EXTENDING IMPLICIT NEURAL REPRESENTATIONS FOR TEXT-TO-IMAGE GENERATION

Guanming Liu[1], *Zhihua Wei*[1,†], *Heng Zhang*[1], *Rui Wang*[1], *Aiquan Yuan*[2], *Chuanbao Liu*[2], *Biao Chen*[2], *Guodong Cao*[2]

[1]Department of Computer Science and Technology, Tongji University
[2]Alibaba Group

ABSTRACT

Implicit neural representations (INRs) have demonstrated their effectiveness in continuous modeling for image signals. However, INRs typically operate in a continuous space, which makes it difficult to integrate the discrete symbols and structures inherent in human language. Despite this, text features carry rich semantic information that is helpful for visual representations, alleviating the demand of INR-based generative models for improvement in diverse datasets. To this end, we propose EIDGAN, an Efficient scale-Invariant Dual-modulated generative adversarial network, extending INRs for text-to-image generation while balancing network's representation power and computation costs. The spectral modulation utilizes Fourier transform to introduce global sentence information into the channel-wise frequency domain of image features. The cross attention modulation, as second-order polynomials incorporating the style codes, introduces local word information while recursively increasing the expressivity of a synthesis network. Benefiting from the column-row entangled bi-line design, EIDGAN enables text-guided generation of any-scale images and semantic extrapolation beyond image boundaries. We conduct experiments on text-to-image tasks based on MS-COCO and CUB datasets, demonstrating competitive performance on INR-based methods.

Index Terms— generative adversarial network, implicit neural representation, text-to-image generation, style modulation, cross attention, Fourier transform

1. INTRODUCTION

Generative Adversarial Networks (GANs) [1] have emerged as a prominent approach in the field of computer vision. One particular variant of GANs that has gained popularity is INR-based GANs. Unlike traditional GANs which utilize convolutional operations, INR-based GANs primarily use Multi-Layer Perceptions and take 2D coordinate locations (x, y) as input to generate RGB values for the corresponding locations in continuous images. However, many of them still employ a multi-scale training process similar to StyleGAN. The artificially defined multiple resolution layers significantly hamper the scale-consistency of INR-based GANs. Additionally, the simple weight modulation restricts the model's capacity to learn diverse image signals, as it does not adequately consider the frequency domain representation and only permits approximation of one-order polynomial function in the network.

Various visual features and patterns could be effectively described using discrete human language. Recently, it has been proven [2, 3] that the appropriate introduction of textual information can improve the image quality of different types of generative models. Nevertheless, existing T2I methods narrowly represent images as discrete 2D pixel arrays, which are cropped and quantized versions of the actual continuous natural signals. To delve deeper into continuous image generation with diverse signal characteristics, our research primarily focuses on the multi-modal Text-to-Image (T2I) generation. Extending INRs for T2I generation is a challenging task, which requires improving the model's expressivity while dealing with the discrepancy between the continuous representation of INRs and the discrete nature of linguistic elements.

In this paper, we focus on the style modulation methods as they improve the performance of GANs and stabilize generator with condition introduction insusceptible to mode collapse issues [4]. To this end, we design a novel dual modulation technique consisting of spectral modulation and cross attention modulation to improve the representation ability of INR-based GANs and integrate them with text features smoothly and efficiently. The dual modulation leveraging sentence and word features extracted from CLIP [5] enables full control over both local and global aspects of generated images. Besides, we maintain the inherent properties of INRs by exploiting thick bi-line representations, which help generate text-guided scale-equivariant images.

This paper's key contributions can be summarized as follows: (1) We propose EIDGAN, an Efficient Scale-equivariant Dual-modulated INR-based GAN for text-to-image generation, enabling scale-consistent interpolation outputs and image extrapolation beyond training resolution. (2) The dual modulation, consisting of spectral modulation and cross attention modulation, complementarily provides continuous spectral representations and high-degree polynomial representations, facilitating efficient style modulation with sentence and word features. (3) Extensive experiments show that EIDGAN achieves comparable performance on image fidelity and text-image alignment with significantly fewer parameters.

2. RELATED WORK

2.1. INR-based GANs

INR takes the coordinate information of signals (images, 3D shape, or audio signals) as input and outputs the value at the corresponding position to represent the target.

