TransT2I: Transformer-based GAN for Text-to-Image Synthesis

Anonymous ICCV submission

Paper ID xxxx

Abstract

Although transformers have shown impressive performance in most downstream vision tasks, they haven't yet fully demonstrated their expressivity in cross-modal text-to-image synthesis. In this paper, we seek to explore employing transformers to build a generative adversarial network for text-to-image synthesis. However, by careful study of transformers and related works, there are three obstacles. First, the standard multi-head attention mechanism is usually insufficient to capture high-frequency signals (local details) and brings huge training burden due to its quadratic computational complexity, which hinders its application to broader tasks severely. Second, the text-image fusion modules adopted by previous works almost don't achieve high-efficient cross-modal text-image fusion, which limits model capacity heavily. Third, the semantic alignment tool DAMSM ignores the representation learning of text features, which blocks the generator from further generating text-matching images. To address these, we propose TransT2I, a transformer-based GAN for text-to-image synthesis. In particular, we propose (i) Mix Attention, which can simultaneously capture global relationships and local details while enjoying linear computational complexity; (ii) Conditioned Fusion Instance Normalization (ConIN), which can achieve effective cross-modal text-image fusion with slight computational cost; (iii) Deep Text-Image Contrastive Model (DTMCM), which can better reflect the learning of text features. Extensive experiments on three challenging benchmarks demonstrate the state-of-the-art performance of TransT2I over prior text-to-image works, proving the promise of transformer-based GAN for text-to-image synthesis. Additionally, more experiments and analyses are conducted in the Supplementary Material.

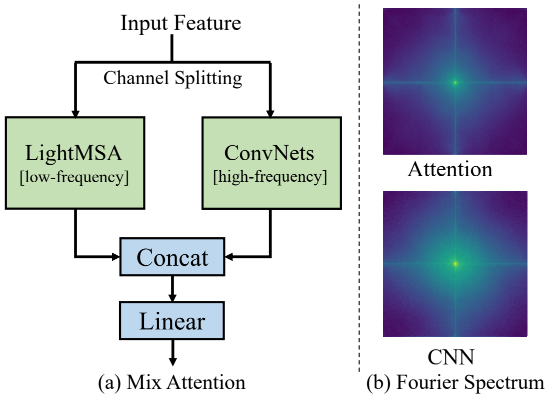
1. Introduction

In recent years, synthesizing images from natural language descriptions has gained widespread attention due to its potential value in many fields, such as computer-aided design, virtual scene generation, photo editing. To reach this, many methods have been proposed, including: Generative Adversarial Networks [11], Diffusion Models [12], Variational Auto-Encoders [21], *etc.* However, it still has a big gap between current results and desired performance.

Inspired by the outstanding performance in NLP [47, 2, 7], vision transformers have been proposed [10, 46] and achieved superior results in numerous vision tasks [26, 49, 9, 16]. This is largely attributed to its strong capability of modeling long-range dependencies in the data with self-attention mechanism. Its success has motivated researchers to apply it to broader tasks. Recently, some transformer-based unconditional GANs have made progress [16, 57, 61, 29, 22], but it‘s still waiting to be explored by transformer-based GAN for text-to-image synthesis.

In this paper, we aim to explore the key ingredients when employing transformers to build a powerful GAN for cross-modal text-to-image synthesis. To achieve this, we found there are three obstacles. First, the standard multi-head attention mechanism is usually insufficient to capture high-frequency signals and brings huge train burden due to its quadratic computational complexity. Vision transformers are beneficial to capture low-frequency signals [30, 41, 28], which indicates global structures and shapes. While it is usually insufficient to capture some high-frequency signals, such as local structures, edges and lines. Furthermore, many studies have pointed out [56, 57, 53] that vision transformers perform poorly when trained from scratch, due to the lack of local inductive bias. Besides, the standard multi-head attention mechanism brings huge train burden. For instance, a feature map of size 56×56×96 costs 2.0G FLOPs in one Multi-head Self Attention (MSA) [47], while the entire model of ResNet-18 [13] only requires 1.8G FLOPs. This problem severely hinders the application of vision transformers to broader tasks.

Second, the text-image fusion modules adopted by previous works almost don't achieve high-efficient cross-modal text-image fusion. Although the previously adopted cross-modal attention [52] extracts long-range relationships and achieves spatial fusion well, it brings a heavy training burden due to its high computational loss. Besides, since image and natural language descriptions belong to different



(a) Mix Attention (b) Fourier Spectrum

Figure 1. (a) The architecture of Mix Attention, which can simultaneously capture global relationships (low-frequency) and local details (high-frequency) while enjoying linear computational complexity. (b) The Fourier Spectrum of CNN [13] and Attention [47]. The lighter the color, the larger the magnitude. A pixel that is closer to the center means a lower frequency. More details are in Section 3.2.

semantic levels [3, 55, 24], we argue that attention-based text-image fusion is not effective enough. Recently, many works achieve text-image fusion by affine transformation [43, 24, 54], which maps text conditions to each channel separately. Although they achieve competitive results, they lack information fusion across channel dimensions after affine transformation. Third, the semantic alignment tool DAMSM ignores the representation learning of text features, which blocks the generator from further generating text-matching images. The DAMSM [52] uses bidirectional LSTM [15] to extract text features. Although LSTM captures context relationships well, it ignores the attention between different words in a sentence, which cause some irrelevant words to impose an impact on model. Besides, the DAMSM supervises the generator with word loss and sentence loss together. We argue that the word loss can't provide positive feedback to the generator, due to the different semantic levels between image and words [3].

