEEE International Conference on Networking, Sensing and Control (ICNSC 2022). He is currently an Associate Editor of IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS. He is the Secretary General of Shanghai Association for System Simulation.

Qihui Xu received the B.S. degree in automation from Tongji University, Shanghai, China, in 2020, where she is currently pursuing the M.S. degree in control science and engineering. Her research focuses on the use of deep learning technology for few-shot learning and image generation.

MengChu Zhou (Fellow, IEEE) received the B.S. degree in control engineering from Nanjing University of Science and Technology, Nanjing, China, in 1983, the M.S. degree in automatic control from Beijing Institute of Technology, Beijing, China, in 1986, and the Ph.D. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

He joined New Jersey Institute of Technology (NJIT), Newark, NJ, USA, in 1990, and has been a Distinguished Professor of Electrical and Computer Engineering since 2013. He has over 1200 publications, including 17 books, over 800 journal papers (over 650 in IEEE TRANSACTIONS), and 32 book chapters. He holds 31 patents and several pending ones. His recently coauthored books include Sustainable Manufacturing Systems: An Energy Perspective (IEEE Press/Wiley, Hoboken, NJ, USA, 2022) (with L. Li) and Supervisory Control and Scheduling of Resource Allocation Systems: Reachability Graph Perspective (IEEE Press/Wiley, Hoboken, NJ, USA, 2020) (with B. Huang). He has led or participated in over 60 research and education projects with total budget over $12M, funded by National Science Foundation, Department of Defense, NIST, New Jersey Science and Technology Commission, and industry. His research interests are in intelligent automation, machine learning, Petri nets, robotics, Internet of Things, big data, cloud/edge computing, and transportation and energy systems.

Dr. Zhou is a recipient of Excellence in Research Prize and Medal from NJIT, the Humboldt Research Award for U.S. Senior Scientists from Alexander von Humboldt Foundation, and the Franklin V. Taylor Memorial Award and the Norbert Wiener Award from IEEE SMC Society, the Computer-Integrated Manufacturing UNIVERSITY-LEAD Award from Society of Manufacturing Engineers, the Distinguished Service Award from IEEE Robotics and Automation Society, and the Edison Patent Award from the Research & Development Council of New Jersey. He has been among most highly cited scholars since 2012 and ranked top one in the field of engineering worldwide in 2012 by Web of Science. He was ranked #99 among the 2023 Top 1000 Scientists in Computer Science in the World, Research.com. He served as the Editor-in-Chief for IEEE/CAA Journal of Automatica Sinica, an Associate Editor for IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION, IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, and IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, and an Editor for IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING. He has served as a Guest Editor for many journals, including IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, and IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING. He is presently an Associate Editor of Research, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, and Frontiers of Information Technology & Electronic Engineering. He is the Founding Chair/Co-Chair of Technical Committee on AI-Based Smart Manufacturing Systems and Technical Committee on Humanized Crowd Computing of IEEE Systems, Man, and Cybernetics Society, the Technical Committee on Semiconductor Manufacturing Automation and Technical Committee on Digital Manufacturing and Human-Centered Automation of IEEE Robotics and Automation Society. He is also a member of IEEE TAB Periodicals Committee and Periodicals Review and Advisory Committee. He was the General Chair of IEEE Conference on Automation Science and Engineering, Washington, DC, USA, August 23–26, 2008, the General Co-Chair of 2003 IEEE International Conference on System, Man and Cybernetics (SMC), Washington, DC, USA, October 5–8, 2003 and 2001 IEEE International Conference on SMC, Bari, Italy, October 6–9, 2019, the Founding General Co-Chair of 2004 IEEE Int. Conf. on Networking, Sensing and Control, Taipei, March 21–23, 2004, and the General Chair of 2006 IEEE Int. Conf. on Networking, Sensing and Control, Ft. Lauderdale, FL, USA, April 23–25, 2006. He was the Program Chair of 2010 IEEE International Conference on Mechatronics and Automation, August 4–7, 2010, Xi‘an, China, the 1998 and 2001 IEEE International Conference on SMC and the 1997 IEEE International Conference on Emerging Technologies and Factory Automation. He is a Life Member of Chinese Association for Science and Technology-USA and served as its President in 1999. He is a Fellow of International Federation of Automatic Control, American Association for the Advancement of Science, Chinese Association of Automation, and National Academy of Inventors.