Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

An unconditional generative model with self-attention module for single image generation

Tek görüntü üretimi için öz-dikkat modüllü koşulsuz üretken bir model

Eyyüp Yıldız[1,*], Mehmet Erkan Yüksel[2], Selçuk Sevgen[3]

[1] *Erzincan Binali Yıldırım University, Computer Engineering Department, 24002, Erzincan Türkiye*
[2] *Burdur Mehmet Akif Ersoy University, Computer Engineering Department, 15200, Burdur, Türkiye*
[3] *İstanbul University-Cerrahpaşa, Computer Engineering Department, 34320, İstanbul, Türkiye*

Abstract

Generative Adversarial Networks (GANs) have revolutionized the field of deep learning by enabling the production of high-quality synthetic data. However, the effectiveness of GANs largely depends on the size and quality of training data. In many real-world applications, collecting large amounts of high-quality training data is time-consuming, and expensive. Accordingly, in recent years, GAN models that use limited data have begun to be developed. In this study, we propose a GAN model that can learn from a single training image. Our model is based on the principle of multiple GANs operating sequentially at different scales, where each GAN learns the features of the training image and transfers them to the next GAN, ultimately generating examples with different realistic structures at the final scale. In our model, we utilized a self-attention and new scaling method to increase the realism and quality of the generated images. The experimental results show that our model performs image generation successfully. In addition, we demonstrated the robustness of our model by testing it in different image manipulation applications. As a result, our model can successfully produce realistic, high-quality, diverse images from a single training image, providing short training time and good training stability.

Keywords: Generative adversarial networks, single image generation, self-attention, image manipulation

Özet

Üretken Çekişmeli Ağlar (GANs), yüksek kaliteli sentetik verilerin üretilmesini sağlayarak derin öğrenme alanında devrim yaratmıştır. Bununla birlikte, GAN'ların etkinliği büyük ölçüde eğitim verilerinin boyutuna ve kalitesine bağlıdır. Birçok gerçek dünya uygulamasında, büyük miktarda yüksek kaliteli eğitim verisi toplamak zaman alıcı ve pahalı bir süreçtir. Buna bağlı olarak, son yıllarda, az veri kullanan GAN modelleri geliştirilmeye başlanmıştır. Bu çalışmada tek bir eğitim görüntüsünden öğrenebilen üretken çekişmeli ağ model önermekteyiz. Modelimiz, farklı ölçeklerde sıralı olarak çalışan birden fazla GAN'ın, eğitim görüntüsünün özelliklerini öğrenip son ölçekte farklı gerçekçi yapılarla örnekler ürettiği bir prensibe dayanmaktadır. Modelimizde, üretilen görüntülerin gerçekçiliğini ve kalitesini artırmak amacıyla bir öz-dikkat ve yeni ölçeklendirme yöntemi kullandık. Deneysel sonuçlar, modelimizin başarılı bir şekilde çalıştığını göstermektedir. Buna ilaveten, modelimizi farklı görüntü manipülasyonu uygulamalarında test ederek model sağlamlığını ortaya koyduk. Sonuç olarak, geliştirdiğimiz GAN modeli; tek bir eğitim görüntüsünden faklı, gerçekçi ve kaliteli görüntü örneklerini başarılı bir şekilde üretebilmekte, kısa eğitim süresi ve iyi eğitim kararlılığı sağlamaktadır.

Anahtar kelimeler: Üretken çekişmeli ağlar, tek görüntü üretimi, öz-dikkat, görüntü manipülasyonu

1 Introduction

Recent advances in deep learning techniques have contributed significantly to the growth of artificial intelligence. However, the acquisition of consistently structured datasets that conform to specified criteria poses a remarkable challenge confronting researchers and developers in the field of artificial intelligence. Creating large problem-specific datasets and performing the necessary pre-processing operations is challenging and time-consuming; in some cases, the dataset may not exist at all. Therefore, developing successful models for learning from small amounts of data has emerged as a significant research area. Recent works have focused on developing methods to learn effectively from limited data, such as transfer learning [1], meta-learning [2], and data augmentation [3]. Besides, several studies have explored the use of generative models to learn from low data regimes. For example, Variational Autoencoders (VAEs) have been utilized in natural language processing to produce new samples [4]. Few-shot learning generates more samples for under-represented classes using Generative Adversarial Networks (GANs) [5].

GANs [6], stand out as a powerful method in machine learning to generate synthetic data that can be used for various applications. However, training GANs with limited data has some challenges. One of these challenges is overfitting. GANs can lead to memorization of training data rather than learning specific features that generate new data. Hence, they are highly susceptible to overfitting when trained on small datasets. The other challenge is the mode collapsing which is a limited set of outputs, drawn by the

* Sorumlu yazar / Corresponding author, e-posta / e-mail: eyyup.yildiz@erzincan.edu.tr (E. Yıldız)
Geliş / Recieved: 27.09.2023 Kabul / Accepted: 15.10.2023 Yayımlanma / Published: 15.01.2024
doi: 10.28948/ngumuh.1367602

generator, that do not reflect the exact distribution of the training data. Mode collapse occurs more frequently when the generator network does not have enough samples to learn all the features in the data. Another challenge is training instability. GANs can be highly sensitive to hyperparameter selection, such as learning rates and batch sizes, and require a significant amount of fine-tuning to achieve stable training [7-9].

