Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 4, No. 1, 2025

Enhancing Vision Transformers for Image Generation: A Hierarchical GAN with Triplet Attention and Consistency Regularization

Jericho Whitman

School of Electrical and Computer Engineering, University of Arizona, Tucson, USA

jericho89@arizona.edu

Abstract: This paper proposes a novel hierarchical generator adversarial network based on Vision Transformers (ViT) for unconditional image generation. To address common challenges such as structural inconsistency and unstable training in GANs, we introduce the Triplet Attention mechanism within the generator, enhancing the structural soundness of generated images without increasing the model's parameter size. Additionally, a consistency regularization term is integrated into the loss function, improving the training stability and robustness to noise while mitigating overfitting. The effectiveness of the proposed method is demonstrated through extensive experiments on the CIFAR-10 and STL-10 datasets, where our framework outperforms TransGAN and other CNN-based GANs in both FID and IS metrics. Despite the simplicity of our architecture, which contains only three transformer layers, we achieve promising results, laying the groundwork for further enhancements in high-resolution image generation.

Keywords: Image Generation; Generative Adversarial Networks; Transformer.

1. Introduction

Generative Adversarial Nets (GAN) [1], as one of the most interesting models in the field of computer vision, has attracted a lot of interest and attention from researchers. Today, there are various generative models that are used for different tasks, such as image generation, image style transfer, text to image, image super-resolution, 3D reconstruction, and even video generation. Generative models not only play a huge role in image processing applications, but also play a crucial role in the development of Artificial Intelligence (AI) and Deep Learning (DL)[2]. It is well known that DL requires a large number of samples for training, and the traditional manual collection methods have been difficult to meet the demand in terms of quantity and huge cost. Thus, the sample generation task is like adding a constant stream of fuel to the speeding train of DL.

Early generative models include Restricted Boltzman Machine(RBM)[3] that have intractable likelihood functions and VAE[4] that based on the variational Bayesian inference. In 2014, Goodfellow et al. proposed GAN, which was trained with backpropagation and no need for any Markov chains[5] or unrolled approximate inference networks[6,7], which makes it stand out from the rest of the models. However, it is difficult for GAN to control the generated results for complex datasets. Therefore, CGAN put a conditional variable c together with both random variable z and real data x to guide the data generation process, and it can inspire subsequent tasks such as image style transfer, text to image etc. Based on CNN and GAN, A. Radford et al. proposed DCGAN[8], which achieved good performance in computer vision(CV) field. Image Transition task is an important

subsequent task of image generation. Pix2pix collects a same dataset in two different styles and one of two styles plays as the conditional input with U-NET[9] and PatchGAN[10]. CycleGAN[11] generates samples for twice to eliminates the requirement for matching images in target domain. StarGAN[12] learns among multi-domains and gains a surprising performance. SyleGAN[13] can separately control different factors of the image appearance. Text image generation tasks have become increasingly interesting in recent years. DCGAN[8] can be used to generate naked-eye acceptable images from text descriptions. StackGAN[14] gave a two-stage generation approach to improve the resolution and stability of text-generated images. Seq2Seq[15] generates images that include the spatial layout of multiple objects and the attributes of each object, including pose, expression, etc.

In recent years, although GANs have yielded good results in recent years for different tasks, its development is often accompanied by some tricky problems, such as model collapse, unstable training, and unreasonable structure of generated images etc. WGAN[16] solves the model collaps and the gradient disappearance problems by modifying the distance in the loss function. ProGAN[17] combines layer-by-layer generation from low- resolution to high-resolution with smooth embedding to improve the stability of the training process for high-resolution image generation. In order to capture the global dependencies from images, attention mechanism was introduced in CV[18,19], SAGAN[20] introduces self-attention mechanism[21,22] to

GAN for generating images and it is effective in modeling long-range dependencies.

Inspired by the properties of the attention mechanism, transformer[23] has made a huge breakthrough in the field of NLP. By introducing transformer to the CV field, Sharir G. el al. proposed ViT[24] that made a major breakthrough by introducing transformer to the CV field. Recently proposed TransGAN[25] used two pure Transformer to generate high-resolution images on GAN architecture, which can surpass some then-popular CNN-based GAN. TransGAN aims to be the first pilot study to build a pure transformer-based GAN, only using some of the techniques necessary to confirm the advantages of transform-based GAN. In subsequent research, many researchers also made many improvements on this foundation. In addition, the transformer-based structure requires more data compared to CNNs, and experiments indicate that a significant advantage can only be shown only on large-scale datasets.