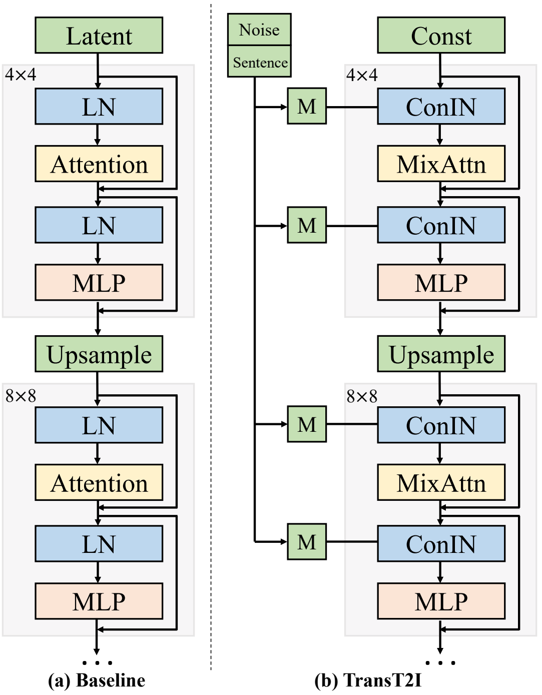
To address them, we propose TransT2I, a transformer-based GAN for text-to-image synthesis. For the first obstacle, we propose a novel attention method (Mix Attention), which can simultaneously capture global relationships (low-frequency) and local details (high-frequency) while enjoying linear computational complexity. In Mix Attention, we disentangle high/low frequency signals by dividing the input feature into two groups. Furthermore, we adopt the lightweight attention module to strike a tradeoff between model capacity and computational efficiency. For the second obstacle, we propose **Con**ditioned Fusion **I**nstance **N**ormalization (ConIN), which can achieve effective cross-modal text-image fusion with slight computational cost. We accept the text condition through affine transformation and add an additional fully connected Layers to facilitate the channel fusion of cross-modal information. In addition, before and after the module, we stack the instance normalization method to improve the stability of training. For the third obstacle, we propose **D**eep **t**ext-**I**mage **C**ontrastive **M**odel (DTMCM), which can reflect the representation learning of text features well. For text encoders, before the original LSTM, we add several transformer encoders to show attention between different words in a sentence to achieve learning of text features. To better guide the generator, we remove the word loss and use the sentence loss to provide feedback to generator. More importantly, our proposed DTMCM can be utilized as a general-purpose cross-modal alignment tool like DAMSM for text-to-image synthesis.

Extensive experiments on three challenging benchmarks demonstrate the superiority over previous text-to-image works, proving the promise of transformer-based GAN for text-to-image synthesis. On CUB and Multi-modal CelebA-HQ, TransT2I achieves a state-of-the-art FID 10.06 and 11.87 that exceeds all previous works. On COCO, TransT2I achieves a competitive FID 11.94 only with 42% trainable parameters compared with current SOTA model Lafite [62].

2. Related Work

Text-to-Image Synthesis. In 2016, Reed *et al.* first proposed using conditional generative adversarial networks to generate images under text conditions [35]. To further generate the desired images, Zhang *et al.* proposed stacking multiple generator-discriminator pairs to gradually generate high-quality images from coarse to fine under text conditions [59]. During training, multiple generator-discriminator pairs are required to coordinate to generate higher quality images. Then, Xu *et al.* followed the architecture and proposed AttnGAN [52] to achieve word-level fine-grained generation by introducing a word-level attention mechanism. For a period of time, the stacked architecture has become the basic method for text-to-image synthesis [5, 37, 60, 23, 31]. Due to the limitations of stacked architecture, Ming *et al.* proposed DF-GAN [44], which aims to utilize single generator to achieve text-to-image synthesis. Recently, some works [34, 8] attempt to train auto-regressive models based on transformers to achieve zero-shot text-to-image synthesis, but the expensive training costs hinder most researchers. Besides, the methods of diffusion model [12, 39] have shown wide potential value.

Transformer-based GAN. Jiang *et al.* [16] first proposed to employ pure transformer to build GAN, in which grid attention was utilized to lower computational loss. Later, more advanced performance was achieved based on vision transformer [22, 29, 61, 57]. Then, the StyleSwin [57] combine StyleGAN [19] and Swin transformer [26] to construct transformer-based GAN, which showed more competitive



(a) Baseline **(b) TransT2I**

Figure 2. The architectures we investigate. (a) The baseline architecture is comprised of a series of transformer blocks hierarchically. (b) Our proposed TransT2I builds a novel transformer-based GAN for text-to-image synthesis, which consists of 7 TransT2I Blocks from 4^2 to 256^2. The "Sentence" stands for sentence vector encoded by pretained text encoder of proposed DTMCM. The "Noise" stands for random noise sampled from Gaussian distribution. The "M" stands for MLP. More details are in Section 3.

performance. But these works almost are only associated with unconditional image synthesis. Unlike previous works, we propose TransT2I to explore transformer-based GAN for text-to-image synthesis.

3. Method

In this paper, we explore building a transformer-based GAN for text-to-image synthesis. We propose TransT2I, which contains three innovative components: Mix Attention, Conditioned Fusion Instance Normalization (ConIN) and Deep Text-Image Contrastive Model (DTMCM).

3.1. Model Overview

The architecture of generator is shown in Figure 2. We propose TransT2I, a transformer-based text-to-image generative adversarial network. First, inspired by StyleGAN family [19, 17], we replace random noise with learnable constants as the input to first layer, and accept cross-modal text conditions in the way of style injection [19]. Next, we propose an innovative basic computational block (called TransT2I Block). Different from the standard vision transformer block [10, 47], our proposed TransT2I Block can achieve the trade-off between computational efficiency and model capability. In TransT2I Block, we propose Mix Attention, which can simultaneously capture global relationships and local details while enjoying linear computational complexity; besides, we propose Conditioned Fusion Instance Normalization, which can achieve effective cross-modal text-image fusion while helping to stabilize the training process.

