TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up

(NeurIPS 2021)

Yifan Jiang, Shiyu Chang, Zhangyang Wang (2021)

Introduction

TransGAN

- Generative adversarial networks(GANs) architecture which is transformer based and use convolutions at all.
- Generator is memory friendly and discriminator is multi scale.

Contributions :

- Introduced new module of grid self-attention.
- Training recipe : Data augmentation, modified normalization, and relative position encoding.
- Competitive performance with s-o-t-a GANs using convolutional backbones.

Architecture

Memory-friendly Generator

Multi-scale Discriminator

Grid Self-Attention

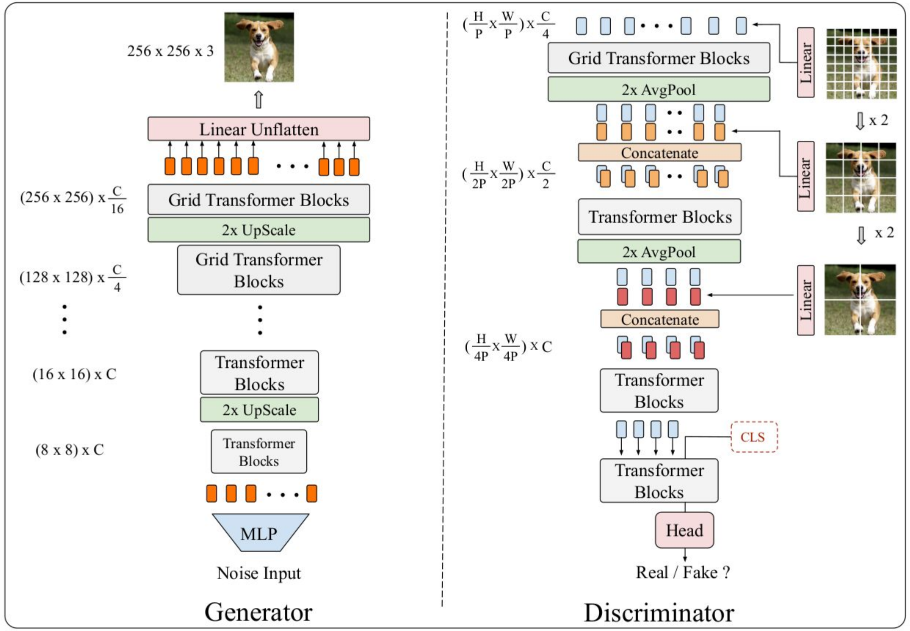
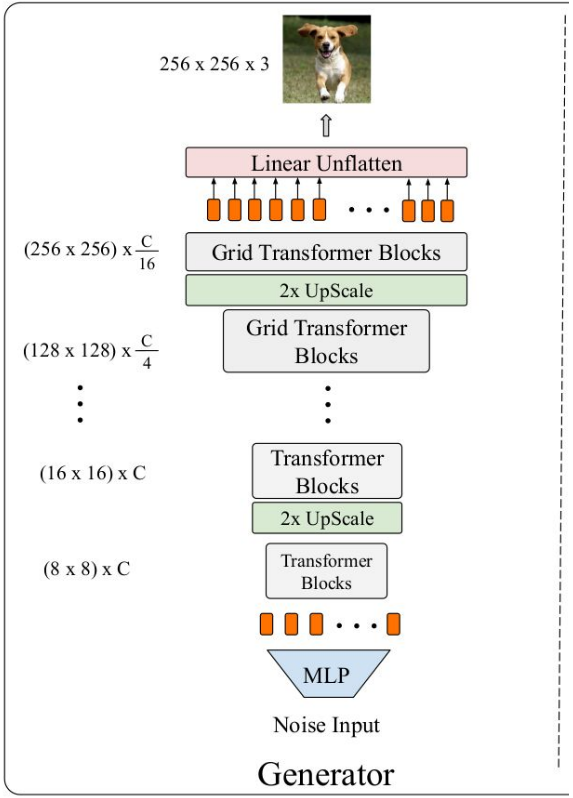


Figure 2: The pipeline of the pure transform-based generator and discriminator of TransGAN. We take 256×256 resolution image generation task as a typical example to illustrate the main procedure. Here patch size p is set to 32 as an example for the convenience of illustration, while practically the patch size is normally set to be no more than 8×8, depending on the specific dataset. `Grid Transformer Blocks` refers to the transformer blocks with the proposed grid self-attention. Detailed architecture configurations are included in Appendix B