INR-based GANs dedicated to 2D image synthesis have acquired an unique transformation capability through coordinate gird

[†] Corresponding author: Zhihua Wei. The work is partially supported by the National Nature Science Foundation of China (No. 61976160, 61906137, 61976158, 62076184, 62076182) and Shanghai Science and Technology Plan Project (No.21DZ1204800) and Technology research plan project of Ministry of Public and Security (Grant No.2020JSYJD01).

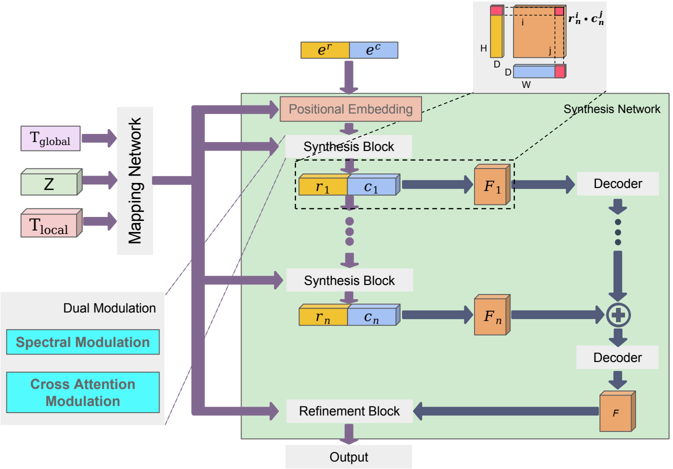

Fig. 1. EIDGAN architecture overview. The mapping network incorporates global and local text features with latent codes. The synthesis network, which employs a thick bi-line representation to improve memory-efficiency, incorporates dual modulation mechanism.

manipulation. INR-GAN [6] employs a factorized multiplicative modulation to enhance performance. Recently, some works such as AnyRes [7], ScaleParty [8], and Creps [9] make efforts on any-scale image generation leveraging benefits of INR-based GANs' design.

The work most similar to ours is HyperCGAN [10], which is also built upon an INR-based GAN architecture. HyperCGAN is the first job to explore Text to Continuous Image Generation (T2CI). However, the multi-scale structure leads to scale-inconsistency, weakening the ability to generate any-scale images. Also, HyperCGAN grapples with the expensive computation costs and high memory usage associated with full-resolution 2D feature maps.

2.2. Text-to-Image Generation

The T2I generation, extended with the maturity of uni-modal generation, has shown great potential in creating human-indistinguishable samples. Many generative models, such as GANs [1] and Diffusion [11], have been expanded for conditional image generation.

The GANs, which were developed earlier, had significantly improved in terms of sampling quality and diversity. Notably, StyleGANv2 [12] innovatively designed a mapping network to modulate the parameters of the synthesis network, enhancing the controllability of image generation. Gansformer [13], StyleFormer [14] and StyleSwin [15] incorporate attention mechanism [16, 17] with GANs to explore the probability of leveraging Transformer in the image generation domain. The T2I GANs can be seen as two stages. Previously, models such as AttnGAN [18] and DF-GAN [19] used specially trained text encoders to generate semantically consistent images. With the emergence of large-scale multi-modal pre-trained models like CLIP [5], many generative tasks [20] including T2I GANs become more powerful. Lafite [21] and StyleGAN-T [2] combines the StyleGANv2 model with CLIP to explore zero-shot T2I generation. Above all, the attention mechanism has increasingly assumed a vital role in T2I generation tasks.

3. METHODOLOGY

INRs model images as natural signals and regard the neural network as a continuous function constrained in image domain. The INR-based GANs use a generator G to form images by the pixel coordinates and latent codes $z \in Z$. Specially, given an image I of a resolution $H \times W$ and the RGB value $c \in [0,1]^3$, the synthesis network generates the image as:

$$I = \{G(x,y,z) \mid (x,y) \in mgrid(H,W), G(x,y,z) \mapsto c\} \quad (1)$$

$$mgrid(H,W) = \{(x,y) \mid 0 \leq x < W, 0 \leq y < H\} \quad (2)$$

where (x,y) is the pixel location sampled from the input coordinate grid related to resolution of the image. The $mgrid(H,W)$ represents the set of pixel locations.

3.1. Efficient Scale-equivariant Network Architectures

The EIDGAN architecture overview shows in Fig. 1. We employ a thick bi-line representation and design a new dual-modulated architecture backbone to efficiently improve the expressivity and smoothly introduce text features. To be specified, we introduce spectral blocks in both generator and discriminator to process image signals in the frequency domain, further aligning with the intrinsic nature of INRs. And we introduce an efficient cross attention modulation in the generator to enhance the multi-polynomial representation of the model. The memory-efficient representation would not change the resolution setting equipped with only linear layers and activation functions during training process.

