Diverse Single Image Generation with Controllable Global Structure

Sutharsan Mahendren*, Chamira U. S. Edussooriya*,#, Ranga Rodrigo*

*Department of Electronic and Telecommunication Engineering, University of Moratuwa, Sri Lanka
#Department of Electrical and Computer Engineering, Florida International University, Miami, FL, USA

Abstract

Image generation from a single image using generative adversarial networks is quite interesting due to the realism of generated images. However, recent approaches need improvement for such realistic and diverse image generation, when the global context of the image is important such as in face, animal, and architectural image generation. This is mainly due to the use of fewer convolutional layers for capturing the patch statistics and, thereby, not being able to capture global statistics well. The challenge, then, is to preserve the global structure, while retaining the diversity and quality of image generation. We solve this problem by using attention blocks at selected scales and feeding a random Gaussian blurred image to the discriminator for training. We use adversarial feedback to make the quality of the generation better. Our results are visually better than the state-of-the-art, particularly, in generating images that require global context. The diversity of our image generation, measured using the average standard deviation of pixels, is also better.

Keywords: Single image generation, generative adversarial networks, scale-wise attention, adversarial feedback.

1. Introduction

Generative Adversarial Networks (GANs) are successful in implicitly learning the underlying statistics of a large dataset and thus enable generating new samples from the same distribution [1], [2]. In such GANs, generating good-quality and diverse images needs a large image dataset. Recently, SinGAN [3] proposed a hierarchical learning-based approach for training with a single image. Here, in each scale, the generator and discriminator with low receptive fields learn to capture the internal statistics of the patch distribution of the image. One of the drawbacks of this method, as the paper itself states, is unrealistic image results when the global structure of the image is important, e.g., in face and animal image generation. The main reason for unrealistic results is the lack of global structure, when the images are generated starting from the coarsest scale. SinGAN can only generate the images without destroying the global structure by feeding the downsampled version of the real image in less-coarser scales instead of feeding noise to the coarsest scale. However, this tends to reduce the diversity in generation.

The level of information that must be captured for the global structure of the image to produce realistic looking results varies with the input image. For example, the generated samples from the images with small structures in the foreground (e.g., balloons in air, flocks of birds) and natural scenes (e.g., landscapes, foliage) do not need to maintain its original global structure. In contrast, the global structure plays a major role in realistic generation of images of large objects like faces and buildings.

Prior works [3, 4] in single image GAN domain depend on the receptive field of convolutional layers to incorporate local information to the network. Increasing the capacity of the network causes overfitting to the single sample and looses the diversity. Our main motivation is increasing the generation quality without degrading the diversity

In this paper, we propose two main strategies for inserting the global context to the network while maintaining the diversity of the image generation. First is using self-attention (SA) as a key to control the level of the insertion of information on the global structure for realistic generations of all type of images while not sacrificing much in the diversity. Second is using a random Gaussian kernel to convolve the real image before feeding it to the discriminator. This helps to improve the diversity in the generated images. Using these strategies we are able to generate diverse set of images starting from the coarsest scale with global context. The level of diversity in our results is significantly higher than the SinGAN, when we start the generation from less-coarse scales. In addition, to generate better reconstructions (in terms of Single Image Fréchet Inception Distance (SIFID)) we leverage on the concept of adversarial feedback in our single-image multi-scale network. Realistic images generated through our method are useful in tasks such as data augmentation [5, 6] and animation. Moreover, our method does not affect downstream tasks such as image editing, harmonization, and arbitrary size generation from the SinGAN architecture.

The main contributions of this work are: (1) proposing a method that retains the global structure in generated images by using SA, (2) increasing the diversity of generated images by depriving the discriminator of high frequency detail by simple random Gaussian smoothing.

(3) enabling the controllability of the global structure by carefully choosing where SA is used and by the standard deviation of the Gaussian kernel. (4) Using adversarial feedback from previous scales to generate better reconstructions in a single-image multi-scale network. To the best of our knowledge, this is the *first time*, feedback from the discriminator is utilized under a multi-scale architecture.

1.1. Related Work

The ability of a GAN to generate a sample from a distribution resembling data resulted in many contributions in GAN-based image generation [7], [2]. These GANs generate novel images by learning from a large database of images, that would have emanated from the same distribution. StackGAN [8] and ProgressiveGAN [9] use progressively growing architectures to improve the stability while generating high resolution images. BiGGAN [10] trains with large number of parameters and a large batch size on ImageNet dataset to attain high fidelity generation. Although GANs generate impressive results, the need for a large dataset, the resulting large training time, dataset specific nature of the generation are concerns.

As our objective in this work is single-image GAN generation, here we concentrate on **Deep Internal Learning**: Training a deep architecture with a single image for image specific tasks comes under deep internal learning. Deep Image Prior (DIP) [11] captures the image statistics through the structure of the generator network to perform image restoration tasks like denoising, inpainiting, and super-resolution. Here, the network learns to map random input to the single image sample. In this reconstruction process, DIP is able to recover the corrected version (denoised, inpainted) of the trained sample. Double DIP [12] extends this idea to decompose a single image into two with a task-specific mask and regularization. In segmentation, the image is decomposed into foreground and background with a binary mask. In image dehazing, the image is decomposed into an airlight map and haze free image with a transmission map. In ZSSR [13] an image specific CNN is learned to super resolve an image. It is trained with high and low resolution pairs from a single image. KernelGAN [14] proposes a method to extract a downsampling kernel from a deep linear convolutional generator and combine with ZSSR to super-resolve an image. [15] extends the ZSSR concepts to video domain for temporal super resolution tasks. These works establish that it is possible to learn from patches in a single image for multiple tasks.

Now, there are prominent approaches for using deep-internal learning to generate images from a single image, making the generation process much faster to learn. Spatial GAN uses [16] a fully convolutional generator and patch discriminator starting from 2D noise to generate an arbitrary sized texture image from single image. In [17], the generator generates larger images condition on small patches from the training image. Here, VGG [18] based perceptual loss is used alongside with adversarial training. [19] uses structural noise at three different levels at local, global, and periodic part at the generator with a patch discriminator. InGAN [20] learns to remap an image to different aspect ratios and sizes while maintaining the same patch distribution by training a GAN based architecture with a multi-scale patch discriminator and a generator with a non-parametric transformation layer which can also learn to remap the output to the input using an inverse transformation. These networks typically map images to images and are, therefore, constrained to a couple of tasks. SinGAN, on the other hand, extends this deep internal learning concept with a hierarchical architecture to train from single image for multiple purposes. Since SinGAN only trains with a single image, it only learns statistics of the patches under its lower receptive field through a few convolution layers. Having a deep architecture with a higher receptive field will easily memorize the trained image and generate less diverse outcomes. In our work, we aim to improve SinGAN by having a controllable global structure insertion while maintaining the diversity of generation.

Although regular GANs, generate diversified images due to the diversity of the training set, single-image GANs suffer from the lack of diversity. In prior work related to GANs on large training samples, loss of diversity in generation is considered as mode collapsing. [21] proposes mini batch discrimination to overcome mode collapse in training, where the similarity between intermediate features from discriminator for real and generated samples is used as additional information to discriminate the real from fake. However, in the single-image generation context, there is no way to incorporate this similarity. [22] proposes regularization to maximize the generator gradient with respect to the latent noise, specifically for conditional generation. While this may be of use, we opt to add the diversity through Gaussian smoothing at coarser scale without using an additional regularization loss. [23] uses latent noise from a mixture model with learnable means and sigmas to improve the diversity. This impose more complexity for other task than generation such as editing and harmonization.

Several prior works leverage different kinds of feedback—some with GANs—to improve the output results in deep learning. [24, 25, 26] use the output of the CNN as an input in an iterative manner to refine the results in different applications like instance segmentation, human pose estimation and medical image segmentation. [27] uses error feedback inside the feature space from back projection units from low to high and vice versa. [28] and [29] are first to explore the concept of feed back in GANs. They use discriminator feedback to improve the generation quality. [28] uses the intermediate features of the discriminator to modify the corresponding features in the generator. First the network is trained without the feedback modules and then the generator is fixed while discriminator and feedback modules are allowed to be updated. [29]

uses adaptive spatial transform layer to use the generated results from previous iteration and its discriminator scores to find the affine parameter to modify the encoded channel features in the generator at the next iteration. Both of these methods have not been explored with multi scale architectures and in the context of single image GAN. Instead of using the discriminator feedback to improve the results in the same scale, we pass it to the generator in the next scale to improve the results in most challenging regions.

