Direct Reading StyleSwin

Chen Lei
G2202273D
MSAI in Nanyang Technological University
chen15552@e.ntu.edu.sg

Abstract

Image generation now is one of the most famous computer vision tasks. Generative Adversarial Network(GAN), among the generative models, provides pleasing high-resolution images. Most current GANs are Conv-based, which adopting convolutional backbones, is proved to need a architecture with large enhanced capacity to generate high-quality images. At the same time, transformer has shown its potential in vision tasks such as image classification.
StyleSwin is a transformer-based GAN for high-resoltuion images generation, which is proposed by B Zhang and his team. It uses Swin transformer in a style-based architecture and local attention to gain a high computational efficiency. StyleSwin is proved to outperform the Conv-based GAN on 256*256 datasets and provide a close performance to ConvNets on 1024*1024 datasets.

Introduction

background

In the field of image generation tasks, Generative Adversarial Network(GAN) has shown a great ability to output high quality images and its researching focus, gradually, transfers from stabilizing the training procedure by using proper regulations or loss functions(Gu et al. 2020)(Gulrajani et al. 2017), as GAN model is sensitive to its model architecture and hyper parameters, to creating models with large capacity.
With the great achievement of transformer in NLP tasks, some research on uses of transformer in computer vision tasks. Though there has been a few works building generative models with transformer(Jiang, Chang, and Wang 2021)(Lee et al. 2021) in order to generate complex images, it is still not competitive to the ConvNets in generating images with high-resolution. One of the challenges is the high quadratic computational cost preventing the network generating high-resolution images, which means it needs to reduce the point-wise multi-layer perceptrons(MLP) when facing large scale. Also, for window-based GAN, the generator doesn't know the position for patches when generating images so that it can't leverage absolute positions for image synthesis. Besides, the blocking artifact when the generating images with high-resolution because of the transformer is also a great challenge in prior works.

Directed Reading Paper

StyleSwin(Zhang et al. 2022) is a transformer-based GAN proposed by Bowen Zhang and his team, in order to build a GAN model with pure transformer for high-resolution image synthesis and look for the key issues during the generating tasks. StyleSwin adopts Swin transformer in a style-based architecture for the sake of a balanced computational efficiency and modeling capacity, and using local attention to gain most modeling capacity and high computational efficiency as double attention can compensate the receptive field. Also, they introduce sinusoidal positional encoding(SPE) to help the generator understand the position for patches. At last, to prevent blocking artifact, it examines the spectral discrepancy by a wavelet discriminator.
StyleSwin has shown a great ability to generate high-resolution images according to Zhang's team. It has been tested on CelebA-HQ 1024 and FFHQ-102 bencmarks. For the former one, StyleSwin reaches an Fréchet inception distance(FID) score of 4.43 and it gets 5.07 FID score on the latter benchmark, representing StyleSwin's competitive performance on high-resolution image generation tasks.

Style Swin Model

The architecture of StyleSwin model is as shown in figure1:

Base Generator Model

The StyleSwin is built based on a simple generative model, which takes a variable $z \sim N(0, I)$ as input and go through several transformer blocks to upsample the feature maps. The Swin transformer(Liu et al. 2021) is used to act as the basic building block using shifted window partition to compute multi-head self-attention locally. Suppose in the layer l, there comes a input feature map $x^l R^{H \times W \times C}$, the Swin blocks act as:

$$RegularWindow \begin{cases} \hat{x}^l = W - MSA(LN(x^l)) + x^l, \\ x^{l+1} = MLP(LN(\hat{x}^l)) + \hat{x}^l, \end{cases} \quad (1)$$