This study aims to propose an enhanced generative model that addresses the aforementioned issues by utilizing a single training image to generate new samples. Our model follows Single Image Generative Adversarial Network (SinGAN) [10], which is trained on a single natural image and produces diverse and visually plausible samples. SinGAN relies on a pyramid structure comprising fully convolutional GANs, wherein each level corresponds to a different scale of the input image. The key breakthrough of SinGAN lies in its utilization of a generative model at each pyramid level, enabling the transformation of feature maps from the previous level into the corresponding feature maps of the current level. However, despite the impressive outcomes, images generated by SinGAN often encounter challenges in preserving the intended overall structure or semantic content of the original image. In our model, we employed the self-attention module [11] to capture pixel dependencies within a single image, resulting in the generation of highly realistic images. Self-attention is a valuable tool as it enables the model to selectively emphasize different parts of the data that are most relevant for rendering the image. Additionally, we introduced a new scaling method for image resizing in our model. This method specifically focuses on enhancing the realistic representation of medium-sized objects in training images. Moreover, this approach facilitates a seamless transition from global dependencies to local dependencies during the image generation phase, thereby improving the overall coherence of the generated results. Consequently, our model aims to improve the quality and fidelity of produced images, addressing the specific challenges related to realism and object consistency.

The contributions of our study are summarized as follows:

- Using the self-attention module: To enhance the realism of generated images and improve the coherence of depicted objects.
- Introducing a new scaling method: To focus on medium-sized images generated in coarse scales of the training.

The rest of the paper is organized as follows. Section 2 presents a related work. Section 3 introduces our generative model using the self-attention module for the single image generation problem. Section 4 provides experimental results. Section 5 presents some image manipulation tasks such as paint-to-image and harmonization. Finally, Section 6 summarizes and concludes the paper.

2 Related Work

Developing GAN models that effectively operate with a limited amount of data poses an intriguing yet challenging task. Zakharov et al. [12] introduced an innovative solution in the form of a few-shot learning approach, which enables the creation of high-quality videos featuring individuals speaking, even when only a small number of images of the target person are available. The proposed model leverages a GAN architecture that operates specifically based on these target person images. Moreover, to improve the generated images' quality and maintain consistency with the target person's appearance, the GAN architecture is combined with a meta-learning approach. The results obtained from both training and test datasets demonstrated that the method is capable of producing impressive talking head videos. These videos exhibit realistic lip synchronization and facial expressions, even when trained on a small number of images. Lucic et al. [13] presented a model aimed at generating high-quality images when confronted with a scarcity of training data. Their approach involved the development of a GAN model incorporating both self-supervised learning and semi-supervised learning techniques. Self-supervised learning is used for extracting semantic features and guiding the training of the learnable GAN. Semi-supervised learning is used for selectively removing labels from a small subset of labelled training images and utilizing this modified dataset as conditional information during GAN training. By using self-supervised learning and semi-supervised learning together, the authors overcame the challenges associated with limited training data and achieved the generation of high-quality images. Noguchi and Harada [14] focused on the challenge of generating high-quality images when working with small datasets. They introduced a method that processes the batch statistics of a pre-trained GAN to adapt the characteristics of the limited dataset. By tailoring the aggregated statistics, their approach enables the model to generate high-quality images that align with the distribution of the small dataset. This adaptation process ensures that the generated images maintain consistency and fidelity with limited data, resulting in improved image quality.

Generating high-quality, diverse images from only a single training image has long been a challenging task in the field of deep learning. In response to this challenge, Shocher et al. [15] proposed the Internal GAN (InGAN) model, which stands as the pioneering generative model designed to operate exclusively on a single training image. Unlike traditional GAN models, InGAN specifically focuses on capturing and manipulating the internal structure of a single natural image, encompassing both low-level and high-level feature representations. By emphasizing the internal structure, InGAN produces diverse samples from a single training image. Shaham et al. [10] introduced SinGAN, which is an unconditional GAN model trained on a single natural image. SinGAN distinguishes itself from InGAN by becoming the first generative model capable of learning and producing new samples from a single training image. The model is structured around a hierarchical arrangement of fully convolutional GANs, with each level corresponding to a different scale of the training image. By training each sample using the input image progressively scaled from small to large, the top-level generative network generates the final output image. SinGAN's main innovation lies in its implementation of a generative model at each hierarchical level, enabling the transformation of feature maps from the previous level to align with the corresponding feature maps at the current level. In addition, the integration of the patch discriminator [16] effectively reduces the memorization, allowing for training on the entire image. This significantly increases SinGAN's performance and image generation capabilities. Hinz et al. [17] presented Concurrent Single Image Generative Adversarial Network (ConSinGAN), which is based on the pyramid structure of SinGAN. However, ConSinGAN distinguishes itself in training method. Instead of networks operating at a single scale, it enables multiple scales to be trained concurrently with different learning rates,