Several works related to GANs show hierarchical training on multiple scales with different resolutions of images helping to achieve high quality, high resolution samples [30, 9, 31]. The use of patch discriminators, where they infer each small patches inside the generated image as fake or real, has been explored in [32, 33]. [34, 35, 36] incorporate attention in the vision models by computing features in channel or spatial dimension using the global pooling operation. In our work we use self attention for incorporating global structure through finding the similarities of pixels at different locations on the trained image. [37] proposes to use SA in both the generator and discriminator to capture the long range dependencies. The vision transformers apply global self-attention to full-sized images [38] using 16×16 patches. It uses multi-head self-attention blocks inside the transformer based architecture for classifying images.

Recently proposed SinGAN [3] uses a hierarchical unconditional GAN approach with single image for performing many tasks like image generation, super-resolution, editing and harmonization. ConSinGAN [4] proposes to concurrently train several scales in SinGAN to increase the conformity of generated images. PatchGenCN [39] explicitly models the internal distribution of patch statistics with hierarchical energy based models using a patch convolutional network. Above works contain the generator with only a stack of a few convolution layers which does not provide the global information at coarser scales. To improve the realism of generated images from more complex structures, ExSinGAN [40] uses external information through GAN inversion at coarsest scale from pretrained BigGAN [10] and perceptual loss from pretrained VGG-19 [18]. MOGAN [41] needs an additional input to specify the region of interest for performing foreground and background generation separately. HP-VAE-GAN [42] uses hierarchical patch variational autoencoder at coarser scales for diverse video and image generation. It also depends on the lower receptive fields of convolutional layers at coarser scales. Above works show limited performance in generating realistic images for which the global structure is important with higher generation quality without using external information, e.g., faces and buildings. External information mostly causes lesser visual diversity in the generation. The main reason for this limitation is that these networks do not have explicit controllable parameters to capture global information while training. Therefore, generating diverse images from single image that need better representation of the global structure still needs exploration.

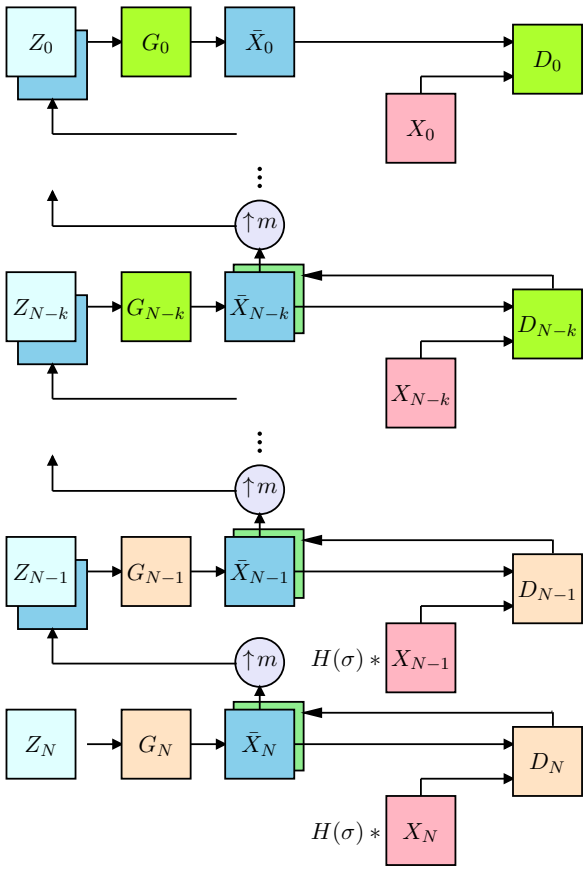


Figure 1: Overall structure of our single image generator. See sec. 2.1 for details on the general single-image generation mechanism. Here, G_0, D_0 to G_{N-k}, D_{N-k} are convolutional blocks *without SA*, and G_{N-k+1}, D_{N-k+1} to G_N, D_N are convolutional blocks *with SA*. Further, X_i, $\bar{X}_i$ and Z_i $(i = 0, 1, \ldots, N)$ are real image, generated image, and 2-D noise at scale i, respectively, and $\uparrow m$ denotes upsampling. $H(\sigma)$ is a Gaussian kernel. Notice that the scales with attention blocks receive Gaussian smooth image as real samples for the discriminator. These infuse global context to image generation.

2. Method

We describe the proposed system in detail in this section. Fig. 1 shows the overall architecture of our system, which generates diverse images based on the training on a *single image* while preserving the global structure. Our architecture contains several scales to train from coarsest to finer scales. In each scale, other than the coarsest one, the generator has the upsampled images generated from the previously trained scale, upsampled feedback from discriminator of previous scale, and noise as inputs. Here, each scale produces residual terms compared to the upsampled image from the previous scale. The generator at the coarser scale receive only 2D noise as an input. Our modification to the network architecture compared to SinGAN and ConSinGAN architectures is the introduction of SA blocks at the coarser scales in both the generator and discriminator and the use of adversarial feedback in all the

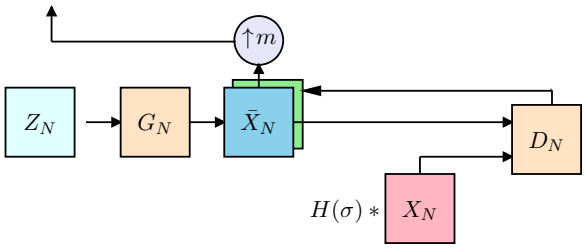


Figure 2: Adversarial feedback: the output from the discriminator of particular scale (e.g., D_N) is fed back through the upsampling block $\uparrow m$ to the generator G_{N-1} of the next scale.

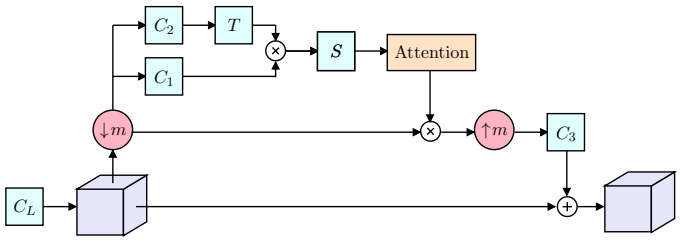


Figure 3: Self-attention block. This encourages the retention of the global structure in the generated image. Here, C_L denotes the previous convolution layer, C_1, C_2 and C_3 are convolution layers. Further, T denotes transpose and S denotes softmax, and $\downarrow m$ and $\uparrow m$ denote downsampling and upsampling, respectively. This self attention block is added to the layers in generator and discriminator at coarser scales which are depicted at Fig.1 in orange tone.

scales above the coarsest one. Because we have a hierarchical structure and a patch-based discriminator, we can generate images based on a single input image. Furthermore, we can preserve the global structure of the input image, essential for intelligibility, due to the SA block. Feeding the input image (of appropriate scale) convolved with a Gaussian to the discriminator enables us to achieve a good diversity in generated images. In summary, our single-image synthesizer is an unconditional GAN with a multiple-scale pipeline similar to SinGAN with a modified SA layers for the first few scales in both the discriminator (D) and generator (G). We use SA blocks to infuse global context into the image generation at selected scales.

To compensate for the reduction of diversity when we impose the long-range dependencies through SA blocks, we introduce a strategy to use Gaussian kernels with varying standard deviation (std.) to convolve the real image before feeding to D at coarser scales only. With this, we can achieve the desired diversity while maintaining the global structure, which remedies one of the major issues in the SinGAN-like architectures.

Fig. 2 (an excerpt of Fig. 1) shows how the adversarial feedback is used. The output from the discriminator of a particular scale (e.g., D_N) is fed back through the upsampling block $\uparrow m$ to the generator G_{N-1} of the next scale. In this fashion, we are able to achieve adversarial feedback in our multi-scale architecture. Here, we add noise only for the upsampled image from the previous scale. In summary, with the inclusion of attention blocks at selected scales and using Gaussian convolution in the discriminator image inputs, we are able to produce a diverse set of images, particularly valuable when global structure is important.

2.1. A Brief Background on Single Image Generation

To create better understanding, here we briefly discuss about single-image generations. For more information on single image generation, following are good reads: [3], [4] . The key idea to generating a generally similar but different image by learning from a single image is to learn the distribution of the patches within the image. In each scale above the coarsest scale, the generator produces the residual term conditioned on upsampled version of generated image from previous scale (see Fig. 8). When the residual from the current scale is added to the upsampled term from the previous term a larger and diverse image is generated.

2.2. Image Generation at Each Scale

The procedure of image generation is the same as SinGAN. Coarsest scale generation starts with random noise. The generated results of each scale is fed to the scale above after upsampling. These scales add noise (Z_{n-k}) to the upsampled generation and use them as a prior to generate the image from the current scale. We attach the upsampled adversarial feedback with it. We can also generate images starting from scales lesser than N. To do that we downsample the original image to the size of the corresponding scale and feed that with added noise. In this situation adversarial feedback is computed from the scale above with the downsampled original image.