In generator, we first generate a random noise sampled from Gaussian distribution. Then, text descriptions are encoded into sentence vectors by our proposed pretrained text encoder. We concatenate random noise with sentence vector to get the input to model. From 4^2 to 256^2 , the generator consists of 7 TransT2I Blocks. The mathematical form of TransT2I Block is shown below:

$$\begin{aligned} z\hat{\mathbf{h}}^i &= \mathrm{MixAttn}(\mathrm{ConIN}(\mathbf{h}^{i-1}, \mathbf{s})) + \mathbf{h}^{i-1}, \\ \mathbf{h}^i &= \mathrm{MLP}(\mathrm{ConIN}(\hat{\mathbf{h}}^i, \mathbf{s})) + \hat{\mathbf{h}}^i, \end{aligned} \tag{1}$$

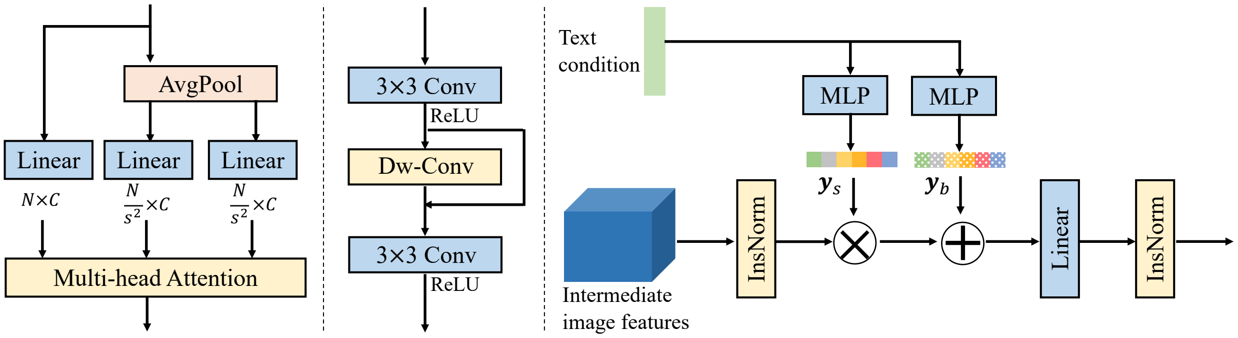
where $\mathbf{h}^i$ is the output feature map of TransT2I Block i, $\mathbf{s}$ is text condition, ConIN is Conditioned Fusion Instance Normalization, MixAttn is Mix attention and MLP is Multilayer Perceptron [45], respectively. We will introduce them next.

3.2. Mix Attention

The architecture of Mix Attention is shown in Figure 1(a). The proposal of Mix Attention is motivated by two aspects. First, many previous studies have shown that attention-based mechanisms can capture low-frequency signals (global relationships) well, while it's incompetent to capture high-frequency signals (local details) [41, 28]. Second, transformers are data-hungry for vision tasks due to the lack of local inductive bias like CNNs [56, 57]. However, some public datasets for text-to-image synthesis aren't large enough for training transformers from scratch, such as: CUB [48], Multi-Modal CelebA-HQ [51]. To address these, we propose Mix Attention.

LightMSA. The architecture of LightMSA is shown in Figure 3(a). In order to achieve a trade-off between computational efficiency and model capacity, we introduce LightMSA in Mix Attention. The standard multi-head attention mechanism brings huge training burden due to its quadratic computational complexity. Unlike the standard multi-head attention and SRA [49], we utilize the average pooling operation on the transformed $\mathbf{K}$ and $\mathbf{V}$ matrix to reduce the computational loss, thus achieving linear complexity. The $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ matrix is then processed by a fully connected layer. Finally, the obtained results are sent to calculate the multi-head dot-product attention.

ConvNets. The architecture of ConvNets is shown in Figure 3(b). To better capture high-frequency signals and introduce local inductive bias into our model, we introduce



(a) LightMSA **(b) ConvNets** **(c) Conditioned Fusion Instance Normalization**

Figure 3. (a) LightMSA: The low-frequency module of proposed Mix Attention, in which we employ avgpool operation to lower computational loss. (b) ConvNets: The high-frequency module of proposed Mix Attention, in which we stack several convolutional networks to introduce local inductive bias. (c) Conditioned Fusion Instance Normalization (ConIN): The module aims to achieve effective cross-modal text-image fusion, in which " InsNorm" stands for instance normalization.

ConvNets in proposed Mix Attention. ConvNet consists of 3×3 convolution networks and depth-wise convolution networks. The purpose of depth-wise convolution networks is to enhance the feature modeling across spatial dimensions.

The Parallel Design. In order to better capture the global relationships and local details, we need to utilize a proper way to combine these two different modules. Some previous works [26, 49] tend to combine the two different attention modules in a successive way. However, we argue that the successive way will make the interaction between the two modules even less interweaved. So, we adopt the parallel design [4, 57] to combine LightMSA and ConvNets.

First, we divide the input feature map into two groups by channel cutting. Each group corresponds to a module. Then, the output feature maps of these two modules are concatenated together. The concatenated results are fed into the following fully connected network to fuse the learnable relations across the channel dimensions. The channel division ratio is set to 0.5 by default.

Therefore, our designed Mix Attention can simultaneously capture low-frequency signals (global relationships) and high-frequency signals (local details) with linear computational complexity. As shown in Table 2, compared with other attention mechanisms, our proposed TransT2I obtains excellent results.

