Generative Adversarial Networks with Attention Mechanisms at Every Scale

FARKHOD MAHMUDKHUDAEEV, (Member, IEEE), AND IN KYU PARK, (Senior Member, IEEE)
Department of Information and Communication Engineering, Inha University, Incheon 22212, Korea
Corresponding author: In Kyu Park (e-mail: pik@inha.ac.kr).

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1A2C1006706). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2020-0-01389, Artificial Intelligence Convergence Research Center (Inha University)). This work was supported by Inha University Research Grant.

ABSTRACT Existing works in image synthesis have shown the efficiency of applying attention mechanisms in generating natural-looking images. Despite the great informativeness, current works utilize such mechanisms at a certain scale of generative and discriminative networks. Intuitively, the increased use of attention should lead to a better performance. However, due to memory constraints, even moving a single attention mechanism to a higher scale of the network is infeasible. Motivated by the importance of attention in image generation, we tackle this limitation by proposing a generative adversarial network-based framework that readily incorporates attention mechanisms at every scale of its networks. A straightforward structure of attention mechanism enables direct plugging in a scale-wise manner and trains jointly with adversarial networks. As a result, networks are forced to focus on relevant regions of feature maps learned at every scale, thus improving their own image representation power. In addition, we exploit and show the usage of multiscale attention features as a complementary feature set in discriminator training. We demonstrate qualitatively and quantitatively that the introduction of scale-wise attention mechanisms benefits competitive networks, thus improving the performance compared with those of current works.

INDEX TERMS Image synthesis, generative adversarial networks, attention, multiscale.

I. INTRODUCTION

The generative adversarial networks (GANs) [1] have attained tremendous improvements in generating realistic, sharp images. Various findings in terms of architectural composition [2]-[7] and training stability [8]-[12] highly contributed to this matter. Moreover, the implication of attention mechanisms [3], [13] assisted GAN to focus on spatial dependencies. Specifically, attentions complementary to ordinary convolutions of adversarial networks facilitated capturing long-range dependencies across image regions. As a result, by adopting the above findings with attention mechanisms, existing methods such as SAGAN [13] and BigGAN [3] have shown impressive performance where generated images convey fine details. However, the usage of attention mechanism is limited due to memory constraints [3], [14]. Intensive computations that require correlation between feature maps add to it. Thus, it is applied at a certain scale (layer) of generator and discriminator networks. For instance, BigGAN [3] introduces attention mechanism only at the scale of 64 × 64 for a generating final image of 128 × 128. Whilst, moving the attention mechanism into a different scale (256 × 256) of the discriminator is prevented by constraints.

In such circumstances, benefiting from attention at higher resolutions is quite challenging, whereas considerations on the usage of attention at all scales are impracticable. However, as the generator and discriminator learn image composition and decomposition in a coarse-to-fine manner and vice versa, respectively, drawing attention to the features of every scale should increase the data representation power of networks. Although attention mechanisms facilitate a close look at certain information, one more problem may lay down in the discriminator. Specifically, a network structure is designed to focus on learning data distribution that describes it by the most distinctive characteristics in compact form. Hence, it often leans more on global or local information and does not maintain sufficient data representation power to capture both. In fact, this causes overlooking several useful features that might be valuable for performing more accurate
real/fake image classifications.

To alleviate the aforementioned problems in a single form, we propose a GAN-based framework that considers utilizing scale-wise information in a more optimal manner. Unlike existing GANs that apply attention mechanisms at a single layer, first, we propose a framework incorporating attention mechanisms at every scale of the generator and discriminator. Considering that drawing attentions should not affect computational resources, we construct a straightforward mechanism that can be readily incorporated into our generative model. As attentions are drawn in a scale-wise manner, they hold coarse-to-fine details necessary for capturing global and local variations. In turn, this encourages the generator to synthesize semantically and structurally coherent images with finer details.