The remarkable success of deep learning-based models in various domains has intensified the interest in this field. Alongside their achievements, researchers have proposed numerous methods to further optimize the performance of these models. Among these methods, attention has emerged as a prominent technique for enabling models to identify both local and global dependencies within datasets. The attention mechanism addresses a limitation in convolutional neural networks, where the filters that underpin their success excel at capturing local dependencies but struggle to detect global dependencies due to their limited receptive fields. By incorporating attention, models can selectively focus on different parts of the data, allowing them to capture and utilize both local and global information effectively. This advancement has proven instrumental in enhancing the capabilities of deep learning models, enabling them to handle a wider range of complex tasks by appropriately addressing both local and global dependencies within the data. The initial version of the attention mechanism was introduced by Bahdanau et al. [18]. The authors applied their model to the machine translation problem [18]. Subsequently, Vaswani et al. [19] proposed an attention mechanism for machine translation, which achieved remarkable success. This breakthrough paved the way for the widespread adoption of attention mechanisms beyond text-based models, extending their application to tasks such as image recognition [20], image classification [21], and image segmentation [22]. The introduction of the self-attention mechanism in the context of image generation was accomplished by H. Zhang et al. with Self-Attention Generative Adversarial Network (SAGAN) [11]. This study demonstrated the successful utilization of self-attention, showcasing its effectiveness in image-generation tasks with fewer iterations. In terms of detecting dependencies within images and producing high-quality results, SAGAN stands out as a pioneering and original approach among GAN models. Its incorporation of self-attention has significantly advanced the field of image generation and demonstrated its potential for generating visually compelling outputs.

3 Material and method

Our model is an unconditional GAN that generates diverse images. It performs image generation by employing a sequential approach with multiple GANs that cater to various image sizes. The overall architecture of our model is showed in Figure 1. We followed a similar image generation procedure with SinGAN [10]. Training across all sizes follows the conventional principles of GAN training, maintaining consistency throughout. In each dimension, the training process remains consistent. Specifically, for the smallest scale, the generating network receives input in the form of noise-only data. However, as we move through sequential scales, the generative networks take as input the sum of the image generated at the preceding scale and the noise data. This iterative process ensures a progressive generation of images with increasing levels of complexity and detail, resulting in high-quality outputs. Let's denote the scaled image set of x with $\mathrm{K} = \{x_0, x_1, \cdots, x_s\}$, although the image used for training is x, the number of scales is s. For each scale, we denote generator and discriminator pairs as $\{(G_0, D_0), (G_1, D_1), \cdots, (G_s, D_s)\}$. According to this, the model operates starting from the initial network (G_0, D_0) and continues training until the completion of network (G_s, D_s). Once the training of (G_i, D_i) is completed, the parameter values are frozen, and the training transitions to (G_{i+1}, D_{i+1}). For ith scale, to generate the image $\tilde{x}_{i+1}$ produced by (G_i), upsampling is applied to the image using interpolation, and noise data is added. This resulting image is then fed into the generator network G_{i+1} of (G_{i+1}, D_{i+1}) for training.

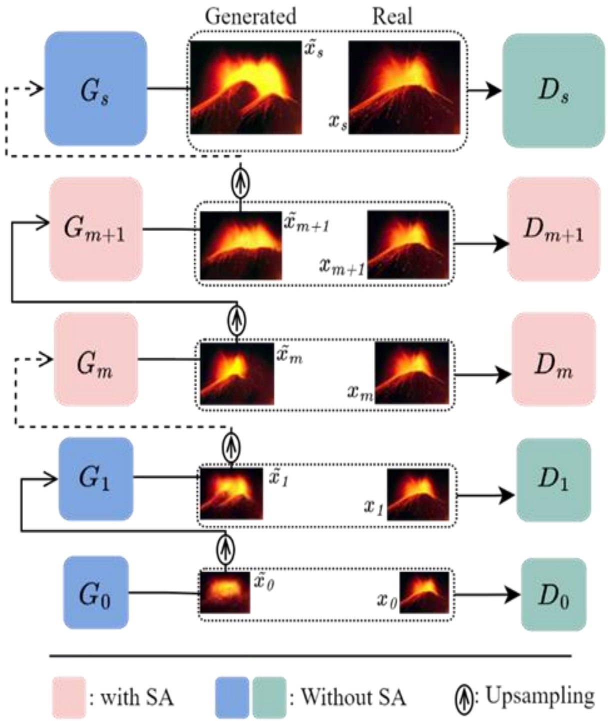

Figure 1. The system architecture of our model.

For model training, we employed a combination of adversarial and reconstruction loss functions, as specified in Equation (1). The patch discriminator was utilized as the means to split the image into patches for evaluation. Unlike evaluating the entire training image, the patch discriminator focuses on assessing individual patches within the image, enabling more targeted analysis. The reconstruction loss function was employed to ensure that specific (fixed) noise values converge to the training image within the infinite noise space. To implement this, we utilized the $\mathcal{L}2$ quadratic difference function described in Equation (2). As for the adversarial loss function, we adopted the Wasserstein Generative Adversarial Network Gradient Penalty (WGAN-GP) [23] formulation, which has proven to be effective in promoting adversarial learning and generating high-quality results. WGAN-GP stands as a variant of the traditional GAN framework and offers significant improvements by introducing the Wasserstein distance metric and the gradient penalty technique. These additions address several limitations commonly encountered in traditional GANs, such as training instability and mode collapsing. The Wasserstein distance metric serves to quantify the dissimilarity between the distributions of real and generated data, offering a more informative and easier-to-optimize measure compared to conventional Jensen-Shannon or Kullback-Leibler deviations. By leveraging the Wasserstein distance, the model gains valuable insights into the quality of the generated samples.