Let e^r and e^c denote a single 2D grid of normalized (x,y) coordinates for row and column, respectively. The image generation can be rewritten as:

$$I = G(e^r, e^c, Z, T) \quad (3)$$

where T combines both global and local text features. The more details will be described next section.

The e^r and e^s are first processed into sinusoidal positional encoding, and then feed into synthesis block, getting two thick bi-line representation $r^{H\times D}$ and $c^{W\times D}$(D means the small thickness to enrich network's performance). As a result, the single feature map on block n is the dot product of the corresponding elements:

$$F_n = r_n \cdot c_n^\top \quad (4)$$

This composition process is illustrated in the top right corner of Fig. 1. For an input image at resolution $256 \times 256 \times 3$, using a small thickness of $D = 8$ results in a parameter size that is 1/32 of the original one. Although this representation may not enough to model all details of some complex images, it is adequate for constructing the overparameterized feature space according to [9].

The intermediate feature map F_n follows the coarse-to-fine design by StyleGANv2 [12]. The feature map fusion process is carried out in conjunction with the residual connection by the decoder, which links the full feature map output from each layer. Finally, the refinement block receives the final feature map and performs style modulation on the full size.

3.2. Dual Modulation for T2I Generation

Formally, let $X^{n\times d}$ denote input features with the shapes of n vectors and dimension d. On the image case, all coordinates are considered as an input sequence length, resulting in $n = W \times H$. In StyleGANv2's setting, a mapping network is used to turn latent codes $z \in Z^d$ into style codes $y \in Y^d, M : z \mapsto y$, which then modulate the weights of synthesis layers. A general formula for common modulation update rule can be defined as:

$$u^c(X,Y) = \mathrm{Mod}(W,Y) \cdot X \quad (5)$$

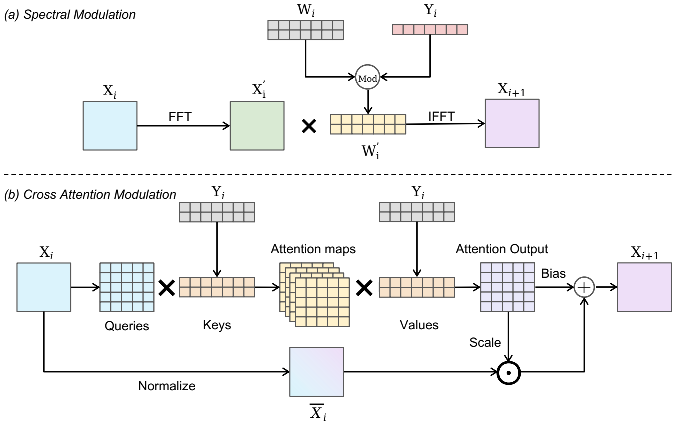

Fig. 2. Dual Modulation, including Spectral Modulation (a) and Cross Attention Modulation (b). Sentence features concatenate with style codes, and word features calculate cross-attention with them.

where W represents weight matrix at different layers. The modulation function Mod is commonly expressed using Hadamard product.

But this modulation method causes style codes to impact only spatial image features, which restricts the capability of image generation. We design the dual modulation to improve model's representation power while effectively fusing sentence features and word features. The dual modulation consisting of spectral modulation and cross attention modulation is illustrated in the Fig. 2.

Motivated by [22], spectral modulation aims to incorporate style codes with frequency domain images. INRs demonstrate that employing only MLP layers within the channel dimension can markedly help augment the discriminability of global representations, contributing to performance improvements. The spectral modulation converts physical space into spectral space using a 1D Fast Fourier Transform (FFT) layer. Particularly, FFT translates image features into frequency domain, providing a smoother and more continuous spectral representation X'. Then style codes are trained to determine the weights of decomposed frequency components. Lastly, Inverse Fast Fourier Transform (IFFT) brings the spectral features back to physical space. The spectral modulation can be defined as:

$$u^s(X, Y) = \mathrm{IFFT}(\mathrm{Mod}(W, Y) \odot \mathrm{FFT}(X)) \tag{6}$$

At the text-to-image setting, the sentence features bypass the mapping network and concatenate with style codes.