2.3. Self Attention (SA) Block

The SA block, shown in Fig. 3, is the important component that infuses global context to our image generation in addition to the convolution features which help to capture the information in the local neighborhood. The advantage of the SA block is that it implicitly increases the receptive field of the features using the attention map which is computed considering all the features in space. It mitigates the need for a higher number of convolutional layers followed by a downsampling layers to increase the receptive field. In hierarchical architectures, the coarser scales play the major role in capturing the global structure of the large objects inside the image. It is a much harder goal to achieve only by using stack of convolutional layers. Although adding self attention blocks helps to capture the global structure, it introduces small local artifacts in the generations. Due to this reason, we use the self-attention blocks in the coarser scales only. Rest of the scales in the GAN having only the convolution layers help to remove the noise and add more local details to the generation.

Unlike the implementation in [37], we compute the SA features from the downsampled version of features from the last convolutional layer. In downsampling map the features to the predefined fixed output size $m \times m$,

$m \in [8, 16, 32]$ irrespective of the input. Instead of using additional convolutions to compute the value features, we directly pass the downsampled features to be modified by the attention. Here, each feature is a weighted sum based upon its similarity with others in key and query features with an optional channel reduction. SA features are directly added to the convolutional feature without having a learnable parameter like in SAGAN [37]. Experiments show this form is enough to capture the required global structure to pass it to the next scales.

2.4. Increasing the Diversity with Gaussian Smoothing with Kernel with Random Standard Deviation

Adding SA to a greater number of scales capture more and more long-term dependencies inside a single image. However, it reduces the diversity in the generation. Depending on the nature of the image, particularly the comparative size of the global structure (e.g., some faces, Eiffel tower image), we need to have SA in more scales than for others. Then, the final realistic outcomes become less diverse. It is hard to tune other parameters like the level of downsampling size or the positions of the SA block inside each scale to balance the trade-off between realistic generation and diversity. SinGAN uses the same downsampled version of the real image to be fed to D depict the real sample for discriminator. It is one of the main reasons why SinGAN architecture has to maintain a lower receptive field with lesser layers to reduce the overfitting and loss in diversity. To solve this problem, we propose convolving with a Gaussian which can maintain both image quality and diversity: Instead of feeding the same image as the real sample, we feed it after convolving with a Gaussian kernel with random value for the std. sampled from a uniform distribution between a predefined min. and max. value. For the computation of the reconstruction loss, we keep a fixed std. value for generating the real sample.

2.5. Adversarial Feedback

As the discriminator too holds a significant amount of information on the distribution of patches, making the generator aware of the discriminator's spatial information improves the reconstructions quality [28].

To achieve this feedback information transfer, we concatenate the discriminator's score for each patch from the previous scale with the generator input for the current scale. Generators at higher layers (above the coarsest scale) only generates the residual images from the upsampled generation of the previous scale. Generation at the coarsest scale does not depend on adversarial feedback because it only depends on the noise that we feed. We connect the feedback to the scales above it. Diversity of the generations highly depends on the variations in the coarsest scale. Scales above the coarsest scale help to add more details on its upsampled versions by generating the residuals. In the higher scales we do not use self attention. So the concatenated feedback in higher scales only affect the results in the local neighbourhood because of the few convolutional layers. Therefore, in our approach, feedback does not degrade the diversity significantly while improving the generation quality.

2.6. Loss

Eq. 1 shows the original GAN loss used in Goodfellow *et al.* [1].

$$\min_G \max_D \mathcal{L}_{\text{adv}}(G, D) = \mathbb{E}_{x \sim P_r}[\log(D(x)] + \mathbb{E}_{z \sim P_z}[\log(1 - D(G(z)))] \tag{1}$$

Here, $D(x)$ is output of discriminator for real images which aims to give a probabilistic score for x that belongs to the real data distribution, $G(z)$ is the fake image from the random vector z. Typically, several iterations of interchangeable generator and discriminator optimizing happen until they reach Nash equilibrium. This leads to the generator fully approximating the the distribution of data P_r and discriminator being unable to differentiate between the two distributions. See Goodfellow *et al.* [1] for more details.

$$W(P_r, P_g) = \inf_{\gamma \in \pi(P_r, P_g)} \mathbb{E}_{(x,y) \in \gamma}[\|x - y\|] \tag{2}$$

$$W(P_r, P_g) = \max_{w \in W} \mathbb{E}_{x \in P_r}[f_w(x)] - \mathbb{E}_{x \in P_g}[f_w(x)] \tag{3}$$

When the discriminator is optimal, the original GAN loss (Eq. 1) relates to the Jensen–Shannon (JS) divergence between the real and fake distributions. Since real and fake distributions mostly lie in a lower dimensional manifold, it may contains non overlapping regions. This scenario make the discriminator to learn separate real and fake images easily and gradients of JS divergence become very smaller for the generator. To overcome this issue WGAN [43] proposes Wasserstein distance instead of using JS divergence bases loss function. Wasserstein distance is defined as the minimum of effort to move from one distribution to another among a possible joint distributions which have the marginals as real and fake distributions (P_r, P_g) are the options for each transport plan as in Eq. 2. WGAN [43] authors use Kantorovich-Rubinstein duality to evaluate the Wasserstein distance by formulating the discriminator as a parameterized family of functions $(f_w(x))$ with the k-Lipschitz constraint which maximizes the difference between the expected socres for real and fake images by updating the discriminator weights as in Eq. 3. [44] introduce gradient penalty loss term to impose the 1-Lipschitz constraint instead of using weight clipping in [43].

We use the WGAN-GP [43] [44] for the adversarial loss and reconstruction loss to make the network generate the real samples from a particular fixed sample of noise at each scale, modified to accommodate the convolution with the

Gaussian $H(\cdot)$. For each scale (see Fig. 1) $n(< N)$

$$L = \min_{G_n} \max_{D_n} \mathcal{L}_{\text{adv}}(G_n, D_n) + \alpha \mathcal{L}_{\text{rec}}(G_n)$$
$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\vec{x}_f \sim P_g}[D(\tilde{\vec{x}})] - \mathbb{E}_{\vec{x}_r \sim P_h(\sigma_1,\sigma_2)}[D(\tilde{\vec{x}})] + \lambda \mathbb{E}_{\hat{\vec{x}} \sim P_{\hat{\vec{x}}}} \left[(\|\nabla_{\hat{\vec{x}}} D(\hat{\vec{x}})\|_2 - 1)^2 \right]$$
$$\mathcal{L}_{\text{rec}} = \|G_n(0, (\vec{x}_{n+1}^{\text{rec}} \uparrow^r) - \tilde{\vec{x}}_n * H(\sigma_3)\|^2$$

where P_h is the distribution of images convolved with a Gaussian, $\vec{x} * H(\sigma)$ and $\sigma \sim \text{U}(\sigma_1, \sigma_2)$ and $\sigma_3 \in [\sigma_1, \sigma_2]$.The convolution with the Gaussian $H(.)$ is for increasing the diversity of generation (while maintain the global structure). Note that convolution with $H(\cdot)$. is not used for scales without the attention block. See Sec. 2.4 for more details.

2.7. Selection of Parameters

Several parameters select the insertion level of global information. (1) No. of scales with SA starting from the coarsest scale (k). With this, SA is on scales N to $N-k+1$. (2) Choices of layers in each G scale to add SA. (3) Max. and min. value for std. for the Gaussian kernel σ_1 and σ_2. (4) Output size after downsampling with a factor m inside the SA blocks. Items (1) (2) and (3) are impactful. The default choice for (4) is 16 which worked well for all the images that we tested. (1) and (2) directly control the global structure in the image generation.

3. Experimental Results

We carried out experiments to show 1. how the attention blocks and Gaussian smoothing of the input to the discriminator generate high-quality diverse images, 2. the effect of the hyper-parameters (k and σ), 3. impact of using Gaussian smoothing only, 4. impact of using adversarial feedback only, 5. overall impact of using feedback, self-attention, and Gaussian smoothing together comparing to SinGAN and ConSinGAN 6. how our system can perform editing, harmonization, and arbitrary-sized generation.

We keep the LR of the generator (G) and discriminator (D) at 0.0001 and train for 6000 epochs with updating G and D one time in each epoch. We use 10 for α as a weight for reconstruction loss. In each epoch, G and D are updated with the loss on a single real and fake pair. In our experiments we use instance normalization [45] instead of using batch normalization [46] as in SinGAN. We keep this configuration as baseline for SinGAN and our approaches. We also analyze ConSinGAN with our choice of parameters as mentioned above.