3.3. Conditioned Fusion Instance Normalization

It's our belief that a powerful cross-modal text-image fusion module is not only beneficial to generate higher-quality images, but also facilitates the stabilization of training process. Drawing by this, we propose the **Con**ditioned Fusion **I**nstance **N**ormalization module (ConIN). The architecture of ConIN is shown in Figure 3(c).

First, inspired by DCGAN [33] and StyleGAN family [19, 17, 18], in order to stabilize training process, we use instance normalization to process the input feature map $\mathbf{F} \in \mathbb{R}^{B\times C\times H\times W}$, the mathematical form is as follows:

$$\text{InsNorm}(\mathbf{f}_i) = \frac{\mathbf{f}_i - \mu(\mathbf{f}_i)}{\sigma(\mathbf{f}_i)}, \tag{2}$$

where $\mathbf{f}_i$ is the i-th channel of input feature map, $\mu(\mathbf{f}_i)$ is the mean of $\mathbf{f}_i$, $\sigma(\mathbf{f}_i)$ is the variance of $\mathbf{f}^i$, each channel is separately normalized, respectively. Then, in order to achieve effective cross-modal text-image fusion, we stack the affine transformation [6, 27] to accept text conditions. For the affine operation, we stack two MLPs [45] to predict channel-wise scaling parameters and bias parameters for a given input feature map $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$. The mathematical form is shown in the following:

$$\begin{aligned} \mathbf{y}^s &= \text{MLP}_1(\mathbf{s}), \\ \mathbf{y}^b &= \text{MLP}_2(\mathbf{s}), \\ \hat{\mathbf{x}}_i &= \mathbf{y}^s_i \cdot \mathbf{x}_i + \mathbf{y}^b_i, \end{aligned} \tag{3}$$

where $\mathbf{y}^s$ are the scaling parameters, $\mathbf{y}^b$ is the bias parameters, $\mathbf{s}$ is the text condition, $\hat{\mathbf{x}}_i$ and $\mathbf{x}_i$ represent the i-th channel of the output and input feature map, respectively. After the affine transformation, we employ a ReLU function to add nonlinearity to fusion process [43]. Since the affine transformation only fuses the text condition separately for each channel, it lacks information fusion across channel dimensions. Therefore, we additionally add a fully connected layer functioned on channel dimensions to facilitate the text-image fusion. Finally, we use instance normalization again to normalize the output feature map.

As shown in Table 3, we still explore other style injection methods for our TransT2I. Extensive experiments demonstrate the effectiveness and superiority of ConIN. Compared with previous attention-based fusion methods [52, 23], ConIN has lower computational loss. Compared with concatenation [36] and affine [44, 54] methods, ConIN enhances the text-image fusion across channel dimensions with slight computational cost.

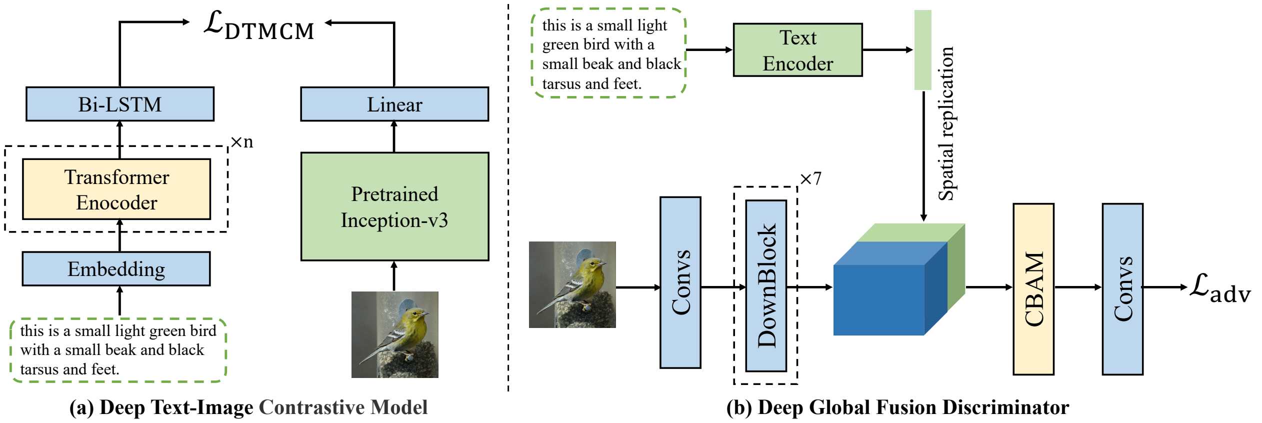

$\mathcal{L}_{\text{DTMCM}}$
Bi-LSTM
Linear
×n
Transformer Encoder
Pretrained Inception-v3
Embedding
this is a small light green bird with a small beak and black tarsus and feet.
(a) Deep Text-Image Contrastive Model
this is a small light green bird with a small beak and black tarsus and feet.
Text Encoder
Spatial replication
×7
Convs
DownBlock
CBAM
Convs
$\mathcal{L}_{\text{adv}}$
(b) Deep Global Fusion Discriminator

Figure 4. (a) Deep Text-Image Contrastive Model (DTMCM): We propose a new text-image aligned tool DTMCM, which can achieve better representation learning of text features than DAMSM [52]. (b) Deep Global Fusion Discriminator: In order to achieve global text-image fusion in discriminator, we introduce the CBAM module [50] to enhance text-image fusion across spatial and channel dimensions.