The self-attention mechanism can be viewed as one variant of third-order polynomials, and it can also enhance the expressivity by capturing long-range dependencies for both image-only and image-text relationships. Nonetheless, the expansive cost of attention calculation remains a limitation for regular resolutions. Instead of modulated self attention mechanism [14, 15], the cross attention modulation performs $\mathrm{Attention}$ calculation between the lower-rank style codes and bi-line feature maps, reducing the computational cost of self-attention, especially in high-resolution feature maps. Also, the cross attention modulation can be viewed as second-order polynomials, helping to model high-dimension distributions effectively. The cross attention modulation can be defined as:

$$u^a(X, Y) = \gamma(a(X, Y)) \cdot \overline{X} + \beta(a(X, Y)) \tag{7}$$

$$a(X, Y) = \mathrm{Attention}(q(X), k(Y), v(Y)) \tag{8}$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ are linear layers that compute scale factors and bias factors, and $\overline{X} = \frac{X-\mu(X)}{\sigma(X)}$ normalizes the features of X using means and variances. The $q(\cdot)$,$k(\cdot)$ and $v(\cdot)$ are linear layers that map elements into queries, keys and values. Intuitively, the style codes could contribute to the generation of spatial attended regions, significantly enhancing the controllability of style modulation.

At the text-to-image setting, we extend Z from $1 \times d$ to $k \times d$. The word features perform cross attention modulation with style codes, and the style codes perform cross attention modulation with feature maps, establishing style codes as the immediate bridge.

Spectral modulation and cross attention modulation allow each prompt to get fine-grained control over the image generation. Based on CLIP's multi-modal alignment capability, we extract the sentence features $T^{1\times d}_{global}$ from CLIP text encoder. Furthermore, we extract word features $T^{l\times d}_{local}$ (the tokenized prompts' maximum sequence length l is 77 at CLIP's settings) from the penultimate layer of a frozen CLIP text encoder. We incorporate them with spectral modulation and cross attention modulation, termed as dual modulation.

4. EXPERIMENTS

4.1. Experimental Setup

4.1.1. Datasets.

We conduct experiments on two challenging T2I datasets: CUB bird [23] and MS-COCO [24]. The CUB dataset is challenging in fine-grained object generation, and the MS-COCO dataset is challenging in the diverse generation of complex scenes and multiple objects.

4.1.2. Evaluation Metrics.

We adopt the $Fr\acute{e}chet$ Inception Distance (FID) [25] to evaluate the image fidelity, and adopt the CLIP-R [26] and CLIP-S [27] to evaluate text-image semantic consistency. Since HyperCGAN is not open-source, we use the ratio of the clip-R of the generated results to the clip-R of the real datasets, denoted as CLIP-R', for comparison.

4.1.3. Implementation Details.

We adopt a standard GAN training procedure similar to the StyleGANv2 framework to formulate a conditional image generation baseline. We choose the ViT-B/32 [5] model as the frozen CLIP model in EIDGAN. We reduce the dimension of latent codes to 64 similar to [2] to save parameters and memory usage.

First, We utilize a wavelet discriminator to detect the periodic artifact pattern in the spectral domain. Slightly different from StyleSwin [15], we let the first layer of the discriminator process input images at their original resolution. Then, we hierarchically downsample the images and apply discrete wavelet transformation to examine the frequency discrepancy of multi-scaled input.

Second, we use additional contrastive losses to ensure GAN's feature space is supervised by the multi-modal alignment of the pretrained CLIP. We enforce the discriminator-extracted fake image feature $f_D(x)$ aligned with CLIP processed image feature f^I_{clip} by contrastive regularizer as follow:

$$L_{ConD} = -\tau \frac{exp(Sim(f_D(x_i), f^I_{clip}(x_i))/\tau)}{\sum_{j=1}^{n} exp(Sim(f_D(x_j), f^I_{clip}(x_i)/\tau)} \tag{9}$$

where $Sim(\cdot)$ denotes the cosine similarity, τ is a non-negative hyper-parameter and x_i is the real image sample.

And we further utilize contrastive loss for generator to improve the semantic correspondence between the synthetic image sample x' and the sentence features extracted from CLIP text encoder f^T_{clip}:

$$L_{ConG} = -\tau \frac{exp(Sim(f_{clip}(x'_i), f^T_{clip}(t^{global}_i))/\tau)}{\sum_{j=1}^{n} exp(Sim(f_{clip}(x'_j), f^T_{clip}(t^{global}_i)/\tau)} \tag{10}$$

Table 1. T2I comparison results. The best results are indicated in bold, and the second best results are indicated with an underline.