The parameter values of the model were used based on [10]. However, we determined the iteration and filter number parameters as a result of our experiments. In our model, both the generator and discriminator networks are trained for 2000 iterations at each scale. We employ a 3x3 convolution-Batch Normalization- Leaky Rectified Linear Unit (Leaky ReLU) layer in both networks. The Tanh activation function is used solely in the last layer of the generator network for each scale. Our networks consist of 6 convolution blocks, with 32 filters used for images up to half of the training image scale and 64 filters used for larger images at other scales. In networks without self-attention, we substitute self-attention with a convolution layer. In cases where the same number of filters are used, we set the initial parameters of the networks to the final values obtained in the previous scale. However, for other scenarios, we initialize the networks with random parameter values. The learning rate employed is set at 5e-4. We used Adam optimizer [24] with $\beta_1 = 0.5, \beta_2 = 0.9$. We determined the coefficient of reconstruction loss values as α=10 [10]. The average run time of our model on the NVIDIA TITAN X PASCAL GPU is 35 minutes.

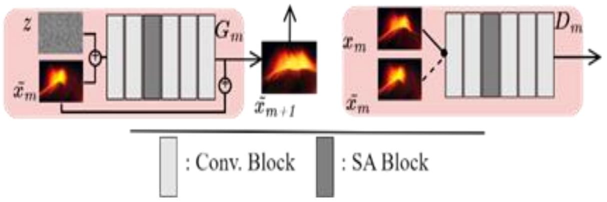

Figure 2. The self-attention module in the generator and discriminator.

$$\min_{G_s} \max_{D_s} \mathcal{L}_{adv}\,(G_s, D_s) + \alpha\mathcal{L}_{rec}(G_s) \qquad (1)$$

$$\mathcal{L}_{rec} = \|G(z_i) - x_i\|_2^2 \qquad (2)$$

The success of our work depends on the number of scales needed for training and the size of the educational image at each scale. In addition to enhancing the model structure, we focused on improving the scaling method. Generally, the receptive fields of convolution filters in the generator and discriminator networks match the size of medium-sized objects in the image. Hence, during the image scaling process, our objective was to generate a greater number of scales that aligned with the sizes of the targeted objects within the processing areas. This approach allows us to achieve improved results and a better representation of objects at various scales. In SinGAN, the scaling factor for an image with size x_S to obtain an image with size x_m in dimension m is given by r^m, where r is the scaling factor. Therefore, the scaling operation in SinGAN can be expressed as $x_m = x_S \times r^m$. obtain scales with a higher representation of medium-sized images, we determined the scale coefficient using the piecewise function outlined in Equation (3). This approach enables our model to effectively transfer global features learned from small-scale images to large-scale local features. In our proposed model, we utilized a minimum image size of 25 pixels and a maximum image size of 250 pixels for the scaling process. Figure 3 illustrates a comparison between the scaling methods employed by SinGAN and our model. Through experimentation, we determined the scaling factor ($r_0 = 0.7$ and the scale add-on $\varepsilon = 0.1$ in the scaling method based on the fundamental base. These parameters were identified as a result of rigorous testing and analysis.

$$r^m = \begin{cases} (1+\varepsilon) \times {r_0}^m, 0 \le m \le 2 \\ (1-\varepsilon) \times {r_0}^m, 3 \le m \le 5 \\ {r_0}^m, m > 5 \end{cases} \qquad (3)$$

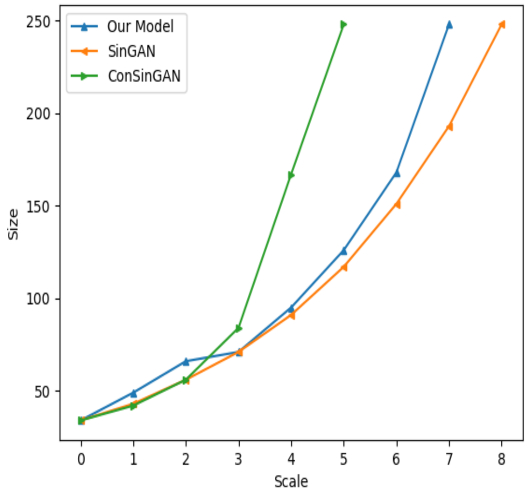

Figure 3. Comparison of the scaling methods.

In our model, we incorporated the self-attention block to create medium-sized networks (G_i, D_i). Figure 2 represents internal structure of each (G_i, D_i) network pair. Every G_i has a residual connection that aggregates input image $(\tilde{x}_i)$ with output image. Each network contains the self-attention block, which was implemented with reference to [11]. This block establishes a connection between each element of the input array, such as individual pixels of the image, and a vector representation within the self-attention mechanism. By utilizing these vector representations, we assess the significance of each element by calculating attention weights. The self-attention mechanism provides a crucial advantage: the ability to selectively focus on different components of the input array. This selective focus is especially valuable when certain regions of the image have varying importance in generating specific outputs. With the aid of self-attention, our model concentrates on the most critical areas of the image for generating each pixel, resulting in higher-quality images with enhanced levels of detail. In the scaling method we use, our model focuses more on medium-sized images. Using a self-attention block at all scales reduces the variety of images produced and requires excessive processing power. For this reason, we used the self-attention block only in medium-sized networks.