An attention feedback that provides the ability to amplify the importance of features in critical regions for refining image quality in the generator and to capture the most distinguishable features in the discriminator.

A structural modification in the U-Net based discriminator to increase its data representational power.

An experimental comparison revealing the better performance of our proposal on well-known datasets qualitatively and quantitatively.

II. RELATED WORKS

A. GENERATIVE ADVERSARIAL NETWORKS

GANs [1] and its conditional variant (cGANs) [15] have shown the first possibilities of generating images within adversarial game. This gave rise to GAN-based methods, especially, in the image synthesis task. Since then, various techniques regarding advanced training strategies and improved objective functions [8], [9], [16]–[18] have been proposed. Additionally, an adversarial game between the generator and discriminator was empowered by the current architectural modifications [2], [3], [7]. Such improvements facilitated achieving impressive results where generated images depict more realism [3]–[7], [9], [19]. A recently proposed framework, U-Net GAN [7], improved current image synthesis performance by modifying the discriminator structure to act like a U-shaped network. The method added an encoder network to the original discriminator structure and made the overall network produce per-sample and per-pixel predictions. Exploiting such outputs allowed providing global/local feedback to the generator that enabled the generation of images with finer details.

B. ATTENTION MECHANISM

In general view, attention can be exploited as a guider that contains useful weights towards the most informative regions of feature maps. There are two categories of attention mechanisms: post-hoc network-based analysis [20]–[23] and learnable attention modules [13], [24]–[28]. The works belonging to the first category have been mostly utilized in tasks like object recognition to understand and explain network behavior. In turn, in the second category, learnable attention...
mechanisms have been used in various tasks because it allows to be trained in an end-to-end manner. The aforementioned advances in GANs have also assisted in constructing reliable attention applicable GAN. One of the representative learnable mechanisms that built on such advances is self-attention mechanisms [13]. This method forces the generator to consider long-range dependencies in the feature maps for producing globally coherent images. The importance and performance gain of using attention mechanisms in image synthesis has been shown in the literature [13]. BigGAN [3] further applied such a mechanism in its architecture and generated realistic images. Unlike BigGAN which used attention in novel image synthesis task, GANimation [29] found application of attention for anatomically-aware facial animation (expressions).

By embedding an attention mechanism to the generator network, this method regressed the attention mask specifying the importance of each RGB pixel for synthesizing the novel expression. Hence, GANimation could focus only on the pixels necessary for facial movements. U-GAT-IT [30] used the attention in challenging selfie-to-anime translations where anime drawings have unique properties. To deal the task, U-GAT-IT learnt the weight for auxiliary classifier and exploited such weight with feature maps of image to calculate a set of domain specific attention maps for further usage in generator. The work of [31] proposed a Masked Spatial-Channel Attention (MSCA) module that stands for cross-attention mechanism to model the semantic correspondence between two scenes used in example guided image synthesis.

III. PROPOSED METHOD
A. OVERVIEW
A GAN comprises two networks: generator $G$ and discriminator $D$, where one attempts to fool another in a min-max game. The aim of $G$ is to map a latent variable $z$ drawn from random distribution $p(z)$ to a image space such that synthesized image looks realistic and depicts characteristics of real data distribution $p(x)$. Simultaneously, $D$ aims to differentiate real image $x \sim p(x)$ from synthesized $G(z)$. Generally, $G$ and $D$ are constructed as upsampling and downsampling deep convolutional neural networks, respectively. We follow such standard way, but include the recent advances in GAN [3], [7], [9]. In the generator and discriminator, we incorporate an attention mechanism for every scale in order to unveil the significant feature information learned at the particular layer. Additionally, on the top of U-Net based discriminator, we consider using a composition of multi-scale attentions as complementary information for discriminator training. We provide an overview of our proposed method in Fig. 2.
B. ATTENTION MODULE