3.4. Deep Text-Image Contrastive Model

The previous DAMSM [52] suffers from two problems. First, the text encoder adopted by DAMSM ignores the attention between different words in a sentence. Second, DAMSM loss includes word loss and sentence loss. We argue that the word loss can't provide positive feedback to generator. The word loss computes the similarity between sub-regions of image and words in a sentence. Natural language descriptions are high-level semantics, while the sub-regions of image are relatively low-level [3, 55, 24]. Thus, it's unreasonable to use word loss to supervise generator to synthesize text-matched images. To solve these, we propose Deep Text-Image Contrastive Model (DTMCM), the architecture is shown in Figure 4(a).

The Image Encoder. The image encoder we follow the design of DAMSM [52], which is built upon the Inception-v3 model [42] pretrained on ImageNet [38]. Meanwhile, the global feature vector $\mathbf{g} \in \mathbb{R}^{2048}$ is extracted from the last average pooling layer of Inception-v3. Finally, we convert the image features to a common semantic space of text features by adding a linear network: $\mathbf{v} = \mathbf{W}\mathbf{g}$, where $\mathbf{W}$ is the learnable parameter matrix, $\mathbf{v}$ is the global vector for the whole image.

The Text Encoder. Unlike DAMSM [52], we propose a novel text encoder to facilitate representation learning for text features. First, we get the initial word matrix. Then, the word matrix is processed by n standard transformer encoders [47] (default n = 3). The self-attention module can calculate the attention of different words in a sentence and enhance the impact of important words on model. Finally, a bidirectional LSTM is used to extract the context relationships in a sentence, and the result is the output of our text encoder. Compared with the text encoder in DAMSM, we can better reflect the attention of different words in a sentence, and thus better realize the representation learning of sentence features.

The DTMCM Loss. The DTMCM loss is employed to compute the semantic loss between entire image and sentence, which is aimed to supervise the generator to synthesize text-matching images. Unlike DAMSM [52], the DTMCM loss only calculates sentence loss. First, for the image-sentence pairs $\mathbf{I}$ and $\mathbf{T}$, we calculate the cosine similarity, the mathematical form is as follows:

$$\text{Sim}(\mathbf{I}, \mathbf{T}) = \frac{\mathbf{v}^\top \mathbf{e}}{||\mathbf{v}||||\mathbf{e}||}, \tag{4}$$

where $\mathbf{v}$ is image vector and $\mathbf{e}$ is sentence vector, respectively. Following previous work [62, 52], we define the loss function as the negative log posterior and prior probability that the images are matched with their corresponding text descriptions. For a batch of image-sentence pairs$\{(\mathbf{I}_i, \mathbf{T}_i)_{i=1}^M\}$, the posterior and prior probability of sentence $\mathbf{T}_i$ being matching with image $\mathbf{I}_i$ is computed as:

$$\begin{aligned} \mathcal{L}_1 &= -\sum_{i=1}^{M} \log \frac{\exp(\gamma_1 \text{Sim}(\mathbf{I}_i, \mathbf{T}_i))}{\sum_{j=1}^{M} \exp(\gamma_1 \text{Sim}(\mathbf{I}_i, \mathbf{T}_j))}, \\ \mathcal{L}_2 &= -\sum_{i=1}^{M} \log \frac{\exp(\gamma_2 \text{Sim}(\mathbf{T}_i, \mathbf{I}_i))}{\sum_{j=1}^{M} \exp(\gamma_2 \text{Sim}(\mathbf{T}_i, \mathbf{I}_j))}, \end{aligned} \tag{5}$$

where γ_1, γ_2 are hyper-parameters. The total DTMCM loss is defined as:

$$\mathcal{L}_{\text{DTMCM}} = \mathcal{L}_1 + \mathcal{L}_2. \tag{6}$$

By this design, DTMCM loss can provide better feedback to the generator than DAMSM. As shown in Table 4, extensive experiments confirm the effectiveness and superiority of DTMCM. As shown in Table 5, compared with CLIP [33], DTMCM is more lightweight and efficient, which confirms that DTMCM is more suitable for our

Method	CUB			COCO		MM CelebA-HQ		-	
	FID↓	IS↑	R-precision↑	FID↓	R-precision↑	FID↓	R-precision↑	Params	Speed
StackGAN++ [60]	15.30	4.04 ± .06	-	81.59	-	-	-	-	-
AttnGAN [52]	23.98	4.36 ± .03	0.246	35.49	0.183	125.98	0.233	169M	0.07s
DM-GAN [63]	16.09	4.75 ± .07	0.287	32.64	0.236	131.05	-	46M	-
DAE-GAN [37]	15.19	4.42 ± .04	0.321	28.12	0.257	-	-	-	-
XMC-GAN [58]	-	-	-	29.63	0.278	-	-	166M	
TediGAN [51]	-	-	-	-	-	106.37	0.275	-	-
DF-GAN [43]	14.81	5.10 ± .- -	0.306	19.32	0.278	137.60	-	19M	-
VQ-Diffusion-B [12]	11.94	-	-	19.75	-	-	-	370M	6.42s
SSA-GAN [24]	15.61	5.17 ± .08	0.326	19.37	0.264	-	-	110M	-
Lafite [62]	11.27	**5.58 ± .- -**	0.350	**10.32**	0.318	12.54	-	75M	0.06s
TransT2I (ours)	**10.06**	5.42 ± .02	**0.369**	11.94	**0.338**	**11.87**	**0.335**	32M	0.03s

Table 1. The results of IS, R-precision and FID compared with the state-of-the-art methods on the test set of CUB, Multi-Modal CelebA-HQ and COCO. ↓ means lower is better. ↑ means higher is better. The speed refers to the time to generate one image, which is tested on an NVIDIA RTX 3090.

model. More importantly, the proposed DTMCM also can be utilized as a general-purpose tool to achieve cross-modal semantic supervision for text-to-image synthesis.