Since the data in the hidden layer is presented as multiple channels, and it also contains a large amount of information on different channels, it is not enough to carry out attentional computation in the spatial dimension alone. However, too much attention mechanism is bound to seriously affect the computation speed. Recently Diganta M. et al.[26] have proposed a lightweight but effective attention mechanism called triplet attention, by which different network architectures with almost no parameter growth can obtain better classification results. In addition, some researches indicate that there are two significant problems in DL for a long time. One is that the model can easy to cause over-fit and the other one is that the results may be affected when a model is affected by a tiny noise. To solve these problems, semi-supervised learning introduces consistent regularization[27,28] Inspired of that, we introduce the bCR[29] into our loss function in this paper, in which the results are free from the affects of noises and the over-fitting is avoided as far as possible.

To sum up, our contributions are summarized as follows:

1.The Triplet-attention mechanism is introduced to the generator to increase the structural soundness of the generated images without increasing the size of the parameters.

2.The loss function is improved by adding consistency regularization to make the training process more stable.

3.The proposed approach is evaluated on two public datasets including CIFAR-10 and STL-10, where FID and IS evaluation metrics of the proposed framework outperformed TransGAN as well as some CNN-based GANs.

2. Related Works

2.1 Generative Adversarial Nets (GANs)

The GANs usually contains at least two neural network models. One is called generator (G) and the other is called discriminator (D). The role of G is to learn to capture the distribution Pdata of real data x and generate a new data G(z), and D acts like a classifier to give the probability that its input data, G(z) and x, is from the real data set and and provide feedbacks for the learning of G through the loss function. The goal of G is to make the generated data distribution Pz as similar as possible to the real data distribution so that D believes that the generated data comes from the real data set, and the goal of D is to distinguish the true source of its input data, G(z) and x, as much as possible. D and G play the two-player minmax game until a Nash equilibrium is reached. The loss function can be expressed as follow Equation:

$$\min_{G}\min_{D}\max V(G,D) = E_{x\sim P_{data}}[\log D(x)] + E_{z\sim P_z}[\log(1-D(G(z)))]$$

During the training process, G and D are trained alternately until eventually D is unable to distinguish the source of the data. The original GAN is implemented by Fully Connection (FC) Networks and piecewise linear units. As more and more researchers have proved that the development of CNN is better than FC [30,31], DCGAN replaces all fully connected layers with CNN on the GAN, and uses batch normalization (BN) [32] to accelerate convergence and reduce over-fitting. This is the first GAN network structure with full convolution, which improves the resolution of the generated images. ProGAN adopts a progressive training network architecture from low to high resolution, and performs smooth embedding and pixel-wise normalization between different resolutions to ensure the stability of training. Meanwhile, ProGAN removes the BN layer and only adds mini-batch to the last layer of D to improve the diversity of the generated images. Since CNN networks usually have small convolutional kernels, it is necessary to stack multiple layers of convolution for obtaining long-range dependencies. However, multiple layer convolutions make it difficult to optimize the algorithm and parameters. SAGAN introduces Self-attention mechanism to obtain dependencies at a distance at one level instead of multi-layer convolution operation. Meanwhile, SAGAN also introduces Spectral normalization for both G and D to stabilize the training process. Following the confirmation of the good performance of transformer on image classification tasks [24,33], TransGAN is a pure transformer-based GAN. Its G can incrementally increase feature resolution while reducing embedding size. It also introduces multi-task co-training for G with self-supervised auxiliary loss, and localized initialization for self- attention to make the training better.

2.2 Vision in Transformer (ViT)

Transformer has become the model of choice in Natural Language Processing (NLP) by 2021, such as BERT [34] andGPT-3[35]. However, there is no good breakthrough in the CV field. Although Cordonner et al. [36] had already applied transformer to CV tasks before ViT was published, ViT's model was simpler and more effective and scalable in comparison. Therefore, it became a milestone for transformer in the CV field, and also attracted extensive

research. Before ViT, it was widely believed that CNNs were more suitable for processing image information than transformers, because the structural properties of transformers made it possible to only process sequential information.