Our ultimate goal is to enhance the quality of generated images by the strength of attention information. To this purpose, we construct a straightforward attention mechanism (A). As shown in Fig. 3(a), we have two $1 \times 1$ convolutional layers learning $q$ and $k$, which we denote as query and key following the conventional terminology, respectively. These convolutions are intended to decide where and how much importance on the value $v_i$ (i.e., feature maps of a $i$-th scale) should be given by network. Note that we consider feature-maps as value information and do not add another convolutional layer for it. We think feature maps already convey valuable information, and so, should be directly weighted by attention. In this way, subsequent layers are instantly obtain knowledge on important regions. Besides, this way also simplifies the structure and allow reducing computations. To draw attention, we apply channel pooling on the Hadamard product of $q$ and $k$. Here, channel pooling is similar to the operation defined in [24] that aims to apply max pooling and average pooling across the channels. We perform this operation to obtain spatial attention. Such attention highlights the informative regions within the feature map that are essential to learning. In contrast to [24], we add max and average pool outputs to highlight informative regions more. We convert the resultant matrix into a probability distribution through an activation function and obtain the attention-weighted feature map ($\bar{v}_i$) of a particular scale by multiplying it with spatial attention.

Incorporating this structure at every scale of the generator and discriminator allows benefiting from scale-wise spatial attention maps. The main advantage of attentions is that they force the generator and discriminator to take a close look at attended regions as they highlight the most informative and crucial part of feature maps. We visualize attention maps in Fig. 4 and Fig. 5 drawn in the discriminator and generator respectively. As can be seen, learned attentions contain scale-specific information. For instance, as the generator goes to upsampling layers, we observe that attention starts unveiling the importance on more global information (at lower layers), and gradually moves to highlight the significance of more local features (i.e., eyes, mouth, skin color, hair style, etc.), and reaches more finer details at the resolution near the final layer.

C. ARCHITECTURE

To construct the base network for our framework, we adopt the generator from BigGAN [3] and U-Net based discriminator [7]. Although we opt for these architectures, our attention mechanism can be readily incorporated into different ones. While it is straightforward to introduce attention mechanisms (A) at every scale of generator, the discriminator part requires several considerations. The U-Net based discriminator is constructed using encoder-decoder. The encoder performs per-sample predictions whereas the decoder makes per-pixel predictions where each pixel depicts real or fake values. To get benefit from per-pixel predictions and avoid affecting its process, we add our attention mechanism at every scale of the encoder and keep the decoder part unchanged. We provide architectural structures and layouts for generator and discriminator networks in Table 1 and 2 and their detailed
description in Section IV-A3.

Nevertheless, we consider that encoder part of the discriminator may overlook some useful information. To address it, we modify the encoder of the discriminator according to the structure shown in Fig. 3(b). Specifically, this structure shows adding global average (GAP) and global max-pooling (GMP) layers at each scale of the encoder. We consider these layers to get compact representations of attention-weighted features at every scale. We utilize concatenation of such multi-scale features to have a complementary representation encompassing global and local characteristics of the image. The U-Net based discriminator already has a two-headed function. We add additional head at the bottleneck of this discriminator, where it makes per-sample predictions using an additional linear layer. For training, we use the optimization and regularization functions defined in [7] and adapt computing them on all heads. For more details on the objectives, we refer to the work by [7]. Here, we define main objective functions needed to optimize both generator and discriminator networks. As discriminator from U-Net GAN contains encoder and decoder parts, we update them by

$$\mathcal{L}_{D_{enc}} = - \mathbb{E}_x \left[ \log D_{enc}(x) \right] - \mathbb{E}_z \left[ \log(1 - D_{enc}(G(z))) \right],$$

$$\mathcal{L}_{D_{dec}} = - \mathbb{E}_x \left[ \log[D_{dec}(x)] \right] - \mathbb{E}_z \left[ \sum_{i,j} \log(1 - [D_{dec}(G(z))]) \right],$$

where $D_{enc}$ provides a per-sample output of discriminator through encoder, and $[D_{dec}()]_{i,j}$ represents the decision made by the decoder part of the discriminator at pixel position $(i,j)$. In order to incorporate multi-scale ($ms$) features into optimization process described in the above sections, we utilize a following function which performs a similar operation as (1)

$$\mathcal{L}_{D_{ms}} = - \mathbb{E}_x \left[ \log D_{ms}(x) \right] - \mathbb{E}_z \left[ \log(1 - D_{ms}(G(z))) \right].$$