4 Results and discussions

To evaluate the performance of our model, we utilized a dataset of 50 images selected from the Places dataset [25], as well as additional training images obtained from the Internet. Figure 4 shows several image samples generated by our model by using the training images in [25-27]. The results demonstrate that our model can produce diverse and realistic images, maintaining high variability while being situated in the same space as the training images. For instance, upon examining the waterfall image in the first row, our model successfully generates realistic patterns with varying structures and locations. Similarly, in the second row featuring the colosseum image, the generated samples showcase the model's ability to maintain the structure and position of the Colosseum while producing distinct and realistic variations. This highlights the model's capability to capture pixel dependencies effectively. Additionally, we evaluated our model's performance by feeding inputs of different sizes to the generator network, consisting of convolutional layers. We randomly selected and presented some of the resulting outputs in Figure 4. These demonstrate how our proposed model successfully generates diverse and realistic images while preserving pixel dependencies, even for inputs of variable sizes.

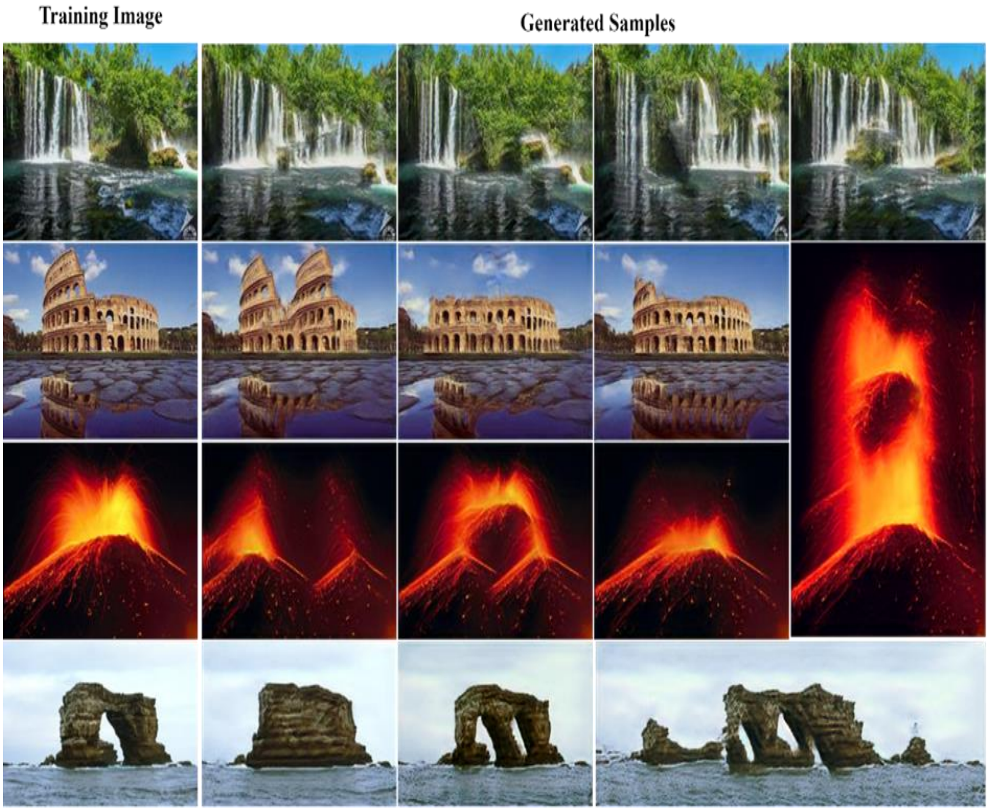

Figure 4. Images samples generated by our model.

Table 1. Quantitative results for Places dataset.

	SinGAN [10]	ConSinGAN [17]	Our Model
SIFID(↓)	0.09	0.06	0.08
SSIM(↓)	0.46	0.44	0.47
# Scale	~8-9	~5-6	~7-8
Training Time	~50 dk.	~25 dk.	~35 dk.
# Parameter	~1.350.000	~650.000	~1.050.000

To quantitatively evaluate the images produced by our model, we employed the Single Image Fréchet Inception Distance (SIFID) method [10]. SIFID is an adaptation of the Fréchet Inception Distance [29] specifically tailored for single images. By utilizing feature maps extracted from intermediate layers of a pre-trained Inception network, SIFID compares the real and generated images. A lower SIFID value suggests that the generated images are more realistic. Additionally, to assess diversity, we employed the Structural Similarity Index Metric (SSIM) [30]. SSIM serves as a metric for measuring the similarity between two images, taking into account structural information, brightness, and contrast. It achieves this by dividing the images into small pixel windows and calculating their respective means, variances, and covariances. The mean represents the average brightness within the window, the variance reflects the contrast level, and the covariance measures the correlation between windows in the two images, indicating how well the images align in terms of structure.