If a visual problem is transformed into a seq2seq problem, patch embedding is necessary. ViT divides the input image into multiple patches, and then projects each patch into a fixed-length vector. ViT also needs to join the positional encoding, which can be understood as a list of N rows (N is the length of the input sequence), and each row represents a vector with the same dimension as the input sequence embedding. Then the patch and position embedding are acted as input to combine into transformer blocks. In a transformer block, it is worth mentioning the multi-head attention, which is to improve the information extraction ability, and therefore combines the results of attention operation of these ways. The input and output dimensions of the transformer block are the same, in which multiple layers can be stacked to increase the depth of the network.

3. Methodology

In this section, we introduce our architecture, which consists of a Transformer-based hierarchical generator, triplet attention and balanced consistency regularization in loss function.

3.1 Transformer-based Hierarchical Generator

We construct a hierarchical generator based on the transformer, as shown in Figure 1, where we take a 256-channels random noise $z \in N(0,1)$ as the input signal and feed it into the generator. Through an MLP, we reshape a 1-dimensional sequence into a 2-dimensional image feature $Xi \in RHi \times Wi \times C$, and then through a learnable position embedding layer, the tensor is fed into the transformer blocks as the input signal. In the transformer blocks, we use a 4-head attention mechanism to obtain the relevance of the data, and then reshape the feature as $Xi \in R2Hi \times 2Wi \times C/2$. In this way, we boost our feature map layer by layer by pixel-shuffle after the fist and second blocks. Finally, after a triplet attention, we convert our output signal into a 32 × 32 resolution image with RGB 3 channels using a convolutional layer.

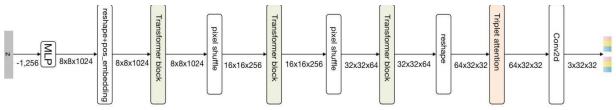


Figure 1. The architecture of our generator

3.2 Pixel-shuffle

Considering the full utilization of channel information, we use pixel shuffle for all of the upsampling processes, moving pixels from the channel dimension to the aspect dimension to achieve upsampling. This implementation process is not to generate high-resolution image directly by interpolation, but to get the feature map of r2 channels by convolution (r is the upscaling factor i.e. image expansion ratio), whose size is the same as the input low-resolution image, and then get this high-resolution image by aperiodic shuffling method.

3.3 Triplet Attention

Triplet attention[26] consists of 3 parallel Branches, in which two are responsible for capturing the cross-dimensional interactions between channel ***C*** and space ***H/W***, and the last one is used to construct Spatial Attention. The outputs of the final 3 Branches are aggregated using averaging. The advantage is that it is possible to model inexpensive but effective channel attention without involving any dimensionality reduction. Unlike CBAM[37] and SENet[38], which require a certain number of learnable parameters to build dependencies between channels, Triplet attention models channel attention and spatial attention almost parameterlessly. It first transfers the input tensor $X_i \in R^{C \times H \times W}$ into three branches. In each branch, a two-by-two interaction is established between C and ***H/W***. It can be represented by the following Equation:

$$y = \frac{1}{3}\left(\overline{y_1} + \overline{y_2} + y_3\right)$$

where the y_1 represents the interactions between ***H*** and ***C***, and y_2 represents the interaction between C and W, both with the 90° clockwise rotation to retain the original input shape of $C \times H \times W$, y_3 represents the interaction between ***H*** and ***W*** without any rotation.

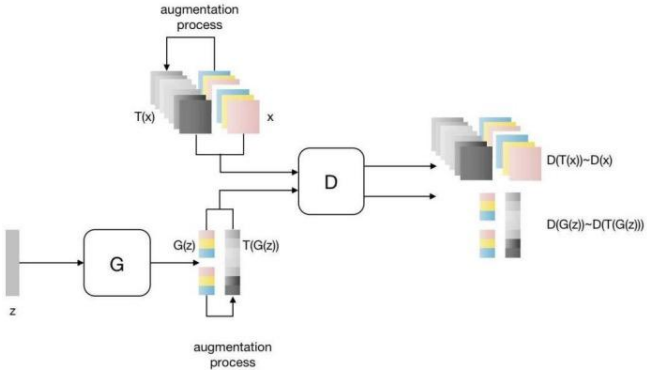


Figure 2. The architecture of our GAN

3.4 Loss Function with Balanced Consistency Regularization

According to CR-GAN[39], an encouraged discriminator can produces similar outputs for an image and their augmentation according to the characteristics of consistency regularization. So we fed real images x and their augmented ones T(x) into the discriminator. However, when augmented images contain visual artifacts during the process of data augmentation, it can also result in generated images. To avoid this phenomenon, we also put augmented generated images T(G(z)) into discriminator,