Hence, a new discriminator loss is computed as a combination of above defined functions:

$$\mathcal{L}_D = \mathcal{L}_{D_{enc}} + \mathcal{L}_{D_{dec}} + \mathcal{L}_{D_{ms}}.$$  \hspace{1cm} (4)

Note that our proposed additions providing multi-scale attention-weighted features in the discriminator part are intended to serve as complementary information for helping to capture and understand meaningful properties of data distributions in GAN training. In order not to disrupt the training process of U-Net based discriminator with its functions for encoder and decoder parts, we follow the strategy of [7] on weighting different terms of Eq. (4). We do not put any special emphasis on $\mathcal{L}_{D_{ms}}$, and use it in the same way as other loss functions by treating them equally.

Correspondingly, the generator optimization can be performed by

$$\mathcal{L}_G = - \mathbb{E}_z \left[ \log D_{enc}(G(z)) + \log D_{ms}(G(z)) \right] + \sum_{i,j} \log[D_{dec}(G(z))]_{i,j},$$

$$\mathbb{E}_z \left[ \sum_{i,j} \log[D_{dec}(G(z))]_{i,j} \right].$$

FIGURE 5. Generator’s scale-wise attentions for synthesizing the leftmost images. Attentions are shown with overlaying on given leftmost images. From the second column to the last, attentions calculated at the resolutions from $4 \times 4$ to $256 \times 256$. Note that brighter regions depict the most attentive features.
IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

1) Datasets

In our evaluations, we consider two well-known face datasets, CelebA [32] and FFHQ [4]. CelebA contains 200k face images of 10k different celebrities (identities). The face images depict different expressions in various facial poses. FFHQ is a relatively new dataset providing 70k high-quality face images. Compared with CelebA, this dataset is more challenging as it contains images with high diversity in terms of age, ethnicity, and viewpoints. For additional demonstration purposes, we consider CelebA-HQ [2], and newly released AFHQ [33] datasets. CelebA-HQ is a high-quality version of the CelebA dataset, but consists of only 30k images. AFHQ also contains 30k animal images classified into cat, dog, and wild categories. The dataset comes at a 256 × 256 resolution that makes it possible for use in synthesizing more detailed animal images. As the CelebA has been released with lower resolution images, we perform experiments by setting the resolution to 128 × 128, whereas we maintain 256 × 256 for others.

2) Training details

To train our framework, we follow parameter settings of [3], [7]. We employ Adam [44] optimizer with momentum parameters $\beta_1 = 0$, $\beta_2 = 0.999$. The learning rates of generator and discriminator are set to $1 \times 10^{-4}$ and $5 \times 10^{-4}$ respectively. We initialize the weights of both networks using orthogonal initialization [35]. For training stability, we use spectral normalization [9]. We sample input to the generator from uniform distribution $z \in [-1, 1]$. Similar to [7], we set the batch size of 20 for FFHQ, CelebA-HQ, and AFHQ, 50 for CelebA.

In evaluation, to generate new images, we utilize an exponential moving average over parameters of the generator network with decaying parameter of 0.9999. We perform the whole task on Tesla V100 GPU with our implementation in PyTorch [36].