3.1. Impact of Self Attention Blocks and Gaussian Smoothing in Single Image Generation

Here we qualitatively and quantitatively explain how the self attention helps to increase the generated image quality alongside with Gaussian smoothing for increasing the diversity from single image. Fig. 4 compares our results with SinGAN. Top five rows show images that need the global structure to be realistic. In column c and d we show our results with its hyper-parameters. We increase the number of scales with SA standard deviation values from column c to column d, and add self-attention to first four layers inside G and D for the results in column d. Note that our results in column c is visually better than SinGAN for the image in the first three rows. For the images in the 4th and 5th row, our results in column d performs better as these images required another SA layer to recover the global structure and a higher σ to maintain the diversity. Last three rows show images that do not need the global structure to be realistic. In column g where we use SA in the coarsest scale only. Even in this scenario our method produces diverse images on par with SinGAN. However, while we increase the number of scales with SA to 3 in column h, diversity becomes low as the constraint on global structure becomes higher.

Table 1: Average SIFID score (lower the better) of generated images from SinGAN [3] and ours starting with scale N and $N-1$.

Generation starting scale	SinGAN	Ours
N	0.02371	0.01828
$N-1$	0.01396	0.01120

We explain how diversity and image quality vary between SinGAN and ours. Diversity is computed among 50 generated images by as the channel wise std. of each pixel and averaging. Fig. 5 compares diversity of images which are generated from scale N to scale $N-7$. We use SA blocks in first 4 scales. These scales are trained with random Gaussian blurred input to maintain diversity compared to the corresponding scales in SinGAN. SinGAN is able to generate images with global context if it starts the generation from scale $N-1$ or above, but it looses the diversity severely compared to the generation from coarsest scale with random noise. So, in SinGAN the maximum achievable diversity with global context can only be attained at scale $N-1$. In all above images the diversity of generated images of SinGAN from scale $N-1$ is lower than the diversity of images with global contexts which are generated by our method from scale N and even in scale $N-1$. This validates that our method is able to produce images with global context without loosing diversity as in SinGAN. We use Single Image Frechet Inception Distance (SIFID) as proposed in SinGAN to compare the quality of generated samples from each image. It uses statistics of features from 2nd layer of the inception network [47]. It compares generated single image with real one by calculating distance between distribution which is created by features before the second pooling layer of Inception Network. It is low when the generated image quality is same as the real image. Table 1 shows the average SIFID for the images which need global structure in column d in Fig. 4.

(a) Input (b) SinGAN [3] (c) $k = 3, \sigma_c$ (d) $k = 4, \sigma_d$

(e) Input (f) SinGAN [3] (g) $k = 1, \sigma_g$ (h) $k = 3, \sigma_h$

Figure 4: Here we compare the baseline with only using Gaussian smoothing and attention. Attention retains the global structure. Attentions needs to couple with Gaussian smoothing to have diversity in results. (a) to (d): for images that needs to maintain the global structure. (e) to (h): for images that do not need to maintain the global structure. (c), (d), (g) and (h): our results with two sets of hyper-parameters. $\sigma_c \in [1, 3]$, $\sigma_d, \sigma_g, \sigma_h \in [3, 7]$. Notice that our method retains the global structure better than SinGAN [3] for the images which require global structure for its realistic look.

Our method has lower scores in generations from scale N and $N-1$. This shows that our method can generate high quality samples compared to SinGAN in the context of images which need global structure. In summary, the low SIFID and high diversity show that our results are of better quality than SinGAN.

3.2. Selection of Hyper-Parameters

Fig. 7 show pairs of training and fake images at the end of training at each scale. Column a shows the results of using SA block in first three scales. This G is not able to capture the global structure from the scales with self attention. Here the self attention blocks are added to first three coarser scales $[N, N-1, N-2]$. So, the generations after this scale are not able to preserve the global structure. In the next step, we add SA to scale $N-3$ and to compensate for the reduction of diversity, we increase the standard deviationσ range from $[0.5, 1]$ to $[3, 7]$. It assists G to capture the global structure fully, which makes the following generations realistic.

Table 2: Comparison of SinGAN [3] baseline with our proposed improvements in terms of SIFID (lower the better). SinGAN-G: SinGAN with Gaussian smoothing the input to the discriminator. SinGAN-F: SinGAN with adversarial feedback. Adversarial feedback only, and Gaussian smoothing only improve the generation quality in most of the image.

Image	SinGAN	SinGAN + G	SinGAN + F
balloons	0.049	0.053	**0.041**
birds	0.029	0.022	**0.021**
Colosseum	0.044	**0.038**	0.038
cows	**0.037**	0.047	0.063
Eiffel	0.041	**0.031**	0.032
Elon Musk	0.014	**0.005**	0.007
face	0.031	0.018	**0.016**
kid	**0.022**	0.037	0.025
lightning1	0.065	0.043	**0.031**
Mona Lisa	0.03	0.047	**0.026**
mountains	0.143	0.038	**0.034**
tree	0.027	0.015	**0.012**
Mean SIFID	0.044	0.033	**0.029**

3.3. Impact of Using Gaussian Kernel with Random Standard Deviation:

While using self attention (without random Gaussian blur) the system captures the portion of the image very well from the coarsest scale due to the SA blocks and keeps that portion unchanged in generation. Random Gaussian blur mitigates this issue and adds more diversity. Intermediate results are shown in Fig. 8 for two training samples. First row shows the images from the generator output (residual term) and second row shows the upsampled images from previous scale. Here, self attention blocks with random Gaussian augmentation is applied to first 3 scales. In scale 3, the network with a stack of convolution layers with small receptive field is able to learn the remaining high frequency residual term according to its low frequency input from the previous scale. Fig. 9 shows the impact of using the Gaussian smoothed input (see Fig. 1) in generating images while maintaining the global structure. The figure shows both the sale-0 generation (small images) and the last scale (large images, scale-8). Column 1 is with no smoothing, column 2 is with $\sigma \in [1, 3]$ and column 3 is with $\sigma \in [3, 7]$. All the three experiments use self-attention in

Table 3: SIFID measured (lower the better). Ours outperforms SinGAN and ConSinGAN all except in one.

Image	SinGAN	ConSinGAN Def. Param.	ConSinGAN 6000	ConSinGAN lr_scale_0.5	Ours
balloons	0.049	0.167	0.282	0.104	**0.042**
birds	0.029	0.112	0.221	0.131	**0.014**
Colosseum	0.044	0.103	0.144	0.056	**0.029**
cows	0.037	0.125	0.097	0.126	**0.022**
Eiffel	0.041	0.139	0.047	0.033	**0.027**
Elon Musk	0.014	0.031	0.03	0.023	**0.004**
face	0.031	0.040	0.045	0.032	**0.012**
kid	**0.022**	0.104	0.113	0.092	0.029
lightning1	0.065	0.111	0.123	0.086	**0.024**
Mona Lisa	0.030	0.139	0.099	0.141	**0.012**
mountains	0.143	0.154	0.146	0.096	**0.037**
tree	0.027	0.059	0.079	0.045	**0.018**
Average	0.044	0.107	0.119	0.08	**0.023**

Table 4: Diversity measured using the average standard deviation of pixel. Note that the diversity in our work is better when compared with ConSinGAN with the same hyper-parameters needed for maintaining the global structure. Here the diversity values are normalized with mean of pixel values in each image.

Image	SinGAN	ConSinGAN Def. Param.	ConSinGAN 6000	ConSinGAN lr_scale_0.5	Ours
balloons	0.591	**0.628**	0.418	0.584	0.576
birds	**0.35**	0.346	0.205	0.224	0.285
Colosseum	**0.834**	0.803	0.568	0.455	0.777
cows	**0.834**	0.764	0.602	0.75	0.779
Eiffel	0.42	**0.487**	0.208	0.203	0.370
Elon Musk	0.216	**0.254**	0.192	0.169	0.201
face	**0.772**	0.502	0.362	0.302	0.296
kid	**0.555**	0.462	0.369	0.404	0.446
lightning1	**1.548**	1.014	0.628	0.629	0.823
Mona Lisa	**1.172**	1.055	0.422	0.907	0.634
mountains	0.791	**0.855**	0.557	0.513	0.527
tree	0.312	**0.370**	0.313	0.305	0.221
Average	**0.700**	0.628	0.408	0.454	0.495