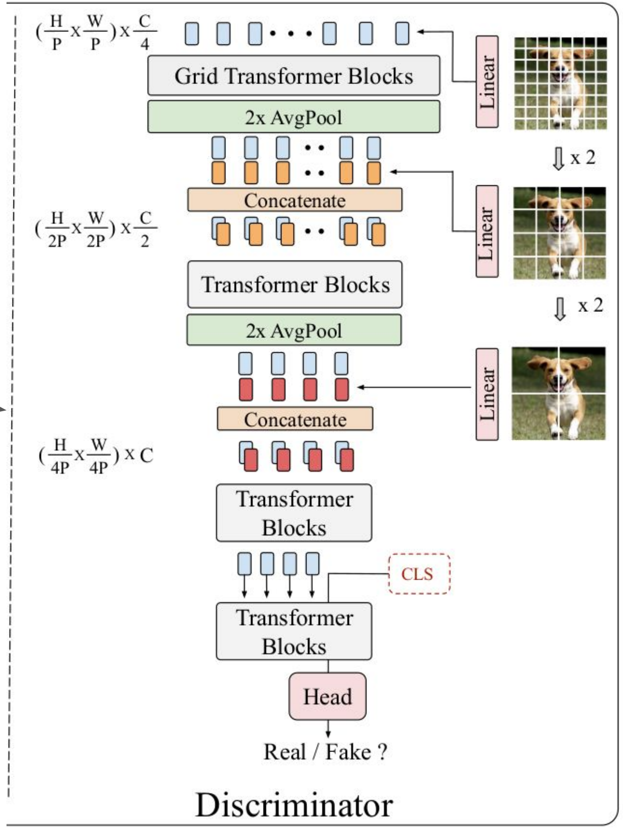
Architecture : Generator

Memory-friendly Generator



Architecture : Discriminator

Multi-Scale Discriminator



Architecture : Grid Self-Attention

Grid Self-Attention

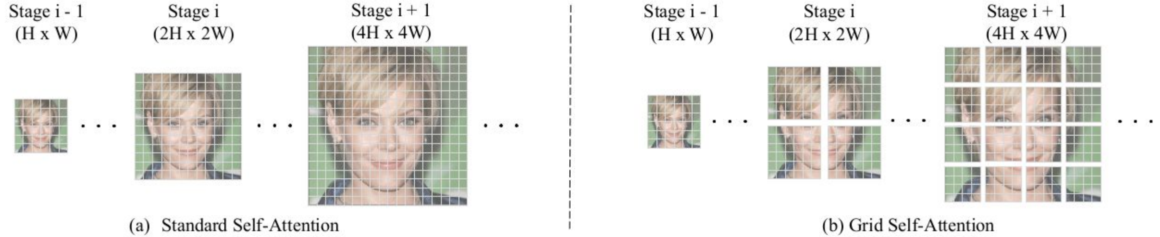


Figure 3: Grid Self-Attention across different transformer stages. We replace Standard Self-Attention with Grid Self-Attention when the resolution is higher than 32×32 and the grid size is set to be 16×16 by default.

Training Recipe

Data Augmentation

Differential augmentation with three basic operators {Translation, Cutout, Color} leads to surprising performance improvement for TransGAN, while CNN-based GANs hardly benefit from it.

Relative Position Encoding

$$Attention(Q, K, V) = softmax(((\frac{QK^T}{\sqrt{d_k}} + E)V)$$

Modified Normalization

$$Y = X/\sqrt{\frac{1}{C}\sum_{i=0}^{C-1}(X^i)^2 + \epsilon}, \text{ where } \epsilon = 1e-8$$

Results

Table 1: Unconditional image generation results on CIFAR-10, STl-10, and CelebA (128×128) dataset. We train the models with their official code if the results are unavailable, denoted as "*", others are all reported from references.

Methods	CIFAR-10		STL-10		CelebA
	IS↑	FID↓	IS↑	FID↓	FID↓
WGAN-GP [1]	6.49 ± 0.09	39.68	-	-	-
SN-GAN [48]	8.22 ± 0.05	-	9.16 ± 0.12	40.1	-
AutoGAN [18]	8.55 ± 0.10	12.42	9.16 ± 0.12	31.01	-
AdversarialNAS-GAN [18]	8.74 ± 0.07	10.87	9.63 ± 0.19	26.98	-
Progressive-GAN [16]	8.80 ± 0.05	15.52	-	-	7.30
COCO-GAN [74]	-	-	-	-	5.74
StyleGAN-V2 [69]	9.18	11.07	10.21* ± 0.14	20.84*	5.59*
StyleGAN-V2 + DiffAug. [69]	**9.40**	9.89	10.31*± 0.12	19.15*	5.40*
TransGAN	9.02 ± 0.12	**9.26**	**10.43** ± 0.16	**18.28**	**5.28**

other "modern" normalization layers [76–78] that need affine parameters for both mean and variances, we find that a simple re-scaling without learnable parameters suffices to stabilize TransGAN training – in fact, it makes TransGAN train better and improves the FID on some common benchmarks, such as CelebeA and LSUN-Church.