3) Implementation

To implement our framework, we use publicly available implementations of generator and discriminator networks from BigGAN [3] and U-Net GAN [7] respectively. Note that adversarial training of these networks might require about a week or more based on the training configurations.

https://github.com/ajbrock/BigGAN-PyTorch
https://github.com/boschresearch/unetgan

TABLE 1. Generator architecture with scale-wise attention mechanisms for generating images at resolution 256 × 256. $ch$ stands for channel multiplier and is set to 64.

| Layer                        | Width × Height | $ch$ |
|------------------------------|----------------|------|
| Encoder                      |                |      |
| ResBlock down                | 16ch → 8ch     |      |
| Attention                    | 8ch → 8ch      |      |
| ResBlock up                  | 8ch → 8ch      |      |
| Attention                    | 8ch → 8ch      |      |
| ResBlock up                  | 8ch → 8ch      |      |
| Attention                    | 8ch → 8ch      |      |
| BN, ReLU, 3 × 3 Conv         | $ch$ → 3       |      |
| Tanh activation              |                |      |

TABLE 2. Discriminator architecture with attention mechanisms and multiscale features (ms) used for generating images at resolution 256 × 256. $ch$ stands for channel multiplier and is set to 64.

| Layer                        | Width × Height | $ch$ |
|------------------------------|----------------|------|
| Encoder                      |                |      |
| ResBlock down                | $ch$ → 2ch     |      |
| Attention                    | 2ch → 2ch      |      |
| ms block                     | 2ch → 2ch      |      |
| ResBlock down                | 2ch → 4ch      |      |
| Attention                    | 4ch → 4ch      |      |
| ms block                     | 4ch → 4ch      |      |
| ResBlock down                | 4ch → 8ch      |      |
| Attention                    | 8ch → 8ch      |      |
| ms block                     | 8ch → 8ch      |      |
| ResBlock down                | 8ch → 8ch      |      |
| Attention                    | 8ch → 8ch      |      |
| ms block                     | 8ch → 8ch      |      |
| ResBlock down                | 8ch → 16ch     |      |
| Attention                    | 16ch → 16ch    |      |
| ms block                     | 16ch → 16ch    |      |
| ReLU, Global sum pooling, Linear $e$ → 1 |                  |      |
| Concat, Linear $ms$ $e$ → 1  |                |      |

Decoder

| Layer                        | Width × Height | $ch$ |
|------------------------------|----------------|------|
| ResBlock up                  | 16ch → 8ch     |      |
| ResBlock up                  | (8 + 8)ch → 8ch|      |
| ResBlock up                  | (8 + 8)ch → 4ch|      |
| ResBlock up                  | (4 + 4)ch → 2ch|      |
| ResBlock up                  | (2 + 2)ch → ch |      |
| ResBlock up                  | $ch + ch$ → $ch$|      |
| ResBlock                     | $ch$ → 1       |      |
| Sigmoid                      |                |      |
We adopt BigGAN [3] generator as a base for constructing our generative model (Table 1). To enhance generated images produced by this generator, we incorporate our proposed attention mechanism at every scale of the network. All blocks are in the form of residual units (ResBlock up) [37] as shown in Figure 6(a), where our mechanism do not interrupts its feature learning process. Similar to existing GANs, we use batch normalization [38], and ReLU activation within this blocks. The upsampling is achieved by applying the nearest neighbor interpolation.

In discriminator part, we make use of U-Net based discriminator (Table 2) as competitor to our generator as discussed earlier. Note, this architecture follows discriminator and generator setups from BigGAN [3]. The network is built on residual units (ResBlock up and down) as shown in Figure 6. We introduced attention mechanism at every scale of encoder sub-network of the discriminator, and kept unchanged the decoder part for per-pixel predictions. To provide complementary information for discriminator, we introduce multi-scale features that are being composed using attention mechanism. In order to make a compact representation for additional adversarial training of discriminator, we add ms block shown in Figure 3(b) that consists of global average and global max pooling layers (see Section III-C). For these features, we add additional linear layer (Linear\(_{ms}\)) as a complementary to encoder’s original layer (Linear\(_e\)). The decoder part of the discriminator produces a single channel output with resolution same as input.