Table 1 presents the results obtained for the Places dataset, providing insights into the performance of our proposed model. Notably, our model achieved 0.08 SIFID value, outperforming SinGAN. Furthermore, when examining the diversity values, all models achieved similar levels of SSIM value. These findings highlight the ability of our model to effectively learn global dependencies in real images, resulting in the generation of highly realistic images. Notably, our proposed model also boasts a reduced parameter count and requires less training time, showcasing its superiority over SinGAN in terms of both performance and efficiency.

5 Applications

We tested our model on two different applications: paint-to-image and harmonization. For training images [27, 28] used in these applications, we conducted unconditional image generation training. Once the training phase was completed, each application went through the inference phase without any modifications or adjustments made to the model. The inference stage of the applications was carried out using the trained model as is, without any changes or fine-tuning.

5.1 Paint-to-image

Figure 5 presents the visual outcomes acquired for the paint-to-image application. Through the employing of a trained model, the objective of this application is to generate realistic drawings. A drawing is created that possesses the same textural and structural features as the training image. The generative model is trained on a single image. Once the model training is complete (i.e., in the test time), the drawing is provided to the model as input at a specified scale. The model's task is to generate a realistic version of the input drawing, progressing from the initial scale to the final scale. Notably, this application does not prescribe a specific input scale level. In the case of drawing, all available scales are consecutively employed as the input scale, and the model generates corresponding outputs. The choice of the input scale depends on the scale at which the model yields the most realistic output. In the domain of the paint-to-image application, the input scales range between 2 and 4. The selection of the input scale significantly influences the resulting output.

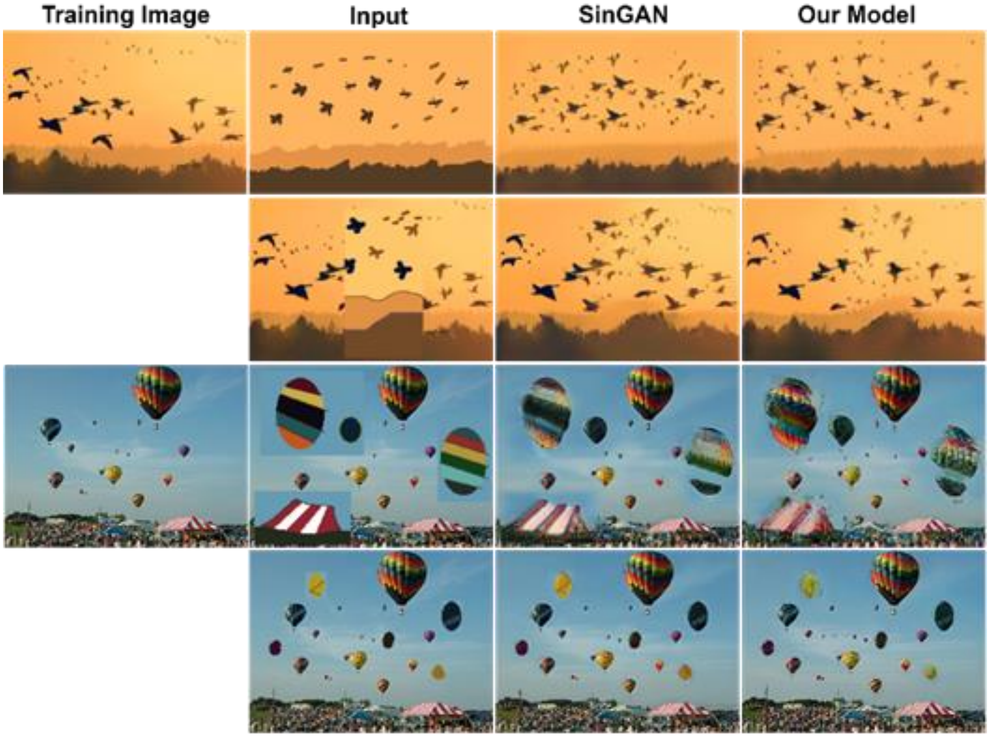

Figure 5. Paint-to-image.

Table 2. Quantitative results for paint-to-image.

	SinGAN [10]	Our Model
SIFID(↓)	3.2	3.1
SSIM(↑)	0.55	0.57

Two different implementations have been utilized to carry out this application. In the first implementation, the model receives the complete drawing as its input and produces the transformation of this drawing into highly plausible and visually coherent images. This approach showcases the model's ability to process the essential textural and structural features of the drawing.

In the second implementation, drawing patches are meticulously embedded within the training images. With this hybrid input, the trained mode attempts to generate more integrative and contextualized images. Figure 5 shows the visual outcomes of paint-to-images for two single training images. Successful outputs of different types of input images underscores the model's versatility and adaptability in handling different input scenarios, further underscoring its potential applicability across a range of use cases within the realm of computer vision and image generation. Table 2 indicates our model achieves similar SIFID and SSIM values with SinGAN for these images.