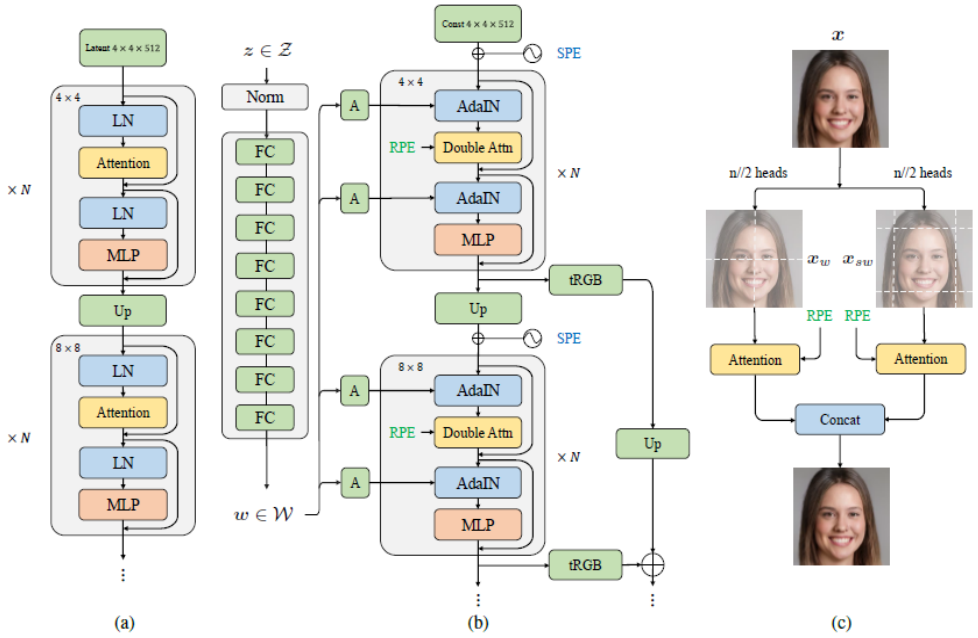

Figure 1: The architecture of StyleSwin model(Zhang et al. 2022), (a) is the baseline generator model, (b) is the StyleSwin model using styled architecture, (c) is the proposed double attention

$$ShiftedWindow \begin{cases} \hat{x}^{l+1} = SW - MSA(LN(x^{l+1})) + x^{l+1}, \\ x^{l+2} = MLP(LN(\hat{x}^{l+1})) + \hat{x}^{l+1}, \end{cases} \tag{2}$$

where W-MSA and SW-MSA represent the regular window and shifted window partition and LN is the layer normalization. However, accroding to the team, the discriminator impairs the stability of training, so StyleSwin uses several tricks to improve the model be more competitive.

Style Injection

As shown in figure1(b), StyleSwin uses a style-based architecture to enhance model's capacity, which needs a nonlinear function $f : Z \rightarrow W$ mapping the input features from space Z to W to help inject style into network. Several injection methods have been tried by the team as show in table1:

According to the result, all injection methods improve the modeling capacity except AdaBN makes training not convergent might because the batch size need to be small when synthesizing high-resolution images. Also, although AdaRMSNorm has a sufficient style injection by using style information in both attention block and FFN, Zhang's team decides not to use it for the sake of efficiency. So based on the FID score, AdaIn is chosen as it can be modulated independently as well as supply better feature modulation with normalized feature maps.

Style Injection methods	FID
Baseline	15.03
AdaIN	6.34
AdaLn	6.95
AdaBN	¿100
AdaRMSNorm	7.43
Modulated MLP	7.09
Cross attention	6.59

Table 1: Comparison of different style injection methods on FFHQ-256(Zhang et al. 2022)

Double Attention

StyleSwin introduces the double attention to gain a larger receptive field, which enables single transformer block cares about shifted windows and local context at the same time. As shown in figure1(c), the attention heads are split into two groups computing regular window attention and shifted window attention separately. Then two parts' results are concatenated as the output.

Suppose the patches under regular window as x_w and patches under shifted window as x_{sw}, x_w and x_{sw} are not overlapping and both of them are in $R^{\frac{HW}{K^2} \times K \times K \times C}$, the double attention can be formulated as:

$$DoubleAttention = Concatenate(head_1, ..., head_n) \cdot W^O \tag{3}$$

W^O is the projection matrix mixing heads to output. For each attention head $head_i$, it can be computed by:

$$head_i = \begin{cases} Attention(x_w W_i^Q, x_w W_i^K, x_w W_i^V) & i \leq \lfloor \frac{h}{2} \rfloor \\ Attention(x_{sw} W_i^Q, x_{sw} W_i^K, x_{sw} W_i^V), & i > \lfloor \frac{h}{2} \rfloor \end{cases} \tag{4}$$

where W_i^Q is the query projection matrix, W_i^K is the key projection matrix and W_i^V is the value projection matrix of

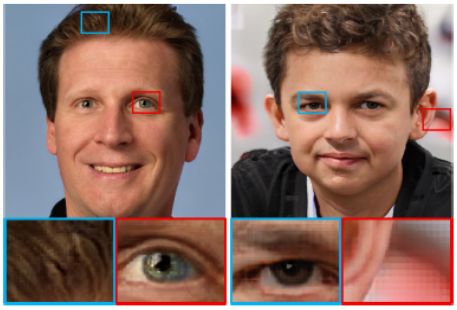

Figure 2: Obvious blocking artifacts on 1024×1024 images(Zhang et al. 2022)