Table 3: The ablation study of proposed techniques in three common dataset CelebA(64×64), CelebA(128×128, and LSUN Church(256×256)). "OOM" represents out-of-momery issue.

Training Configuration	CelebA (64x64)	CelebA (128x128)	LSUN Church (256x256)
(A). Standard Self-Attention	8.92	**OOM**	**OOM**
(B). Nyström Self-Attention [64]	13.47	17.42	39.92
(C). Axis Self-Attention [67]	12.39	13.95	29.30
(D). Grid Self-Attention	9.89	10.58	20.39
+ Multi-scale Discriminator	9.28	8.03	15.29
+ Modified Normalization	7.05	7.13	13.27
+ Relative Position Encoding	6.14	6.32	11.93
(E). Converge	**5.01**	**5.28**	**8.94**

larger than CIFAR-10, suggesting that transformer-based architectures benefit much more notably from larger-scale data than CNNs.

Table 2: The effectiveness of Data Augmentation on both CNN-based GANs and TransGAN. We use the full CIFAR-10 training set and DiffAug [69].

Methods	WGAN-GP		AutoGAN		StyleGAN-V2		TransGAN	
	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓
Original	**6.49**	39.68	8.55	**12.42**	9.18	11.07	8.36	22.53
+ DiffAug [69]	6.29	**37.14**	**8.60**	12.72	**9.40**	**9.89**	**9.02**	**9.26**

Examples

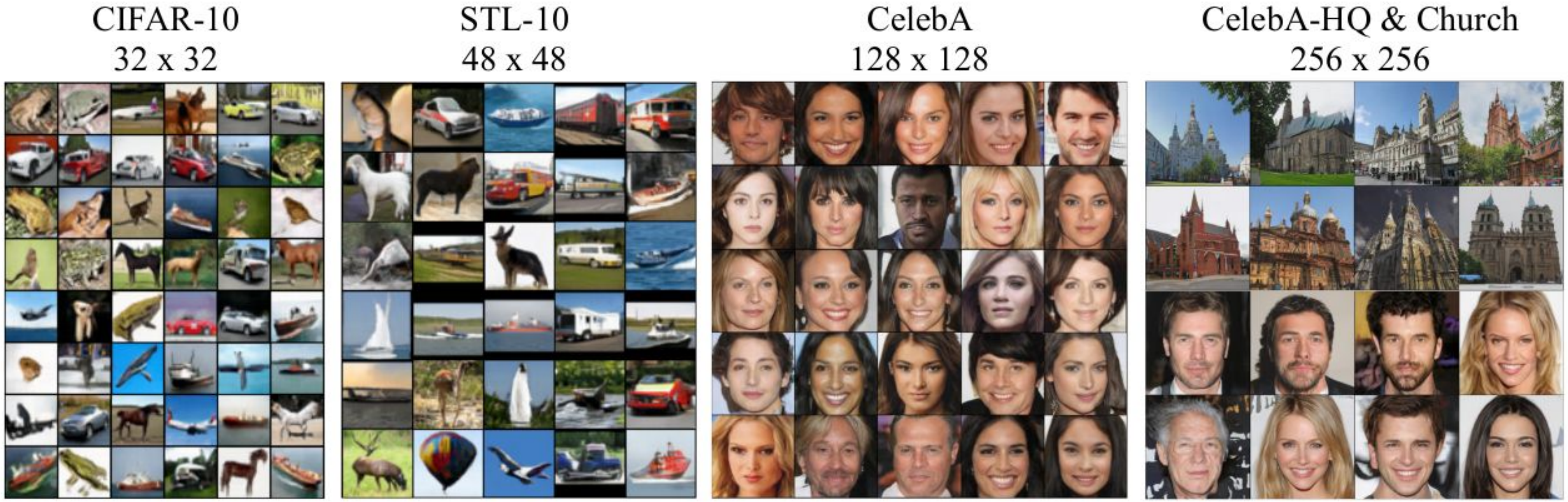


Figure 4: Representative visual results produced by TransGAN on different datasets, as resolution grows from 32×32 to 256×256. More visual examples are included in Appendix F.

Thank You