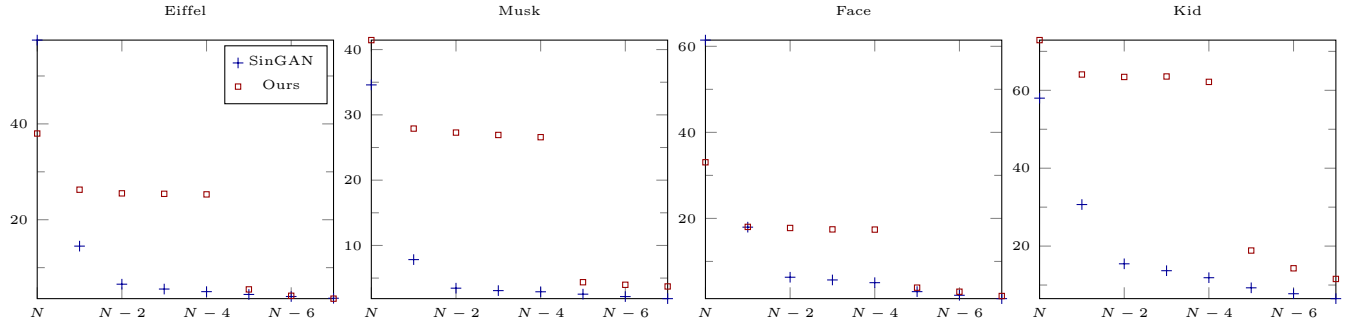


Figure 5: Diversity (mean std. of pixel values of 50 generated images, higher the better diversity) with scale: ours show more diversity even at lower scales compared to SinGAN [3].

first three scales. Experiments without Gaussian smoothing memorize portion of trained image and have that portion unchanged in all over the generation (e.g., see the top left corner of column 3 where large σ has been successful in generating an image with a large variation in hair in comparison with column 1 and 2). This effect is reduced when adding Gaussian smoothing with larger values of σ. This effect is clearly shown in Fig. 10. Here, the diversity is computed at each spatial location using generated images. Self attention, important to add global information to the generation, has a disadvantage of reducing the diversity. Gaussian smoothing at the discriminator resolves this issue and increases the diversity. In view of this, providing the Gaussian is crucial for maintaining diversity in generation. The third column (SinGAN + G) of Table 2 shows the SIFID scores for the experiments only with Gaussian smoothing. Gaussian smoothing helps to increase the image quality in most of the tested images compared to the SinGAN base line.

3.4. Impact of Using Adversarial Feedback

Figure 11 shows the visualization of the feedback in terms of discriminator score for three scales of the balloons image. The area with lower quality generation, e.g., balloons in the image, result in low scores in the discriminator. We feed this back to the generator in the next scale (See Fig. 2). Due to the low score, the generator will get a cue where the improvements are needed and vise versa. The last column of Table 2 (SinGAN + F) shows the SIFID scores for the experiments only with adversarial feedback. Feedback helps to increase the image quality in most of the tested images. Fig.12 shows qualitative results of using feedback.

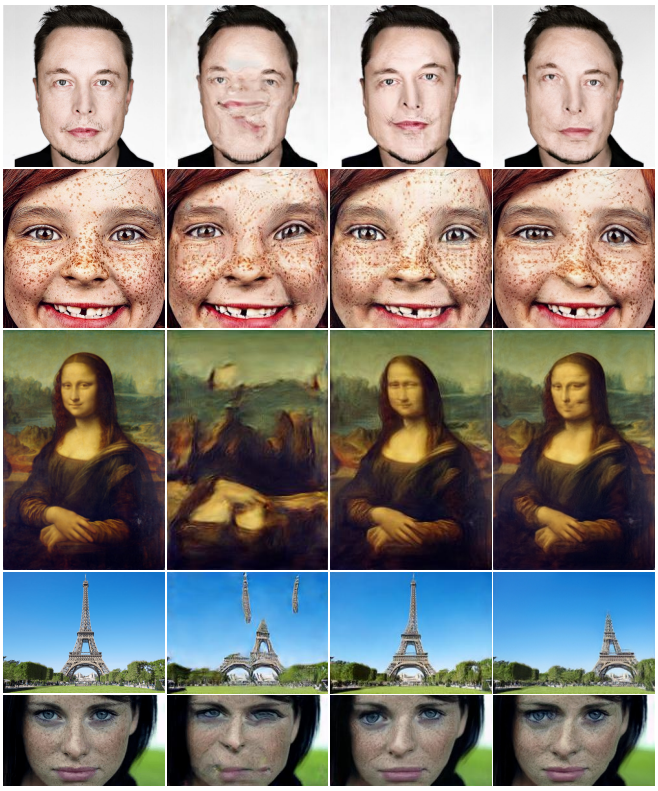
3.5. Overall Impact

In this subsection, we compare the performance of our method with SinGAN [3] and ConSinGAN [4] with respect to the SIFID scores and the diversity scores. In this experiment we use feedback with self-attention and Gaussian smoothing. With the help of feedback, our method is able to produce realistic diverse results when using self-attention blocks only at coarser scales in first three layers of G and D and Gaussian smoothing with σ in the range between 1 and 3.

ConSinGAN has been directly developed from SinGAN architecture with the major contribution of training multiple stages concurrently while propagating features within the stages and using a lower learning rate at coarser scale resulted from specific learning rate scale. We first present the average SIFID scores achieved with our method, SinGAN [3] and ConSinGAN [4] in Table 3. Here, we consider ConSinGAN with the default parameters (2000 epochs with updating D and G three times per epoch, not using normalization layers, using learning rate scale of 0.1), ConSinGAN 6000 which matches hyperparameters of our model (6000 epochs with updating D and G one time per epoch, use instance normalization, using learning rate scale of 0.1) and ConSinGAN lr_scale_0.5 (Here learning rate of lower scales are reduce by the factor of 0.5. Hence coarser scales are trained with higher learning rates than above configuration. This helps to maintain more global structure). It is evident that our method outperforms both SinGAN and ConSinGAN for all the considered images except one.

Table 4 shows the diversity-score comparison with SinGAN and ConSinGAN. Our system gets higher diversity when we consider the ConSinGAN 6000 and ConSinGAN lr_scale_0.5 which generates realistic samples for the images which need global structure to be maintained. Diversity of our system is lower than ConSinGAN with default parameters, but ConSinGAN generated images are not realistic. When compared to SinGAN, our method looses diversity for constraining in the global structure through self attention blocks. As we mentioned in Fig. 5, our method generates highly diverse images compared to SinGAN when it starts the generation not from the coarser scales to retain global structure. Fig. 6 shows the visual results comparison with ConSinGAN. ConSinGAN with default parameters is clearly unable to preserve the global structure. Our method constrains the global structure for the images with larger objects like faces, humans, and buildings while generating diverse images with smaller objects like balloons and cows.

We also test our method in the same 50 images of



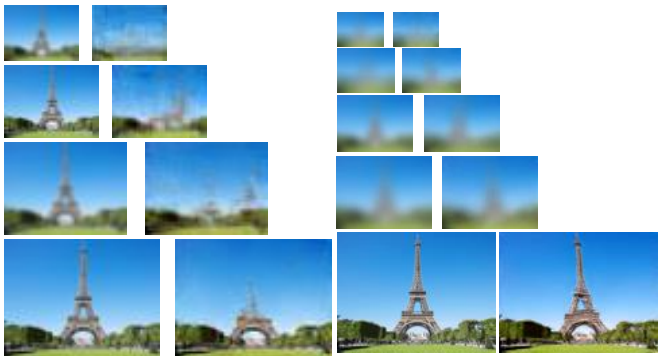
(a) Input (b) ConSinGAN-Def (c) ConSinGAN6000 (d) Ours

(e) Input (f) ConSinGAN-Def (g) ConSinGAN6000 (h) Ours

Figure 6: ConSinGAN vs. ours: (b) and (f) show that ConSinGAN with default parameters (2000 epochs without any normalization layers inside the architecture) is not able to preserve the global structure in generating. Columns (c) and (g) match the hyper-parameters with our with the same number of epochs and instance normalization.

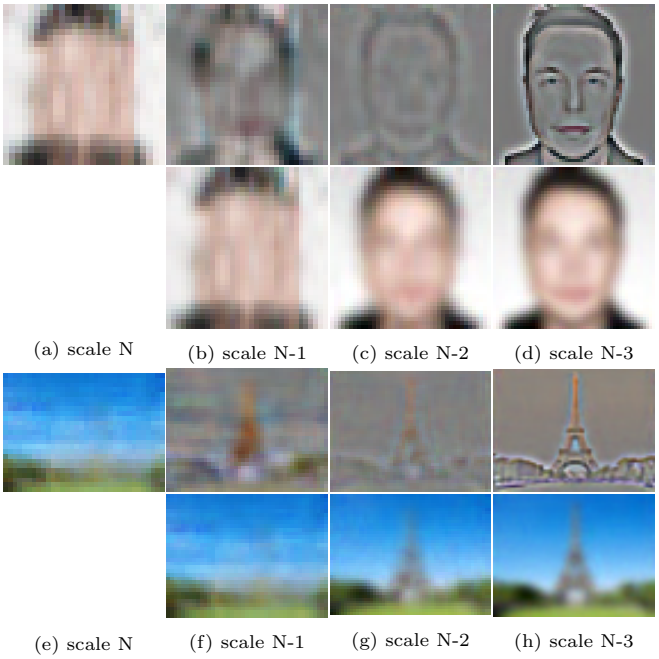
LSUN [48] and Places [49] datasets as in the prior works. In Table 5 and Table 6 we present the average SIFID and the diversity scores for LSUN and Places datasets, respectively. In both datasets we achieve lower SIFID scores without degrading much in the diversity. We show the qualitative comparison in Figure 13.