4) Evaluation metrics
To evaluate the performance of our work quantitatively, we use widely applied metrics: Fréchet Inception distance (FID) [39] and Inception score (IS) [40]. These metrics have been reported [39], [40] to have a correlation with human evaluation on assessing the image properties (fidelity and diversity). Designed for this purpose, both measure the quality of generated images to provide a score defining how realistic GAN can perform. Such measurement involves using a pretrained Inception network [41]. FID measures the difference between two distributions (i.e., real and fake) by using specific layer feature-vectors of the Inception network, whereas IS considers the KL-divergence from conditional and marginal distributions. Following the works of [2], [4], [7], we compute FID and IS on 50k generated images. We report the best FID and IS scores achieved through over training processes.

B. EXPERIMENTAL RESULTS
1) Ablation Study
We provide an ablation study that depicts the importance of adding our changes into GAN. In Table 3, we present the

| Method                        | FID ↓ |
|-------------------------------|-------|
| Baseline GAN                  | 7.63  |
| + Scalewise Attentions        | 7.10  |
| + Multi-scale Features        | 6.95  |

FIGURE 7. FID scores of our framework at the particular training iterations on the FFHQ dataset.

FIGURE 8. Influence of attention on feature maps of 256 × 256 scale.

TABLE 3. Ablation study on FFHQ dataset. Performance of adding our proposals in GAN.
As shown, attention provides useful information that enforces generated images with and without applying learned attentions. Maps of ease of visualization and analysis, we test this on the feature-influence of attention information on the feature maps. For We start presenting qualitative results by demonstrating the results are also supported by the improvements in the quality and achieves lower (better) FID score. The quantitative results are also supported by the improvements in the quality of generations, we classify images accurately, and so penalize the generator to in combination with attentions induces the discriminator to classify images accurately, and so penalize the generator to refine its image composition. Overall, the results demonstrate the efficiency of our proposals and their role in improving GAN performance.

In addition, we present a study demonstrating the behavior of our framework in the training stage. To quantify the performance at the particular training stamps, we make use of the FID metric. In terms of quality of generations, we visualize one exemplar image generated at these stamps for better exploration of quality. For this purpose, we use the same noise as input to our framework. Fig.7 presents the behavior of framework from early to final training steps on the FFHQ dataset. By amplifying feature-maps on most crucial regions, attention facilitates generation of images in better quality. As can be seen the generated images are diverse in terms of background variations, pose, expression, etc.

As can be seen the generated images are diverse in terms of background variations, pose, expression, etc. As can be seen the generated images are diverse in terms of background variations, pose, expression, etc. As can be seen the generated images are diverse in terms of background variations, pose, expression, etc. As can be seen the generated images are diverse in terms of background variations, pose, expression, etc. As can be seen the generated images are diverse in terms of background variations, pose, expression, etc. As can be seen the generated images are diverse in terms of background variations, pose, expression, etc.

2) Qualitative Results
We start presenting qualitative results by demonstrating the influence of attention information on the feature maps. For ease of visualization and analysis, we test this on the feature-maps of 256 × 256 scale. In Fig.8 we demonstrate the generated images with and without applying learned attentions. As shown, attention provides useful information that enforces the generator to focus more on those locations of feature maps. Comparing generated images in this manner shows that many fine details, e.g., in eyes, teeth, hairstyle, etc., are introduced using attention information. This example shows that the generator attempts to obtain maximum benefit from attention to improve the overall image quality.

Additionally, we provide Figure 9 which presents more images generated with and without using attention information. By amplifying feature-maps on most crucial regions, attention facilitates generation of images in better quality. As can be seen the generated images are diverse in terms of background variations, pose, expression, etc.

We present the main qualitative results on the FFHQ and CelebA datasets in Fig.10 and 11. The generated images depict data distribution characteristics (e.g., facial pose, age, ethnicity, expressions, etc.) of FFHQ and CelebA. Thus, generated images are diverse and high quality. It is noticeable that better quality is facilitated with finer details.