Model	Generator	COCO			CUB			Scale-consistent	Beyond-boundary
	Parameters	FID ↓	CLIP-S ↑	CLIP-R' ↑	FID ↓	CLIP-S↑	CLIP-R' ↑	Interpolation	Extrapolation
AttnGAN [18]	230M	35.49	-	32.77%	23.98	-	<u>119.19%</u>		
DF-GAN [19]	<u>19M</u>	21.42	0.2920	29.22%	<u>14.81</u>	0.2972	107.97%		
LAFITE [21]	50M	**8.21**	**0.3335**	-	**14.58**	**0.3125**	-		
INR-GAN$_{\text{CLIP}}$ [6]	73M	31.21	0.2807	68.44%	28.49	0.2721	90.52%		✓
CREPS$_{\text{CLIP}}$ [9]	31M	35.90	0.2882	<u>73.87%</u>	22.95	0.2864	91.78%	✓	✓
HyperCGAN$^{\text{word}}_{\text{CLIP}}$ [10]	65M	27.21	-	71.55%	16.48	-	72.59%		✓
EIDGAN	**18M**	<u>20.16</u>	<u>0.3027</u>	**87.14%**	16.36	<u>0.3074</u>	**120.22%**	✓	✓

Table 2. The ablation study of EIDGAN on CUB dataset.

Configuration	FID ↓	CLIP-S ↑
EIDGAN	16.36	0.3074
w/o Cross attention modulation	19.23	0.2907
w/o Spectral modulation	18.83	0.3020
w/o Dual modulation	22.95	0.2864
Unconditional baseline	28.07	-

4.2. Comparisons

We compare the efficiency of our EIDGAN with other text-to-image models on the generator parameters. The results are shown in Table 1. As a continuous T2I model, we also evaluate the performance of our EIDGAN with several methods in T2I generation [6, 9, 10, 18, 19, 21], including some state-of-the-art methods and self-implemented INR-based methods. We establish a baseline named INR-GAN$_{\text{CLIP}}$ using INR-GAN [6] architecture and CLIP's contrastive loss for T2I generation, compared to the adopted baseline CREPS$_{\text{CLIP}}$, which employs the same CLIP's contrastive loss. As a result, the higher CLIP-S or CLIP-R' score among INR-based methods indicates that the dual modulation contributes to the text-image alignment. Even though our EIDGAN can not beat all of the other models on performance, the proposed model reaches a competitive result without the convolution operation and with very small number of parameters. Moreover, EIAGAN has the intrinsic ability of INRs especially text-guided scale-inconsistent interpolation and beyond-boundary extrapolation.

4.3. Exploring the Properties

4.3.1. Scale-consistent Interpolation Outputs.

We train our models on 256 × 256 images once and this model could generate text-guided scale-consistent examples by changing the sampling rate of gird coordinates in the positional encoding. The super-resolution examples are shown in the right sub-figure of Fig. 3, comparing the scaled 512×512 image with the original 256^2 image using bilinear interpolation. The EIDGAN can produce the same semantic details but sharper when increasing the scale.

4.3.2. Beyond-Boundary Extrapolation Outputs.

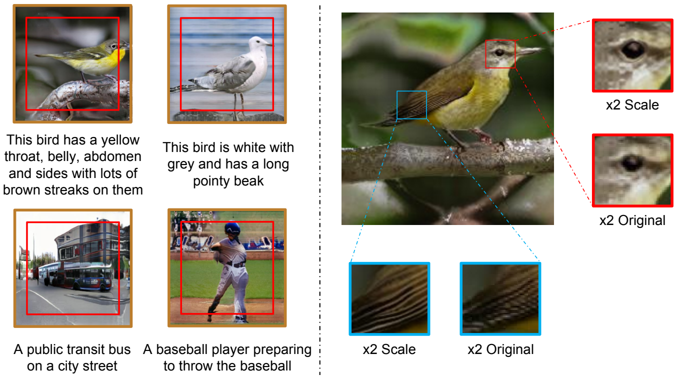
This bird has a yellow throat, belly, abdomen and sides with lots of brown streaks on them
This bird is white with grey and has a long pointy beak
A public transit bus on a city street
A baseball player preparing to throw the baseball
x2 Scale
x2 Original
x2 Scale
x2 Original

Fig. 3. Beyond-boundary extrapolation and Scale-consistent interpolation outputs. We only train models at a 256 × 256 setting.