3.5. Discriminator and Loss Function

Deep Global Fusion Discriminator. The architecture of discriminator is shown in Figure 4(b). Previous works [36, 44, 24] tend to roughly concatenate text features and visual features in the discriminator, and following convolutional networks are used to facilitate the fusion of the two features. We argue that pure convolutional networks are beneficial to extract local features, but can‘t achieve global cross-modal text-image fusion. The recent RAT-GAN [54] notices the problem, but only enhanced the text-image fusion across spatial dimensions. Different from previous works, we introduce the CBAM [50] module in discriminator to facilitate global cross-modal text-image fusion across spatial and channel dimensions.

Loss Function. Similar to DF-GAN [44], we employ hinge loss as the discriminator loss. In order to smooth the gradient, we adopt the MA-GP loss [44]. The training loss of discriminator is as follows:

$$\begin{aligned}\mathcal{L}_{adv}^{D} &= \mathbb{E}_{\mathbf{x}\sim P_{data}}[\max(0, 1 - D(\mathbf{x}, \mathbf{s}))] \\ &+ \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim P_{G}}[\max(0, 1 + D(\hat{\mathbf{x}}, \mathbf{s}))] \\ &+ \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim P_{data}}[\max(0, 1 + D(\mathbf{x}, \hat{\mathbf{s}}))],\end{aligned} \tag{7}$$

where $\mathbf{x}$ is real image, $\hat{\mathbf{x}}$ is generated image, $\mathbf{s}$ is matched sentence, $\hat{\mathbf{s}}$ is unmatched sentence, D is the discriminator, respectively. The training loss of generator is as follows:

$$\begin{aligned}\mathcal{L}_{G} &= \lambda_1\mathcal{L}_{\text{DTMCM}} + \mathcal{L}_{adv}^{G}, \\ \mathcal{L}_{adv}^{G} &= -\mathbb{E}_{\mathbf{x}\sim P_{data}}[D(\hat{\mathbf{x}}, \mathbf{s})],\end{aligned} \tag{8}$$

where λ_1 is a hyper-parameter, respectively.

4. Experiments

In the section, we introduce the datasets, training details, and evaluation details. Then, we conduct our experiments on three challenging datasets to verify the effectiveness and superiority of our proposed TransT2I.

Datasets. We evaluate the proposed model on three challenging datasets, i.e., CUB bird [48], COCO [25] and Multi-Modal CelebA-HQ [51]. The COCO datasets contain 80k images for training and 40k images for testing. Each image has five language descriptions. The CUB bird datasets (200 categories) contain 8855 training images and 2933 testing images. Each image has 10 text descriptions. The Multi-Modal CelebA-HQ datasets contain 24k images for training and 6k images for testing. Each image has 10 language descriptions. All the images are scaled to resolution 256×256.

Training Details. Our model is implemented in PyTorch. The Adam optimizer [20] with $\beta_1 = 0.0$ and $\beta_2 = 0.9$ is used in the training. The learning rate is set to 5×10^{-5} for generator and 2×10^{-4} for discriminator according to TTUR [14]. The hyper-parameters λ_1, γ_1 and γ_2 are set to 0.02, 10 and 10, respectively.

Evaluation Details. The Fréchet Inception Distance (FID) [14], Inception Score (IS) [40], and top-1 R-precision [52] are used to evaluate the performance of our work. For FID, it computes the Fréchet distance between the distribution of the generated images and real-world images in the feature space of a pre-trained Inception v3 network [42]. For IS, it computes the Kullback-Leibler (KL) divergence between a conditional distribution and marginal distribution. Lower FID and higher IS mean model achieves better performance. For R-precision, we use CLIP [32] to calculate the cosine similarity between original image and given description. Following previous works [24, 43], we do not use IS on COCO and Multi-Modal CelebA-HQ datasets because it can’t evaluate the image quality well.

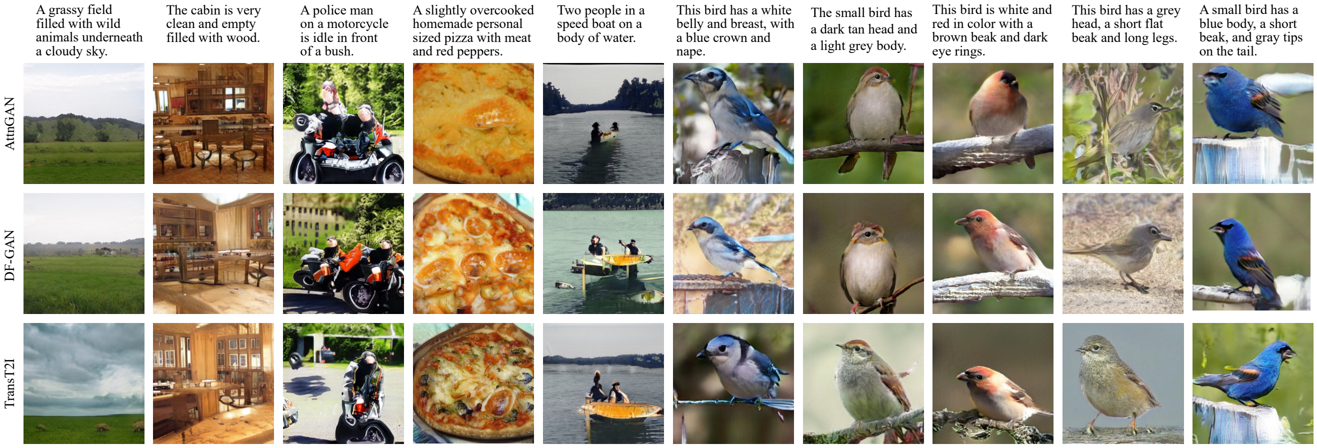

A grassy field filled with wild animals underneath a cloudy sky.
The cabin is very clean and empty filled with wood.
A police man on a motorcycle is idle in front of a bush.
A slightly overcooked homemade personal sized pizza with meat and red peppers.
Two people in a speed boat on a body of water.
This bird has a white belly and breast, with a blue crown and nape.
The small bird has a dark tan head and a light grey body.
This bird is white and red in color with a brown beak and dark eye rings.
This bird has a grey head, a short flat beak and long legs.
A small bird has a blue body, a short beak, and gray tips on the tail.
AttnGAN
DF-GAN
TransT2I