Along with our results, we provide results from U-Net based discriminator [7] on the FFHQ dataset in Fig.10(a) as this method is the closest to ours. As can be observed, results from both U-Net based approach and ours depict highly realistic faces. In addition to 256 × 256 images, we present results of our method on generating samples at 512 × 512 resolution in Fig.10(c). A number of exemplar faces demonstrate the capability of our method on generating images at higher dimensions.

Additionally, we provide generated images for the CelebA-HQ and AFHQ datasets in Fig.11(b) and (c). For AFHQ, we conduct an experiment with a cat group as we are interested in how our framework can learn to mimic and generate images with such “fluffiness” features. As shown, generated images are natural-looking and diverse, especially, the cat images. Observing such favorable results, we verify that our proposals facilitate GAN to perform better in generating novel images.

3) Quantitative Comparison
We start our quantitative evaluations with the CelebA dataset in Table4 where we compare our results with PG-GAN [2], COCO-GAN [5], BigGAN [3], and U-Net GAN [7]. Our

![FIGURE 9. A diverse set of generated images with (w/) and without (w/o) usage of attention information.](image-url)
FIGURE 10. Images generated by U-Net based discriminator approach [7] and our proposed framework on the FFHQ dataset.

(a) 256 × 256 images generated by U-Net based discriminator [7].

(b) 256 × 256 images generated by our proposed method.

(c) 512 × 512 images generated by our proposed method.
proposed method competes favorably with the recent state-of-the-art methods. Notably, we have better performance compared to U-Net GAN. In Table 5, we provide a comparison of our method against BigGAN and U-Net GAN on FFHQ. Similar to CelebA, our method improves the current

### TABLE 4. Quantitative comparison with recent GANs on the CelebA (128 × 128) dataset.

| Method      | FID ↓ | IS ↑ |
|-------------|-------|------|
| PG-GAN [2]  | 7.30  | -    |
| COCO-GAN [5]| 5.74  | -    |
| BigGAN [3]  | 4.54  | 3.23 |
| U-Net GAN [7]| 2.95 | 3.43 |
| Ours        | **2.38** | **3.55** |

### TABLE 5. Quantitative comparison with recent GANs on the FFHQ (256 × 256) dataset.

| Method      | FID ↓ | IS ↑ |
|-------------|-------|------|
| BigGAN [3]  | 12.42 | 4.02 |
| U-Net GAN [7]| 7.63 | 4.47 |
| Ours        | **6.95** | **4.73** |

### FIGURE 11. Images generated by our proposed framework on (a) CelebA (128 × 128), (b) CelebA-HQ (256 × 256) and (c) AFHQ datasets.
performance in this dataset. It is worth mentioning that our framework achieves much better results than BigGAN. By performing such experiments, here, we also demonstrate the better performance of our proposed attention compared with self-attentions [13] applied in original U-Net GAN at the specified layers of generator and discriminator. Overall, the quantitative and qualitative results demonstrate that our proposals force the generator and discriminator to pay attention to the feature maps at every scale, and exploit them in a better manner.