5.2 Harmonization

Harmonization is the process of modifying an inserted object within an image to match the image's visual structure. The application image is created by placing an object within the training image. In this context, the training image can be considered a background image. The application image is given as input at a certain scale to the model trained on the training image. The core objective of the model is to orchestrate a harmonious fusion of the inserted object with the background image, all while faithfully adhering to the characteristic traits inherent in the training image.

The degree to which the added object carries the characteristics of this background image indicates the level of realism achieved. Similar to the paint-to-image application, the scale at which the harmonized image is introduced to the model during the inference stage plays a pivotal role in shaping the outcome.

Figure 6 unfolds a comparative exposition of harmonization outcomes of our model in comparison to SinGAN and ConSinGAN. These results unveil the resounding success of our model in striking a balance that heightens the realism and naturalness of the inserted objects within the image. SIFID and SSIM values for harmonization reveals that our model Works successfully as other models, in Table 3.

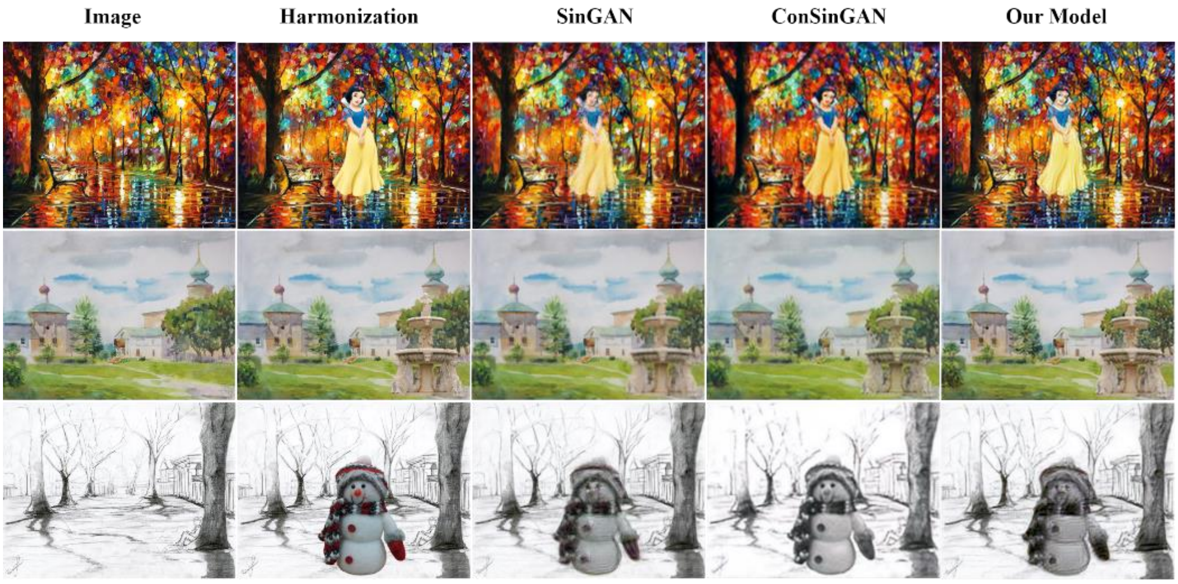

Figure 6. Harmonization.

Table 3. Quantitative results for harmonization.

	SinGAN [10]	ConSinGAN [17]	Our Model
SIFID(↓)	4.6	3.4	4.2
SSIM(↑)	0.24	0.26	0.24

6 Conclusions

In this study, we proposed a GAN model that operates on a training dataset. Our model is based on the SinGAN architecture but incorporates significant modifications. While SinGAN may struggle to capture global dependencies and semantic coherence in training images, our aim was to address this limitation. To achieve this, we introduced self-attention blocks in the intermediate layers of both the generator and discriminator networks. This approach facilitates a smoother transition between the global features learned at smaller scales to the local features learned at larger scales. As a result, our model becomes more capable of accurately detecting both local and global features in images. Furthermore, we employed a scaling function that ensures the dimensions of training images at intermediate scales are closer to each other. This approach allows the convolution filters, which remain fixed in size, to operate more effectively in capturing the details of objects at intermediate scales. The improved structure and scaling method of our model enable it to learn the structure, position, and realism of objects more effectively. We demonstrated through measurement results that having similar dimensions for intermediate-scale images and the use of self-attention blocks in these scales contribute to the enhanced performance of our model. Additionally, ability of our model to produce high-quality outputs in both paint-to-image and harmonization applications demonstrates its suitability for image manipulation tasks in addition to unconditional image generation. In conclusion, our proposed model exhibits potential for applications in various domains with limited or single data availability.

Conflict of interest

The authors declare that there is no conflict of interest.

Similarity rate (iThenticate): 3%

References

[1] S. J. Pan, Q. Yang, A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359, 2010. 10.1109/TKDE.2009.191

[2] C. Finn, P. Abbeel, and S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks. International conference on machine learning PMLR, pp. 1126-1135, Sydney, Australia 6-11 August 2017.

[3] C. Shorten and T. M. Khoshgoftaar, A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48, 2019. https://doi.org/10.1186/s40537-019-0197-0

[4] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz and S. Bengio, Generating sentences from a continuous space. 20th SIGNLL conference on computational natural language learning, CoNLL 2016. Association for computational linguistics (ACL), pp. 10-21, Berlin, Germany, 11-12 August 2016.