Figure 5. Qualitative comparison between AttnGAN [52], DF-GAN [43], and our proposed TransT2I conditioned on text descriptions from the test set of COCO datasets (1st - 5th columns) and CUB datasets (6th - 10th columns).

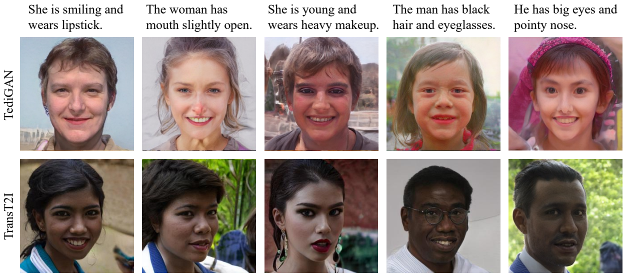

She is smiling and wears lipstick.
The woman has mouth slightly open.
She is young and wears heavy makeup.
The man has black hair and eyeglasses.
He has big eyes and pointy nose.
TediGAN
TransT2I

Figure 6. Qualitative comparison between TediGAN [51] and our proposed TransT2I conditioned on text descriptions from the test set of Multi-Modal CelebA-HQ [51].

Variant	FID ↓	R-precision ↑
Mix Attention (ours)	**10.06**	**0.369**
w/o LightMSA	15.89	0.312
w/o ConvNets	14.88	0.328
w/o Parallel Design	13.83	0.330
→ Grid Self-Attention [16]	18.34	0.258
→ Double Attention [57]	16.24	0.306
→ Multi-axis Self-Attention [61]	14.35	0.351

Table 2. Ablation Study of Mix Attention on the test set of CUB. ↓ means lower is better. ↑ means higher is better.

Moreover, we evaluate the number of parameters (Params) and inference speed (Speed) to compare with current methods.

4.1. Quantitative Evaluation

As shown in Table 1, we conduct the quantitative comparison between our proposed FuseGAN and previous methods, such as: AttnGAN [52], StackGAN++ [60], DF-GAN [43], Lafite [62]. On three challenging datasets, our proposed TransT2I demonstrates the state-of-the-art performance over prior text-to-image works, proving the promise of transformer-based GAN for text-to-image synthesis. On CUB [48] and Multi-modal CelebA-HQ [51], TransT2I achieves a state-of-the-art FID 10.06 and 11.87 that exceeds all previous works. On COCO [25], TransT2I achieves a competitive FID 11.94 only with 42% parameters compared with current SOTA model. On CUB datasets [48], compared with the single-stage baseline DF-GAN [43], our proposed TransT2I decreases FID from 14.81 to 10.06 and improves IS from 5.10 to 5.86, R-precision from 0.306 to 0.369. On Multi-Modal CelebA-HQ datasets, TransT2I achieves state-of-the-art performance, which decreases the current SOTA FID from 12.54 to 11.87. On COCO datasets, compared with current state-of-the-art model Lafite [62], our proposed TransT2I achieves comparable performance, which TrasnT2I improves R-precision from 0.306 to 0.369. Compared with diffusion methods VQ-Diffusion-S [12], our proposed TransT2I achieves more advanced performance, which decreases FID from 30.17 to 11.94. Importantly, our TransT2I only requires 0.03s to generate one image, which is almost 210× faster than VQ-Diffusion-B.

4.2. Qualitative Evaluation

As shown in Figure 5, we conduct the qualitative comparison between our proposed TransT2I, DF-GAN [43] and AttnGAN [52] on CUB datasets and COCO datasets. TrasnT2I generates realistic and semantically consistent images. For instance, in the 3-nd column, our proposed TransT2I generates realistic images, while the image generated by AttnGAN and DF-GAN appears blurry and semantically inconsistent. In the 7-th column, given the sentence “The small bird has a dark tan head and a light grey body”, the images generated by AttnGAN and DF-GAN don’t appear “grey body”, while our proposed TransT2I mentions all attributes. As shown in Figure 6, we conduct the qualita-

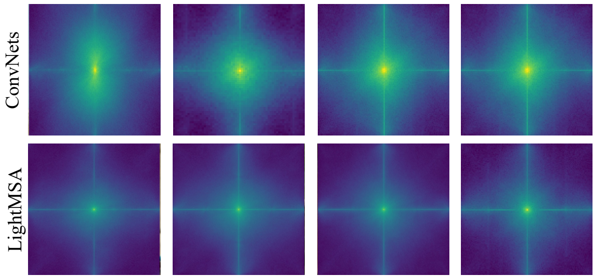


Figure 7. Frequency magnitude from 4 output channels of LightMSA and ConvNets in the proposed Mix Attention. The lighter the color, the larger the magnitude. A pixel that is closer to the center means a lower frequency.