V. CONCLUSION

In this work, we addressed the issue of using attention mechanism in GANs, specifically constraints that limit its application. To alleviate this issue, we reconsider the structure of attention mechanism. In our proposed framework, we introduced a straightforward attention calculation mechanism at every scale of generator and discriminator. By applying such a scale-wise approach, we demonstrated the possibility of drawing attention at the important regions of feature maps at every scale without having memory constraints. Additionally, by leveraging such attention information, we strengthened the discriminator to learn data characteristics more accurately. Overall, exploiting attention in a scale-wise manner demonstrated better performance and improvements in the quality of generated images.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in NeurIPS, 2014, pp. 2672–2680.
[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in Proc. of International Conference on Learning Representations, 2018.
[3] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” ICLR, 2018.
[4] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in CVPR, 2019, pp. 4401–4410.
[5] C. H. Lin, C.-C. Chang, Y.-S. Chen, D.-C. Juan, W. Wei, and H.-T. Chen, “Coco-gan: Generation by parts via conditional coordinating,” in ICCV, 2019, pp. 4512–4521.
[6] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in CVPR, 2020, pp. 8110–8119.
[7] E. Schonfeld, B. Schiele, and A. Khoreva, “A u-net based discriminator for generative adversarial networks,” in CVPR, 2020, pp. 8207–8216.
[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of Wasserstein GANs,” in Proc. of Advances in Neural Information Processing Systems, 2017, pp. 5767–5777.
[9] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” in ICLR, 2018.
[10] L. Mescheder, S. Nowozin, and A. Geiger, “Which training methods for GANs do actually converge?” in ICMIL, 2018.
[11] H. Zhang, Z. Zhang, A. Odena, and H. Lee, “Consistency regularization for generative adversarial networks,” ICLR, 2020.
[12] D. Zhang and A. Khoreva, “Progressive augmentation of gans,” NeurIPS, 2019.
[13] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in ICMIL, 2019, pp. 7354–7363.
[14] G. Daras, A. Odena, H. Zhang, and A. G. Dimakis, “Your local GAN: Designing two dimensional local attention mechanisms for generative models,” in CVPR, 2020, pp. 14.531–14.539.
[15] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.
[16] J. Zhao, M. Mathieu, and Y. LeCun, “Energy-based generative adversarial network,” ICLR, 2017.
[17] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in ICML, PMLR, 2017, pp. 214–223.
[18] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in Proc. of the IEEE International Conference on Computer Vision, 2017, pp. 2794–2802.
[19] T. Miyato and M. Koyama, “cGANs with projection discriminator,” in ICLR, 2018.
[20] K. Simonan, Y. Vedaladi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” arXiv preprint arXiv:1312.6034, 2013.
[21] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in CVPR, 2016, pp. 2921–2929.
[22] R. R. Selvaraju, M. Cosgwell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in ICCV, 2017, pp. 618–626.
[23] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” in WACV. IEEE, 2018, pp. 839–847.
[24] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in ECCV, 2018, pp. 3–19.
[25] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in CVPR, 2018, pp. 7794–7803.
[26] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in CVPR, 2017, pp. 3136–3146.
[27] X. Chen, C. Xu, X. Yang, and D. Tao, “Attention-gan for object transfiguration in wild images,” in ECCV, 2018, pp. 164–180.
[28] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, “Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation,” in CVPR, 2019, pp. 2417–2426.
[29] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, “Ganimation: Anatomically-aware facial animation from a single image,” in ECCV, 2018, pp. 818–833.
[30] J. Kim, M. Kim, H. Kang, and K. H. Lee, “U-gat-it: Unsupervised generative adversarial networks with adaptive layer-instance normalization for image-to-image translation,” in ICLR, 2020.
[31] H. Zheng, H. Liao, L. Chen, W. Xiong, T. Chen, and J. Luo, “Example-guided image synthesis using masked spatial-channel attention and self-supervision,” in ECCV. Springer, 2020, pp. 422–439.
[32] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proc. of the IEEE international Conference on Computer Vision, 2015, pp. 3730–3738.
[33] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, “StarGAN v2: Diverse image synthesis for multiple domains,” in CVPR, 2020, pp. 8188–8197.
[34] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” ICLR, 2015.
[35] A. M. Saxe, J. L. McClelland, and S. Ganguli, “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks,” ICLR, 2014.
[36] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.
[37] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in ECCV, 2016, pp. 630–645.
[38] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167, 2015.
[39] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in NeurIPS, 2017, pp. 6626–6637.
[40] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” in NeurIPS, 2016, pp. 2234–2242.
[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in CVPR, 2015, pp. 1–9.