[5] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr and T. M. Hospedales, Learning to compare: relation network for few-shot learning. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1199-1208, Salt Lake City, USA, 18-23 June 2018.

[6] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville and Y. Bengio, Generative adversarial networks. Communications of the ACM, 63(11), 139-144, 2020. https://doi.org/10.1145/3422622

[7] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen, Improved techniques for training gans. Advances in neural information processing systems, Barcelona, Spain, 5-10 December, 2016.

[8] M. Arjovsky and L. Bottou, Towards principled methods for training generative adversarial networks. Advances in neural Information Processing Systems, Barcelona, Spain, 5-10 December, 2016.

[9] Z. Zhang, M. Li and J. Yu, On the convergence and mode collapse of GAN. SIGGRAPH asia 2018 technical briefs, pp. 1-4, Tokyo, Japan, 4-8 December 2018.

[10] T.R. Shaham, T. Dekel and T. Michaeli, SinGAN: Learning a generative model from a single natural ımage. Proceedings of the IEEE/CVF ınternational conference on computer vision, pp. 4570-4580, Seoul, Korea (South), 27 October-2 November, 2019.

[11] H. Zhang, I. Goodfellow, D. Metaxas and A. Odena. Self-attention generative adversarial networks. 36th international conference on machine learning PMLR, pp. 7354-7363, Long Beach, California, USA, 9-15 June 2019.

[12] E. Zakharov, A. Shysheya, E. Burkov and V. Lempitsky, Few-shot adversarial learning of realistic neural talking head models. Proceedings of the IEEE/CVF international conference on computer vision, pp. 9459-9468, Seoul, Korea (South), 27 October-2 November, 2019.

[13] M. Lučić, M. Tschannen, M. Ritter, X. Zhai, O. Bachem and S. Gelly, High-fidelity image generation with fewer labels. 36th international conference on machine learning PMLR, pp. 4183-4192, California, USA, 9-15 June 2019.

[14] A. Noguchi and T. Harada, Image generation from small datasets via batch statistics adaptation. Proceedings of the IEEE/CVF international conference on computer vision, pp. 2750-2758, Seoul, Korea (South), 27 October-2 November, 2019.

[15] A. Shocher, S. Bagon, P. Isola and M. Irani, InGAN: capturing and retargeting the 'DNA' of a natural ımage. Proceedings of the IEEE/CVF international conference on computer vision, pp. 4492-4501, Seoul, Korea (South), 27 October-2 November, 2019.

[16] P. Isola, J.Y. Zhu, T. Zhou and A.A. Efros, Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125-1134, Honolulu, HI, USA, 21-26 July, 2017.

[17] T. Hinz, M. Fisher, O. Wang and S. Wermter, Improved techniques for training single-ımage gans. Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV), pp. 1300–1309, January 5 – 9, 2021.

[18] D. Bahdanau, K. Cho and Y. Bengio, Neural machine translation by jointly learning to align and translate. Proceedings of the 3rd ınternational conference on learning representations, San Diego, CA, USA, 7-9 May, 2015.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need. Advances neural information processing systems 30, pp. 6000–6010, Long Beach, CA, USA, 4-9 December, 2017.

[20] H. Zhao, J. Jia and V. Koltun, Exploring self-attention for ımage recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10076-10085, Seattle, WA, USA, 13-19 June, 2020.

[21] I. Bello, B. Zoph, A. Vaswani, J. Shlens and Q.V. Le, Attention augmented convolutional networks. Proceedings of the IEEE/CVF international conference on computer vision, pp. 3286-3295, Seoul, Korea (South), 27 October-2 November, 2019.

[22] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang and H. Lu, Dual attention network for scene segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3146-3154, Long Beach, CA, USA, 15-20 June 2019.

[23] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A.C. Courville, Improved training of wasserstein gans. Advances in neural information processing systems 30, pp. 5769-5779, Long Beach, CA, USA, 4-9 December, 2017.

[24] D.P. Kingma and J. Ba, Adam: a method for stochastic optimization. Proceedings of the 3rd international conference on learning representations, San Diego, CA, USA, 7-9 May, 2015.

[25] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba and A. Oliva, Learning deep features for scene recognition using places database. Advances in neural information processing systems, pp. 487–495 Montreal, Quebec, Canada, 8-13 December, 2014.

[26] Düden şelalesi. https://www.kulturportali.gov.tr/contents/images/Yukar%c4%b1%20D%c3%bcden_Servet%20Uygun%20logolu.jpg, Accessed 10 September 2023.

[27] SinGAN github web site. https://github.com/tamarott/SinGAN/tree/master/Input/Images, Accessed 1 September 2023.

[28] ConSinGAN github web site. https://github.com/tohinz/ConSinGAN/tree/master/Images/Harmonization, Accessed 1 September 2023.

[29] 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. Advances in neural information processing systems 30, pp. 6629–6640, Long Beach, CA, USA, 4-9 December, 2017.

[30] Z. Wang, A.C. Bovik, H.R. Sheikh and E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity. IEEE Transactions on image processing, 13(4), 600-612, 2004. 10.1109/TIP.2003.819861