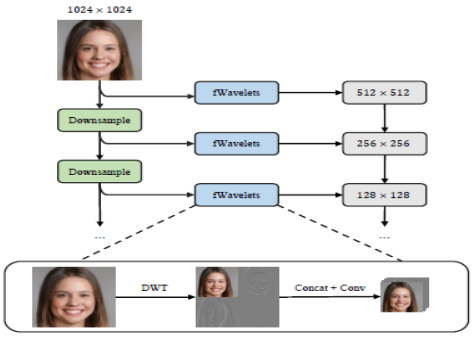

Figure 3: The wavelet discriminator's work flow(Zhang et al. 2022)

$head_i$. With a single double attention, each dimension has more $2.5k$ receptive field, which makes capturing larger context more efficiently.

Local-global positional encoding

As mentioned in Chapter Background, sometimes the generator doesn't know the absolute positions. It is because Swin blocks use default RPE to encode the relative position of pixels and it can't use zero paddings to gain the absolute position like Conv-based GANS does(Islam, Jia, and Bruce 2020)(Kayhan and Gemert 2020).
StyleSwin model introduces a encoding method called sinusoidal position encoding(SPE)(Xu et al. 2021) as in Figure1(b). After scale upsampling the feature maps are encoded with:

$$\underbrace{sin(w_0 i), cos(w_0 i), ...}_{\text{horizontal dimension}} \underbrace{sin(w_0 j), cos(w_0 j), ...}_{\text{vertical dimension}} \in R^C \quad (5)$$

where $w_k = \frac{1}{10000^{2k}}$ and (i, j) refers to the location. StyleSwin uses RPE and SPE together where RPE within each transformer block supply relative positions and SPE is used to gain absolute positions.

Solution of Blocking Artifact

After using StyleSwin to generate 256×256 images, Zhang's team tries to apply it directly to synthesize 1024×1024 images. They find that blocking artifacts are obvious on these images as figure2 shows. Zhang's team guesses it's the transformer which causes blocking artifacts and they confirm their guess after obtaining an 64×64 image without artifacts using the model employing only MLPs to characterize the high-frequency results. This phenomenon may be because Swin transformer break the spatial coherency as it computes attention in non-overlapping local windows.
Zhang's team notices that the blocking artifacts disappear gradually during training, which proves that the current generator is able to be artifact-free. They think that the discriminator can't examine the details with a high-frequency, so they have tried three different discriminators: Patch discriminator(Isola et al. 2017), Total variation annealing and Wavelet discriminator(Gal et al. 2021). They compare the FID score and whether the artifact suppression method removes the blocking artifacts on FFHQ-1024 dataset, and result is shown as table2: The result shows that Patch discriminator fails to remove the blocking artifacts while MLPs isn't able to generator high-resolution images as well as total variation annealing, as both of them get a high FID score. By comparison, Wavelet discriminator outperform than other methods and achieve a pleasing result, which is finally chosen by Zhang's team.

Artifact Suppression methods	FID	Remove artifacts or not
Window-based attention	8.39	No
MLPs after 64×64	12.69	Yes
Patch discriminator	7.73	No
Total variation annealing	12.79	Yes
Wavelet discriminator	5.07	Yes

Table 2: Comparison of different artifacts suppression methods on FFHQ-1024 dataset(Zhang et al. 2022)

StyleSwin uses wavelet discriminator to reduce the size of input image and assesses the frequency discrepancy compared to real images on each scale by discrete wavelet decomposition, as Figure3 shows. The advantage of wavelet discriminator is leading the generator to gain more details without side-effects on distribution matching.

Experiment and Discussion

Experiment

The performance of StyleSwin is tested on the following datasets:CelebA-HQ(Karras et al. 2017), LSUN Church(Yu et al. 2015) and FFHQ(Karras, Laine, and Aila 2019). Zhang's team tries to synthesize 256×256 and 1024×1024 images on CelebA-HQ and FFHQ but only 256×256 on LSUN Church.
StyleSwin' ability is evaluated by Fréchet inception distance(FID) score, which measures the distribution discrepancy. Usually a lower FID score refers to better image quality. For FFHQ and LSUN Church datasets, FID is calculated between 50,000 output images and images in validation sets

randomly sampled from original dataset while for CelebA-HQ, Zhang compares 30,000 output images and all images in the sample. Adam sovler is used as the optimizer with $\beta_1 = 0.0$ and $\beta_2 = 0.99$ and balanced consistency regularization(bCR) is used on FFHQ-256 and CelebA-HQ 256 with $\lambda_{real} = \lambda_{fake} = 10$. The result is shown in table3 and table4: Also, in order to test the effect of each introduced components, Zhang's team conducts ablation studies and the result is shown in table5:

Models	FFHQ	CelebA-HQ	LSUN Church
StyleGAN2	3.62*		3.86
PG-GAN		8.03	6.42
U-Net GAN	7.63		
INR-GAN	9.557		5.09
MSG-GAN			5.20
CIPS	4.38		2.92
TransGAN		9.60*	8.94
VQGAN	11.40	10.70	
HiT-B	2.95*	3.39*	
StyleSwin	2.81*3.25*	2.95	

Table 3: Comparison of generation models on FFHQ,CelebA-HQ and LUSN Church of 256×256 resolution. * refers that bCR is used in training (Zhang et al. 2022)

Models	FFHQ	CelebA-HQ
StyleGAN	4.41	5.06
COCO-GAN		9.49
PG-GAN		7.30
MSG-GAN	5.80	6.37
INR-GAN	16.32	
CIPS	10.07	
HiT-B	6.37	8.83
StyleSwin	5.07	4.43

Table 4: Comparison of generation models on FFHQ and CelebA-HQof 1024×1024 resolution.(Zhang et al. 2022)

Configuration	FID
Swin baseline	15.03
+Style Injection	8.40
+Double Attention	7.86
+Wavelet discriminator	6.34
+SPE	5.76
+Larger model	5.50
+bCR	2.81

Table 5: Comparison of generation models on FFHQ and CelebA-HQof 1024×1024 resolution.(Zhang et al. 2022)

(a)

Figure 4: The generated images by StyleSwin on (a)FFHQ 1024×1024 and (b) CelebA-HQ 1024×1024(Zhang et al. 2022)

Discussion

From the above experiments' results, StyleSwin outperform obviously than other current Conv-based GANs and transformer-based GANs on synthesizing 256×256 resolution images on both FFHQ and LSUN Church datasets. On the other hand, StyleSwin is proved to be able to generate high-resolution images as table4 shows, especially the comparison between StyleSwin and HiT-B, proving that self-attention is beneficial.

According to the Figure4, StyleSwin is able to generator complex images with high resolution. Some details like mouths and eyes proves the advantages of using SPE. Also as shown in table5, style injection brings more model capacity, leading to a better performance. In all, StyleSwin shows a pleasing performance in generating high-resolution images.

However, considering combining transformer in generative modeling is a quite new study field, there must be some improvements. For example, StyleSwin uses an artifact suppression discriminator to solve the blocking effects brought by shifted windows in transformer, so some changes in transformer's architecture is possible like cyclic shifting strategy. Though there is few improvements about StyleSwin, the future work still needs further study.

Conclusion

This paper review study a high-resolution image generative model StyleSwin and introduce the components in StyleSwin like double attention, style injection and artifact

suppression. StyleSwin shows its ability to synthesize complext images with high-resolution and details. It can be an inspiration of more generative models based on transformer.

References

Gal, R.; Hochberg, D. C.; Bermano, A.; and Cohen-Or, D. 2021. Swagan: A style-based wavelet-driven generative model. *ACM Transactions on Graphics (TOG)* 40(4):1–11.

Gu, S.; Bao, J.; Chen, D.; and Wen, F. 2020. Giqa: Generated image quality assessment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 369–385. Springer.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems* 30.

Islam, M. A.; Jia, S.; and Bruce, N. D. 2020. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jiang, Y.; Chang, S.; and Wang, Z. 2021. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074* 1(3).

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kayhan, O. S., and Gemert, J. C. v. 2020. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14274–14285.

Lee, K.; Chang, H.; Jiang, L.; Zhang, H.; Tu, Z.; and Liu, C. 2021. Vitgan: Training gans with vision transformers. *arXiv preprint arXiv:2107.04589*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Xu, R.; Wang, X.; Chen, K.; Zhou, B.; and Loy, C. C. 2021. Positional encoding as spatial inductive bias in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13569–13578.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zhang, B.; Gu, S.; Zhang, B.; Bao, J.; Chen, D.; Wen, F.; Wang, Y.; and Guo, B. 2022. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11304–11314.