which is shown Figure 2. We also introduce the loss function of WGAN-GP[40] to control the gradient during the training process. Therefore, the total loss function of discriminator is given as follow:

$$L_{dis} = L_{wgangp\text{-}eps} + \lambda(D(G(z)) - D(T(G(z)))) + \lambda(D(x) - D(T(x)))$$

where Lwgangp-eps is following the setting of WGAN and WGAN-GP, λ equals to 1.

4. Experiments

4.1 Implementation Details

Firstly, we choose the datasets of CIFAR-10 and STL-10 for "small-scale" dataset image generation task. There are 60,000 images with 32×32 resolution in the CIFAR-10 dataset, which consists of 50,000 training samples and 10,000 testing samples. And the STL-10 dataset consists of 100,000 unlabeled images with 96×96 resolution(we resize it to 48×48). We set batch size 32 for generator while 26 for discriminator. The augmentation strategy is flip and crop. All experiments are set with 2 2080Ti GPUs. Learning rate for both G and D is 1e-4.

4.2 Evaluation Metrics

We select two evaluation metrics to measure the performance of the algorithm, which are respectively Inception Score (IS) [41] and Fréchet Inception Distance (FID) [42].

The definition of IS is given as follows:

$$IS(G) = \exp\left(E_{x\sim P_g} D_{KL}(p(y|x)\|p(y))\right)$$

where $\boldsymbol{E_{x\sim Pg}}$ means to iterate through all the generated samples and find the average. $\boldsymbol{D_{KL}}$ represents the ***KL*** divergence, so $\boldsymbol{D_{KL}(P\|Q)}$ denotes the degree of approximation between $\boldsymbol{P}$ and $\boldsymbol{Q}$. $\boldsymbol{p(y|x)}$ means that for picture $\boldsymbol{x}$, the probability distribution of belonging to all categories and $\boldsymbol{p(y)}$ is the Marginal probability. We want to generate images that are clear enough and generate a variety of categories, the larger the IS the better.

The definition of FID is given as follows:

$$FID = \|\mu_x - \mu_g\|^2 + Tr\left(\sum x + \sum g - 2\sqrt{\sum x \sum g}\right)$$

where $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_g}$ represents the mean value of the features of the real image and generated image respectively. ∑ represents the covariance matrix of the image. ***Tr*** is the Trace. Unlike IS, FID considers more the distance between the generated image and the real image, and the smaller the distance, the better the generated model is.

4.3 Results

The generated samples and samples from real datasets both on CIFAR-10 and STL-10 are shown in Figure 3 and Figure 4:

(a) By our method

(b) Real samples

Figure 3. On CIFAR-10 dataset

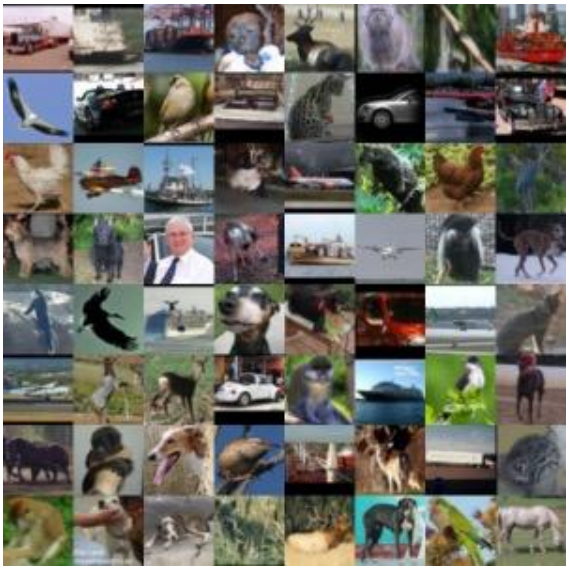
The figure above shows that the samples generated by our model on the CIFAR-10 dataset can be easily recognized as belonging to their respective categories, with very few instances of blurriness or distortion.