Furthermore, we conducted a user study to evaluate the quality and diversity of generated images. Using



(a) SA at the first three scales ($k = 2$), $\sigma \in [0.5, 1]$ (b) SA at the first four scales ($k = 3$), $\sigma \in [3, 7]$

Figure 7: Selection of hyper-parameters k and σ. k is the number of scales having self-attention starting from the coarsest scale $k = 0$. Column 1 and 3 are the the Gaussian smoothed real images (smoothed until $N - k$ scale and second and fourth are the corresponding generated results. σ is the standard deviation of the smoothing filter applied to the real images given as an input to the discriminator. $k = 3$ retains the global structure at $N - 4$th scale. Increasing k requires increasing the σ of the Gaussian smoothing too to have diversity in image generation.



(a) scale N (b) scale N-1 (c) scale N-2 (d) scale N-3

(e) scale N (f) scale N-1 (g) scale N-2 (h) scale N-3

Figure 8: Intermediate outputs: residual term (Fig. 1) from generator at each scale and up-sampled results from previous scale for reconstructed samples. Scale N (coarsest) output is an image-like output. However, higher scales generate resisidual (three top right images). Bottom rows show the summation between the residual and up-sampled images form the immediate lower scales. From scale $N-3$ onward, the network generates the high frequency terms which are subdued in previous scale due to the Gaussian smoothing. As a result the quality of the image generation is not affected by the Gaussian smoothing essential to maintain the diversity.

Amazon Mechanical Turk (AMT), we showed the original image and 5 generated images from different methods (SinGAN[3], ConSinGAN[4], ExSinGAN[40], HP-VAE-GAN[42] and our method) and requested the users to select the one with the highest quality and diversity. We

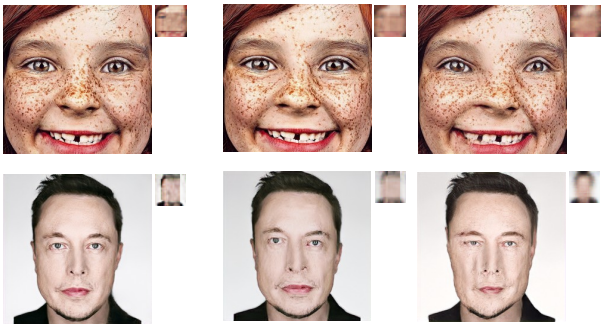
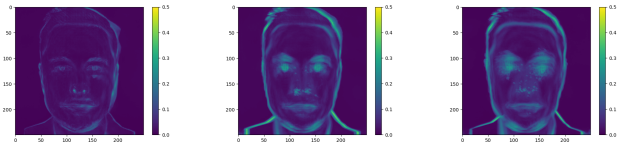
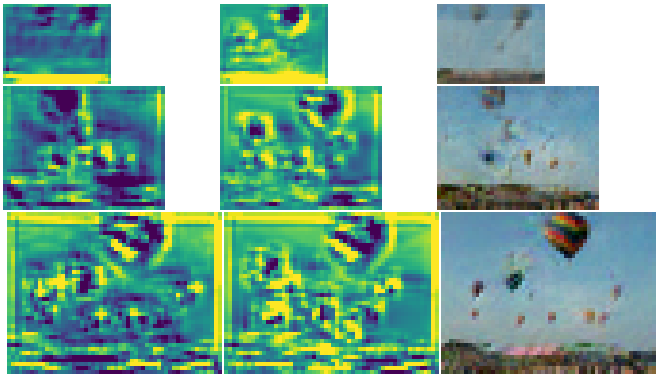


Figure 9: Impact of the σ of the Gaussian kernel in generating diversity while maintaining global structure, particularly at the corners. Column 1 is with no smoothing, column 2 is with $\sigma \in [1, 3]$ and column 3 is with $\sigma \in [3, 7]$. Form the diversity of hair at top-left corner in column 3, in comparison with column 1 and 2, we can see that Gaussian kernel with larger sigma encourages diversity particularly near image corners.



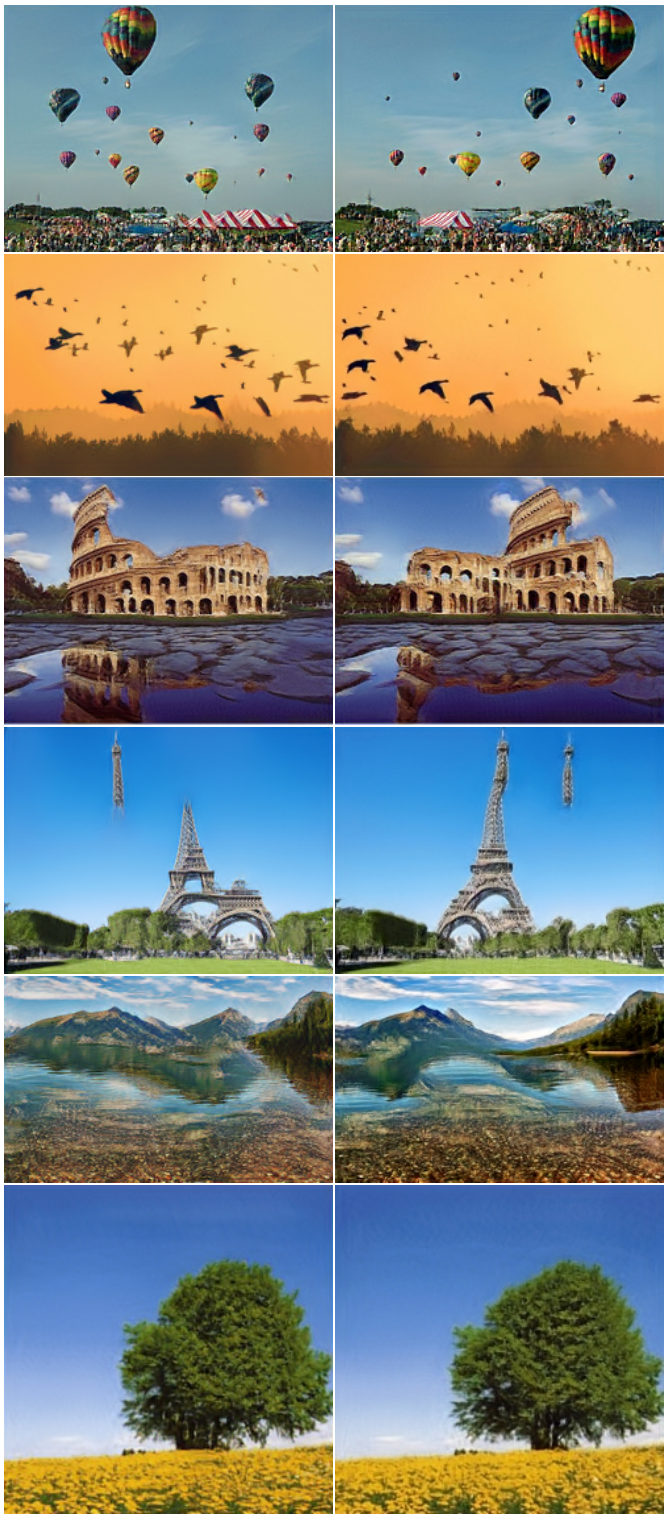
(a) SA only (b) SA3 σ[1,3] (c) SA3 σ[3,7]

Figure 10: The diversity (as shown by the average std. of pixels among 50 generated images). (a) Only with self-attention. (b) Self-attention and Gaussian smoothing. (c) Self-attention and Gaussian smoothing with a larger σ. Notice that as the σ of the Gaussian kernel increases, the diversity too increases.