The left sub-figure of Fig. 3 shows the ability of EIDGAN to extrapolate beyond image boundaries. After training on a regular coordinate grid, EIDGAN can generate meaningful and context-related images, producing out-of-the-region results within a suitable range. The image with description "This bird has a yellow throat..." can be extended to generate the yellow throat and include a complete bird beak outside the original boundary. The image with description "A public transit bus..." can be extended to generate the continuous street mentioned by the description. The image with description "A baseball player..." can also be extended to complete player's feet. The semantic extrapolation results demonstrate the effectiveness of continuous text-image alignment.

4.4. Ablation Study

To verify the effectiveness of EIDGAN, we conduct an ablation study on the CUB dataset in Table 2. The worst result of unconditional baseline indicates text features help improve image quality. The spectral modulation helps achieve better performance on image quality than regular weight modulation. The cross attention modulation, introducing word features while also enabling second-order polynomial representation, helps achieve better performance on both text-image alignment capability and image quality. The dual modulation together enhances the text-to-image generation ability.

5. CONCLUSIONS

In this paper, we present EIDGAN to extend INRs for text-to-image generation. Our findings demonstrate that style modulation plays a crucial role in integrating local and global textual information while preserving the inherent properties of INR-based GANs. The proposed dual modulation, consisting of spectral modulation and cross attention modulation, aims to enhance the ability of continuous modeling and high-degree polynomial representation while leveraging both sentence and word features effectively. EIDGAN achieves a significant advancement in the field of T2CI generation and provides an efficient and scalable solution for generating high-quality, resolution-independent images conditioned on text inputs.

6. REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al., "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[2] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila, "Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis," *arXiv preprint arXiv:2301.09515*, 2023.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[4] Ting Chen, Mario Lucic, Neil Houlsby, and Sylvain Gelly, "On self modulation for generative adversarial networks," *arXiv preprint arXiv:1810.01365*, 2018.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021, pp. 8748–8763.

[6] Ivan Skorokhodov, Savva Ignatyev, and Mohamed Elhoseiny, "Adversarial generation of continuous images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10753–10764.

[7] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang, "Any-resolution training for high-resolution image synthesis," in *European conference on cmputer vision*, 2022, pp. 170–188.

[8] Evangelos Ntavelis, Mohamad Shahbazi, Iason Kastanis, Radu Timofte, Martin Danelljan, and Luc Van Gool, "Arbitrary-scale image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11533–11542.

[9] Thuan Hoang Nguyen, Thanh Van Le, and Anh Tran, "Efficient scale-invariant generator with column-row entangled pixel synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22408–22417.

[10] Kilichbek Haydarov, Aashiq Muhamed, Jovana Lazarevic, Ivan Skorokhodov, and Mohamed Elhoseiny, "Hypercgan: Text-to-image synthesis with hypernet-modulated conditional generative adversarial networks," 2021.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[12] Tero Karras, Samuli Laine, Miika Aittala, et al., "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[13] Drew A Hudson and Larry Zitnick, "Generative adversarial transformers," in *International conference on machine learning*, 2021, pp. 4487–4499.

[14] Jeeseung Park and Younggeun Kim, "Styleformer: Transformer based generative adversarial networks with style vector," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8983–8992.

[15] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo, "Styleswin: Transformer-based gan for high-resolution image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11304–11314.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] Alexey Dosovitskiy, Lucas Beyer, et al. Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International conference on learning representations*, 2021.

[18] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, and Huang, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1316–1324.

[19] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu, "Df-gan: A simple and effective baseline for text-to-image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16515–16525.

[20] Peng Wu, Xiankai Lu, Jianbing Shen, and Yilong Yin, "Clip fusion with bi-level optimization for human mesh reconstruction from monocular videos," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 105–115.

[21] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun, "Towards language-free training for text-to-image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17907–17917.

[22] Badri N Patro, Vinay P Namboodiri, and Vijay Srinivas Agneeswaran, "Spectformer: Frequency and attention is what you need in a vision transformer," *arXiv preprint arXiv:2304.06446*, 2023.

[23] Catherine Wah, Steve Branson, Peter Welinder, and Perona, "The caltech-ucsd birds-200-2011 dataset," 2011.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[26] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach, "Benchmark for compositional text-to-image synthesis," in *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 1)*, 2021.

[27] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan, "Nüwa: Visual synthesis pre-training for neural visual world creation," in *European conference on computer vision*, 2022, pp. 720–736.