As demonstrated in the figure above, our model is capable of generating samples on the STL-10 dataset that produce images which are fairly realistic, but with some degree of distortion.

In order to illustrate the effectiveness of the proposed method, we selected several representative methods as the comparison algorithms, which include TransGAN[25], WGAN-GP[41], SN-GAN[43], AutoGAN[44], AdversarialNAS-GAN[45], ProGAN[17], and StyleGANs[46]. For CIFAR-10 dataset, our GAN's FID is

at the best, and only StyleGANs surpass our method on IS as shown in Table 1, our model scored more than 27% of the score of TransGAN on FID.

(a) By our method



(b) Real samples

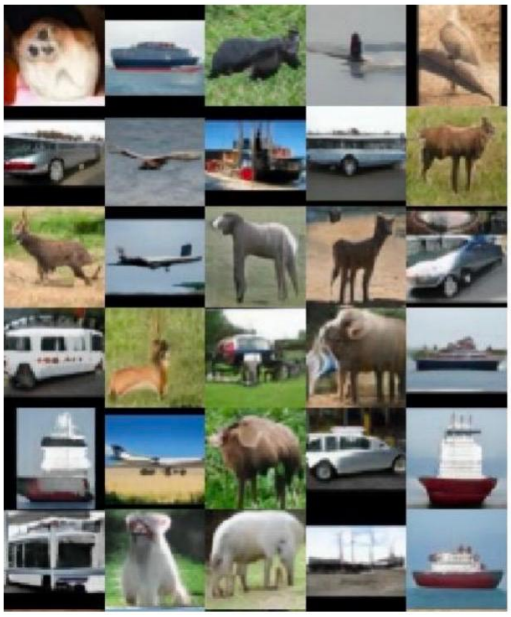
Figure 4. On STL-10 dataset

For STL-10 dataset(48 × 48), since a deeper network and grid self-attention are used for generative tasks with a resolution higher than 32 × 32 in the original TransGAN and our network contains only three layers without grid self-attention, which is a very simple architecture. Therefore, our model's generated samples were not superior to TransGAN's performance on the STL-10 dataset. However, our model did perform better than some of the CNN-based GANs.

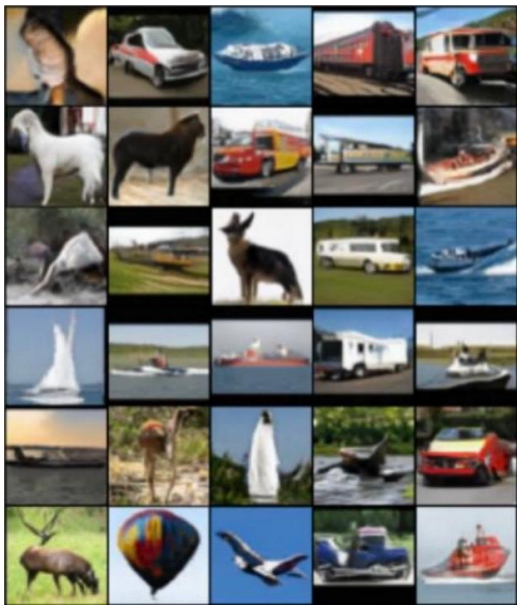
Table 1. FID and IS between some previous GANs and ours

Methods	CIFAR-10		Scheme 3	
	FID↓	IS↑	FID↓	IS↑
WGAN-GP	39.68	6.49±0.09	-	-
SN-GAN	-	8.22±0.05	40.1	9.16±0.12
AutoGAN	12.42	8.55±0.10	31.01	9.16±0.12
AdversarialNAS-GAN	10.87	8.74±0.07	26.98	9.63±0.19
ProGAN	15.52	8.80±0.05	-	-
StyleGAN-V2	11.07	**9.18**	20.84	10.21±0.14
TransGAN	9.26	9.02±0.12	**18.28**	**10.43±0.16**
ours	**7.68**	9.05±0.08	27.43	9.23±0.2

We present a comparison of the sensory perception between the samples generated by our method and those produced by TransGAN on STL-10 dataset, as depicted in Figure 5.