(a) Discriminator score for fake image (b) Discriminator score for real image (c) Fake sample

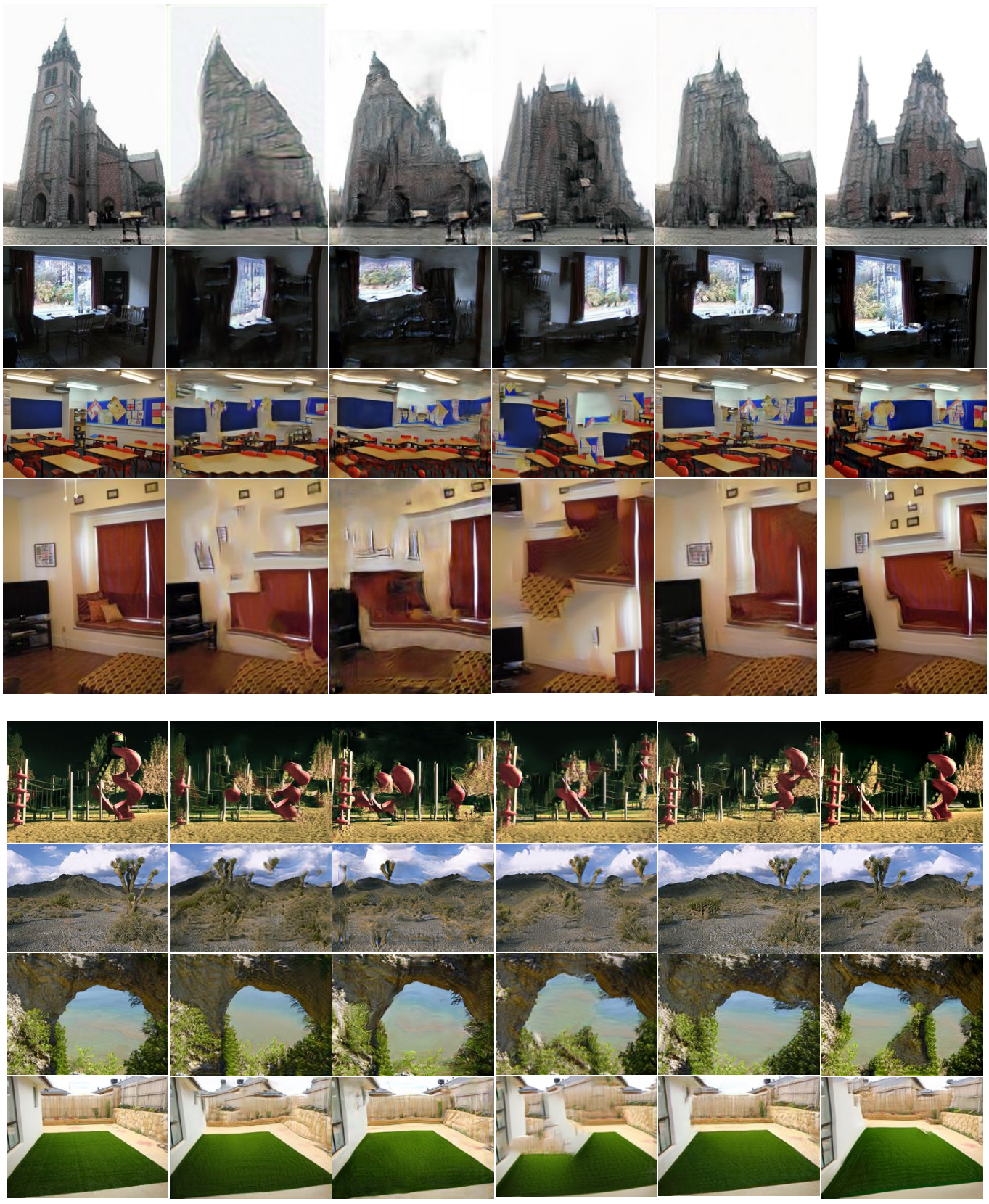
Figure 11: Visualization of adversarial feedback of real and fake image at each scale in terms of the discriminator score (blue: low, yellow: high). Top tow: scale 0, mid row: scale 1, bottom row: scale 2. Notice that the regions in the fake image with low quality generation (e.g., balloons) getting lower values in the feedback. Due to the low score, the generator will get a cue where the improvements are needed.



(a) SinGAN (b) SinGAN with feedback

Figure 12: Effect of incorporating adversarial feedback. The right column with adversarial feedback has better contrast, sharpness (observe the edges) and visual quality. See Table 2 for quantitative results.

showed 49 images from LSUN [48] and 50 images of Places [49] for this user study, and we used one image from LSUN as an example for the users to explain the task clearly. We used AMT to aggregate the responses from 49 participants. For each image, we assigned a score of 1 for the method having the highest user votes, and we present the overall results in Tables 5 and 6, respectively, for LSUN and Places datasets. For both datasets,



(a) Input (b) SinGAN [3] (c) ConSinGAN [4] (d) HP-VAE-GAN [42] (e) ExSinGAN [40] (f) Ours

Figure 13: Comparision on LSUN [48] and Places [49] datasets. Our method is visually more diverse and achieves higher quality than the other methods.

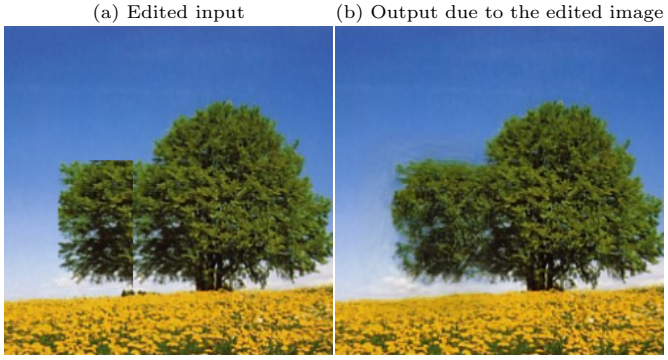
Table 5: Average SIFID score (lower the better), diversity (higher the better), and user votes (higher the better) of generated images from LSUN dataset

Methods	SIFID ↓	Diversity ↑	User votes ↑
SinGAN [3]	0.11	0.60	0
ConSinGAN [4]	0.08	0.55	5
HP-VAE-GAN [42]	0.40	0.78	0
ExSinGAN [40]	0.11	0.50	16
Ours	0.06	0.60	28

Table 6: Average SIFID score (lower the better), diversity (higher the better), and user votes (higher the better) of generated images from Places dataset

Methods	SIFID ↓	Diversity ↑	User votes ↑
SinGAN [3]	0.09	0.52	2
ConSinGAN [4]	0.06	0.50	7
HP-VAE-GAN [42]	0.17	0.62	1
ExSinGAN [40]	0.10	0.47	16
Ours	0.04	0.44	24

(a) Edited input (b) Output due to the edited image



(c) Edited input (d) Output due to the edited image

Figure 14: The performance of our system in image editing: (a) Edited image: Note the rectangular edit (b) successful removal of the edit

our method achieves the highest scores confirming that our method generates images with higher quality and diversity compared to previously proposed methods.

3.6. Image Harmonization, Editing, and Generating Arbitrary-Sized Images

Here we show the results of some other tasks with using self-attention blocks, Gaussian smoothing and feedback. We can do image harmonization by feeding the source image and the image to be blended in at a intermediate scale (e.g., scale 4 or 5). Then the generated image will have

(a) Harmonization input (b) Harmonized output

Figure 15: The performance of our system in harmonization: New mask from different patch (e.g., spacecraft) statistics is harmonized into the trained image.

the image to be blended harmonized into the source image. Fig. 15 shows the ability of our system to harmonize images. Notice how, e.g., the space craft has been harmonized into the background image.

In image editing, an artificially inserted patch at a course scale will be blended in without artifacts. Binary mask indicating the location of inserted patch helps to refine the results by only changing the portion of the inserted patch Fig. 14 shows the ability our our system to edit images. Notice how the edits (light blue patch, and the green branches) have been successfully blended-in.

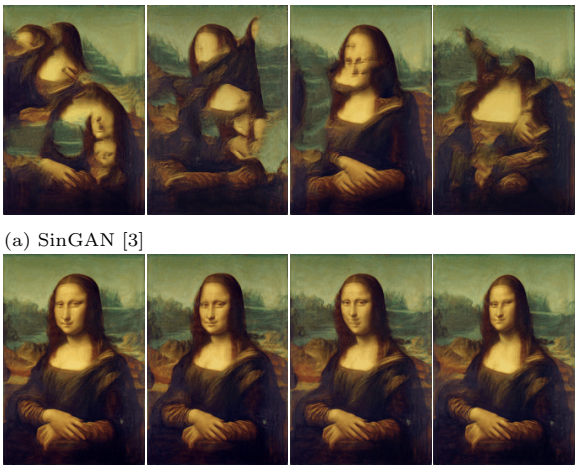
Our method able to produce arbitrary-sized images, this is because of using fixed scale size (m) after downsampling operation in self attention blocks. If we do not use a fixed scale for downsampling, additional features from the arbitray size noise will interfere self attention blocks and reduce the quality of the output. Using a constant scale down size helps to find the inter dependencies of downsampled feature space to compute the re-weighted feature samples. Fig. 16 show four example of arbitrary-sized image generations. Our method cannot produce realistic results when generating arbitrary size images from faces because network cannot preserve the global structure for the arbitrary size input which is not available at the training time. We observed that our animation results also are better in quality than SinGAN.

Figure 16: Arbitrary sized image generation: The new images are generated from noise with a different aspect ratio from the trained image. In this particular example, the width is doubled. Notice how larger images are generated while preserving the global structure.

Above results show that our method is able to produce results on par with SinGAN. Our attention module does not interfere with the the ability to the system in image harmonization, editing, and generating arbitrary-sized images. We are able to do these tasks while preserving the global structure.

3.7. Animation and Image Augmentation

Here we consider two applications of the proposed model: animation and image augmentation. We present four frames of a video generated by our method and SinGAN [3] in Fig. 17. We clearly observe that our method improves the fidelity of an animated video generated with an image which needs global structure to look realistic. This confirms that our method can generate realistic animations with higher diversity while maintaining the global structure compared to SinGAN.



(a) SinGAN [3]

(b) Ours

Figure 17: Four frames of animated videos generated using our method and SinGAN. Our method can generate realistic animations with higher diversity while maintaining the global structure.

We then conduct an experiment to evaluate the feasibility of using our method for data augmentation. To this end, we select two classes (abbey and arch) from SUN database [50]. In each class, we use 4 images and train them to generate 500 images from SinGAN [3] and our method. Next, we train separate classifiers (ResNet18[51]) using these generated images and evaluate the performance with the test set. The classifier trained using the images generated from our method achieves an accuracy of 63.68% whereas the classifier trained using the images generated from SinGAN achieves an accuracy of only 58.42%. The classifier trained only with 8 original samples achieves 52.11%. These experimental results—although produced with eight images—confirm that our model can be successfully used for image augmentation.

4. Limitations

Our work augments the SinGAN architecture to improve its generation quality with an additional attention layer and feedback through the discriminator. Compared to SinGAN, attention layers introduce additional number of trainable parameters. Unlike in SinGAN, we need to forward pass through the previous discriminators for generating the feedback. Having the Gaussian smoothing augmentation also adds an overhead. These reasons make our methodology to take higher training time than SinGAN and ConSinGAN. Even though with our default parameters we achieve the better generation quality without degrading the diversity, image specific parameters will lead to more better results. Since our method trains with only one sample, we cannot control the specific semantic characteristics in the generation from noise.

5. Conclusion

In this work, we improved the technique of image generation in realistic diverse single image generation when global structure is more important. We were able to control the level of global contextual information insertion using self-attention blocks, impose the diversity through convolving the input with a random Gaussian kernel when training the discriminator, and improve the quality of generation with adversarial feedback. This also helped to increase the diversity in generating samples from less-coarse scales significantly compared to SinGAN. Our future work will address generating images that need the global context with varying aspect ratios, denoising, and image inpainting using this work.

Acknowledgement

This work was supported in part by the National Resource Council of Sri Lanka under the grant 19-080.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[2] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: International Conference on Learning Representations, no. 1511.06434, 2016, pp. 1–16.

[3] T. R. Shaham, T. Dekel, T. Michaeli, SinGAN: Learning a generative model from a single natural image, in: IEEE/CVF International Conference on Computer Vision, 2019, pp. 4569–4579.

[4] T. Hinz, M. Fisher, O. Wang, S. Wermter, Improved techniques for training single-image GANs, in: IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1300–1309.

[5] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, D. Rueckert, GAN augmentation: Augmenting training data using generative adversarial networks (2018).

[6] S. Gu, R. Zhang, H. Luo, M. Li, H. Feng, X. Tang, Improved SinGAN integrated with an attentional mechanism for remote sensing image classification, Remote Sensing.

[7] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[8] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. Metaxas, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks (2017).

[9] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, in: International Conference on Learning Representations, 2018, pp. 1–26.

[10] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, in: International Conference on Learning Representations, 2019.

[11] D. Ulyanov, A. Vedaldi, V. Lempitsky, Deep image prior, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9446—-9454.

[12] Y. Gandelsman, A. Shocher, M. Irani, "double-dip": Unsupervised image decomposition via coupled deep-image-priors.

[13] A. Shocher, N. Cohen, M. Irani, "zero-shot" super-resolution using deep internal learning, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3118–3126.

[14] S. Bell-Kligler, A. Shocher, M. Irani, Blind super-resolution kernel estimation using an internal-gan, in: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, 2019, pp. 284–293.

[15] L. P. Zuckerman, E. Naor, G. Pisha, S. Bagon, M. Irani, Across scales and across dimensions: Temporal super-resolution using deep internal learning, in: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII, 2020, pp. 52–68.

[16] N. Jetchev, U. Bergmann, R. Vollgraf, Texture synthesis with spatial generative adversarial networks (2017).

[17] Y. Zhou, Z. Zhu, X. Bai, D. Lischinski, D. Cohen-Or, H. Huang, Non-stationary texture synthesis by adversarial expansion (2018).

[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015.

[19] U. Bergmann, N. Jetchev, R. Vollgraf, Learning texture manifolds with the periodic spatial gan (2017).

[20] A. Shocher, S. Bagon, P. Isola, M. Irani, InGAN: Capturing and retargeting the "DNA" of a natural image, in: IEEE/CVF International Conference on Computer Vision, 2019, pp. 4492–4501.

[21] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, in: Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.

[22] D. Yang, S. Hong, Y. Jang, T. Zhao, H. Lee, Diversity-sensitive conditional generative adversarial networks, in: 7th International Conference on Learning Representations, ICLR 2019, 2019.

[23] S. Gurumurthy, R. K. Sarvadevabhatla, R. V. Babu, DeLiGAN : Generative adversarial networks for diverse and limited data, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, pp. 4941–4949.

[24] K. Li, B. Hariharan, J. Malik, Iterative instance segmentation, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3659–3667.

[25] J. Carreira, P. Agrawal, K. Fragkiadaki, J. Malik, Human pose estimation with iterative error feedback, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, pp. 4733–4742.

[26] K. B. Girum, G. Créhange, A. Lalande, Learning with context feedback loop for robust medical image segmentation, IEEE Transactions on Medical Imaging (2021) 1542–1554.

[27] M. Haris, G. Shakhnarovich, N. Ukita, Deep back-projection networks for super-resolution, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[28] F. Shama, R. Mechrez, A. Shoshan, L. Zelnik-Manor, Adversarial feedback loop, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3205–3214.

[29] M. Huh, S.-H. Sun, N. Zhang, Feedback adversarial learning: Spatial feedback for improving generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1476–1485.

[30] E. L. Denton, S. Chintala, a. szlam, R. Fergus, Deep generative image models using a Laplacian pyramid of adversarial networks, in: Advances in Neural Information Processing Systems, 2015, pp. 1486–1494.

[31] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. Metaxas, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: IEEE International Conference on Computer Vision, 2017, pp. 5908–5916.

[32] U. Demir, G. Unal, Patch-based image inpainting with generative adversarial networks, arXiv (1803.07422) (2018) 1–28.

[33] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[34] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern

Recognition, 2018, pp. 7132–7141.
[35] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Computer Vision – ECCV 2018, 2018, pp. 3–19.
[36] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T. Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6298–6306.
[37] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: International Conference on Machine Learning, 2019, pp. 7354–7363.
[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.
[39] Z. Zheng, J. Xie, P. Li, Patchwise generative convnet: Training energy-based models from a single natural image for internal learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2021.
[40] Z. Zhang, C. Han, T. Guo, Exsingan: Learning an explainable generative model from a single image (2021).
[41] J. Chen, Q. Xu, Q. Kang, M. Zhou, MOGAN: Morphologic-structure-aware generative learning from a single image (2021).
[42] S. Gur, S. Benaim, L. Wolf, Hierarchical patch VAE-GAN: Generating diverse videos from a single sample, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020, pp. 16761–16772.
[43] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, 2017, pp. 214–223.
[44] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of wasserstein GANs, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 5767—5777.
[45] D. Ulyanov, A. Vedaldi, V. S. Lempitsky, Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis, CVPR.
[46] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 448–456.
[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.
[48] F. Yu, Y. Zhang, S. Song, A. Seff, J. Xiao, LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop, arXiv preprint arXiv:1506.03365.
[49] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (6) (2017) 1452–1464.
[50] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba, Sun database: Large-scale scene recognition from abbey to zoo, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3485–3492.
[51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.