(a) By our method



(b) By TransGAN

Figure 5. Comparison on the STL-10 dataset

Based on the figure shown above, although it is noticeable that our model's generated samples display a gap in performance compared to TransGAN's generated results in terms of FID and IS, however, the perceptual difference is not very obvious.

4.4 Ablation Studies

To evaluate the performance of Balanced Consistency Regularization(bCR) [29], Triplet Attention [26], we separately adding these techniques to the basic model and compare their FID score and IS score on CIFAR-10 dataset, which is shown in Table 2. They achieve the best performance on FID. However, when we add the triplet attention, the performances are dropped on IS. The ablation study shows that the model with Consistency Regularization has a significant score improvement on both FID and IS, i.e., 48% and 5% respectively. The performance of Triplet attention is 2% upper and 1% lower on our model with Consistency Regularization. This may be because the measurement of IS is different from that of FID. When we add Triplet attention in our generator, it surely can gain information well on space and channels. So, the FID score performs better, which means that the generated images are more similar to the real images, but the categories may be more homogeneous or the clarity may be affected.

Table 2. The effects of different techniques in our method on CIFAR-10 dataset

Methods	FID↓	IS↑
Ours original	11.65	8.73±0.12
Ours + bCR	7.86	9.21±0.11
Our + bCR + Trip-att	7.68	9.05±0.08

The generated samples for the three methods are shown in Figure 6.

(a) Our method without Triplet-attention and bCR

(b) Our method with bCR but without Triplet-attention

(c) Our method with bCR and Triplet-attention

Figure 6. Samples by different methods in our model

We also try to change the position of Triplet attention, and the results are shown in Table 3. Our final method, i.e., put the Triplet attention after the last layer, gains the best FID score. Although the best IS is to put the Triplet attention after every pixel shuffle layer, but it is too expensive for our device.

Table 3. The results of Triplet attention in different positions of our generator

Positions of Triplet attention	FID↓	IS↑
After first ffn layer	7.87	9.05±0.09
After every pixel shuffle	8.6	9.10±0.07
After block 1	9.04	8.62±0.13
The last layer	7.68	9.05±0.08

5. Conclusion

We construct a hierarchical generator adversarial network based on ViT on the task of unconditional image generation. We have also introduce Triplet attention to the structure, which ensures that the generated images are structured properly. In addition, Consistency Regularization is also shown to perform equally well on the ViT-based model. Experiment results shown that the performance of the proposed method beyond the TransGAN and some other CNN-based GANs. However, our model structure is an ultra-simple structure containing only three layers of transformer, which still presents some challenges in handling higher resolution tasks. We will then further try to gradually improve the resolution of the generated images using some technical means.

References

[1] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[2] Y. LeCun, Y. Bengio, and G. Hinton "Deep learning", NATURE, vol. 521, pp. 436-444, MAY 28 2015.

[3] Zhang N, Ding S, Zhang J, et al. An overview on restricted Boltzmann machines[J]. Neurocomputing, 2018, 275: 1186-1199.

[4] Pu Y, Gan Z, Henao R, et al. Variational autoencoder for deep learning of images, labels and captions[J]. Advances in neural information processing systems, 2016, 29.

[5] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," (Banff, AB, Canada), 2014. Approximate inference; Posterior distributions; Probabilistic models; Recognition models; Reparameterization; Stochastic gradient methods; Variational bayes; Variational inference.

[7] D. J. Rezende, S. Mohamed, and D. Wierstra, " Stochastic backpropagation and approximate inference in deep generative models, " vol. 4, (Beijing, China), pp. 3057 – 3070, 2014. Approximate Bayesian inference; Approximate inference; Generative model; High dimensional data; Posterior distributions; Recognition models; Stochastic variable; Variational inference.

[8] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[9] Weng W, Zhu X. INet: convolutional networks for biomedical image segmentation[J]. Ieee Access, 2021, 9: 16591-16603.

[10] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.

[11] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]// Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.

[12] Choi Y, Choi M, Kim M, et al. Stargan: Unified generative adversarial networks for multi-domain image- to-image translation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8789-8797.

[13] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4401-4410.

[14] Zhang H, Xu T, Li H, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5907-5915.

[15] Tan F, Feng S, Ordonez V. Text2scene: Generating compositional scenes from textual descriptions [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6710-6719.

[16] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//International conference on machine learning. PMLR, 2017: 214-223.

[17] Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation [J]. arXiv preprint arXiv:1710.10196, 2017.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," (San Diego, CA, United states), 2015. Decoder architecture; Hard segments; Machine translations; New approaches; Qualitative analysis; State of the art; Statistical machine translation; Target words.

[19] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 37 (F. Bach and D. Blei, eds.), vol. 37 of Proceedings of Machine Learning Research, pp. 2048 – 2057, 2015. 32nd International Conference on Machine Learning, Lille, FRANCE, JUL 07-09, 2015.

[20] Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks[C]//International conference on machine learning. PMLR, 2019: 7354-7363.

[21] Cheng J, Dong L, Lapata M. Long short-term memory-networks for machine reading[J]. arXiv preprint arXiv:1601.06733, 2016.

[22] Parikh A P, Täckström O, Das D, et al. A decomposable attention model for natural language inference[J]. arXiv preprint arXiv:1606.01933, 2016.

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[24] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[25] Jiang Y, Chang S, Wang Z. Transgan: Two pure transformers can make one strong gan, and that can scale up[J]. Advances in Neural Information Processing Systems, 2021, 34: 14745-14758.

[26] Misra D, Nalamada T, Arasanipalai A U, et al. Rotate to attend: Convolutional triplet attention module [C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 3139- 3148.

[27] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," vol. 0, (Barcelona, Spain), pp. 1171 – 1179, 2016. Benchmark datasets; Data augmentation; Deterministic behavior; Gradient descent; Labeled datasets; Stochastic nature; Time-consuming tasks; Training sample.

[28] D. Berthelot, N. Carlini, I. Goodfellow, A. Oliver, N. Papernot, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 32 (NIPS 2019) (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche Buc, E. Fox, and R. Garnett, eds.), vol. 32 of Advances in Neural Information Processing Systems, 2019. 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, CANADA, DEC 08-14, 2019.

[29] Z. Zhao, S. Singh, H. Lee, Z. Zhang, A. Odena, and H. Zhang, "Improved consistency regularization for gans," in THIRTY-FIFTH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, THIRTY- THIRD CONFERENCE ON INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE AND THE ELEVENTH SYMPOSIUM ON EDUCATIONAL ADVANCES IN ARTIFICIAL INTELLIGENCE, vol. 35 of AAAI Conference on Artificial Intelligence, pp. 11033 – 11041, Assoc Advancement Artificial Intelligence, 2021. 35th AAAI Conference on Artificial Intelligence / 33rd Conference on Innovative Applications of Artificial Intelligence / 11th Symposium on Educational Advances in Artificial Intelligence, ELECTR NETWORK, FEB 02-09, 2021.

[30] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[31] Simonyan K Z A. Very deep convolutional networks for large scale image recognition[J]. Advances in International Conference on Learning Representations. May2015, 2014.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 37 (F. Bach and D. Blei, eds.), vol. 37 of Proceedings of Machine Learning Research, pp. 448 – 456, 2015. 32nd International Conference on Machine Learning, Lille, FRANCE, JUL 07-09, 2015.

[33] Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 568- 578.

[34] Kenton J D M W C, Toutanova L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of naacL-HLT. 2019, 1: 2.

[35] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[36] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers." Attention mechanisms; Building blockes; Grid pattern; Learn+; Longrange dependencies; Numerical experiments; Recent trends; State-of-the-art performance, 2019.

[37] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[38] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[39] Zhang H, Zhang Z, Odena A, et al. Consistency regularization for generative adversarial networks[J]. arXiv preprint arXiv:1910.12027, 2019.

[40] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[J]. Advances in neural information processing systems, 2017, 30.

[41] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training gans[J]. Advances in neural information processing systems, 2016, 29.

[42] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.

[43] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks[J]. arXiv preprint arXiv:1802.05957, 2018.

[44] Gong X, Chang S, Jiang Y, et al. Autogan: Neural architecture search for generative adversarial networks [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3224-3234.

[45] Gao C, Chen Y, Liu S, et al. Adversarialnas: Adversarial neural architecture search for gans [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5680-5689.

[46] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8110-8119.