Variant	FID ↓	R-precision ↑
ConIN (ours)	**10.06**	**0.369**
w/o Linear layer	13.69	0.302
w/o Instance Normalization	19.88	0.227
→ AdaIN [19]	15.31	0.288
→ Affine [44]	16.14	0.303
→ CBN [1]	15.25	0.291

Table 3. Ablation Study of ConIN on the test set of CUB. ↓ means lower is better. ↑ means higher is better.

tive comparison between our proposed FuseGAN and TediGAN [51] on Multi-Modal CelebA-HQ datasets. Compared with TediGAN, TransT2I significantly improves the quality of generated images. For example, in the 4-th column, the image generated by TediGAN doesn't involve the attribute of "eyeglasses" and has some blur, while TransT2I generates clear and text-matching images.

4.3. Ablation Study

To verify the superiority of each component in our proposed TransT2I, we deploy our experiments on the CUB test set [48].

Mix Attention. As shown in Table 2, we conduct experiments to verify the effectiveness of MixAttention. Our attempts to remove LightMSA or ConvNets resulted in performance degradation. Furthermore, our attempt to replace the parallel design with a successive way also resulted in poor performance. Compared with the previous attention methods Grid Self-Attention [16], Double Attention [57] and Multi-axis Self-Attention [61], Mix Attention is more suitable for our work and obtains better performance. As shown in Figure 7, we show the visualization results of LightMSA and ConvNets. We show the frequency magnitude of LightMSA and ConvNets, which indicates that the LightMSA tends to capture low-frequency signals and the ConvNets tends to capture high-frequency signals.

ConIN. As shown in Table 3, we conduct experiments to verify the effectiveness of ConIN. Experiments verify that linear layers and instance normalization have a positive impact on our model. Linear layers enhance the fusion of text-image information across channels, while instance normalization significantly improves performance. Besides, we compare with AdaIN [19], CBN [1] and affine [44]. Our proposed ConIN is more suitable for our model and achieves better results.

Variant	FID ↓	R-precision ↑
DTMCM (ours)	**10.06**	**0.369**
w/ Text Encoder in DAMSM [52]	14.72	0.287
w/ Words Loss	17.38	0.247
→ DAMSM	17.91	0.233
AttnGAN w/ DAMSM	23.98	0.246
AttnGAN w/ DTMCM (ours)	**19.93**	**0.265**

Table 4. Ablation Study of DTMCM on the test set of CUB. ↓ means lower is better. ↑ means higher is better.

Variant	Params	Time ↓	FID(CUB) ↓	FID(COCO) ↓
DTMCM	32M	**13mins**	**10.06**	**11.94**
CLIP [33]	150M	20mins	12.14	12.79

Table 5. The comparison between the proposed DTMCM and CLIP(ViT-B/32) when used in our TransT2I. Time refers to the cost of training TransT2I one iteration on CUB with an NVIDIA RTX 3090 GPU. DTMCM is more efficient and lightweight for our TransT2I.

DTMCM. As shown in Table 4, we conduct experiments to verify the effectiveness of DTMCM. First, the experiments verify the superiority of the designed text encoder of DTMCM and the removal of word loss. Then, DTMCM achieves more advanced performance compared with previously adopted alignment tools DAMSM [52] and CLIP [62, 32]. Finally, we try to adopt DTMCM in AttnGAN [52], which significantly improves the performance. As shown in Table 5, we compare our DTMCM with CLIP, and the experimental results demonstrate the effectiveness for our TransT2I.

5. Conclusion

In this paper, we propose TransT2I, a transformer-based GAN for text-to-image synthesis. We propose a novel attention mechanism Mix Attention, which can simultaneously capture global relationships and local details while enjoying linear computational complexity. Besides, we propose Conditioned Fusion Instance Normalization and Deep text-image Contrastive Model to further improve model capacity. Extensive experiments on three challenging benchmarks demonstrate the state-of-the-art performance. In the future, we try to employ novel attention mechanisms to build a transformer-based GAN for text-to-image synthesis.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 8

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020. 1

[3] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5659–5667, 2017. 2, 5

[4] Qiang Chen, Qiman Wu, Jian Wang, Qinghao Hu, Tao Hu, Errui Ding, Jian Cheng, and Jingdong Wang. Mixformer: Mixing features across windows and dimensions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5249–5259, 2022. 4

[5] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10911–10920, 2020. 2

[6] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 4

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[8] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. 2

[9] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12124–12134, 2022. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the Advances in neural information processing systems (NeurIPS)*, volume 27, 2014. 1

[12] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10696–10706, 2022. 1, 2, 6, 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 2

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 6

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[16] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 14745–14758, 2021. 1, 2, 7, 8

[17] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12104–12114, 2020. 3, 4

[18] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. 4

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 2, 3, 4, 8

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 6

[21] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 1

[22] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *arXiv preprint arXiv:2107.04589*, 2021. 1, 2

[23] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 2, 4

[24] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 18187–18196, 2022. 2, 5, 6

[25] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 6, 7

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 1, 2, 4

[27] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018. 4

[28] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. *arXiv preprint arXiv:2205.13213*, 2022. 1, 3

[29] 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 (CVPR)*, pages 8983–8992, 2022. 1, 2

[30] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 1

[31] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1505–1514, 2019. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 6, 8

[33] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 4, 5, 8

[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8821–8831. PMLR, 2021. 2

[35] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 49–58, 2016. 2

[36] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1060–1069, 2016. 4, 6

[37] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13960–13969, 2021. 2, 6

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[40] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. 6

[41] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. *arXiv preprint arXiv:2205.12956*, 2022. 1, 3

[42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 5, 6

[43] 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 (CVPR)*, pages 16515–16525, 2022. 2, 4, 6, 7

[44] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020. 2, 4, 6, 8

[45] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 24261–24272, 2021. 3, 4

[46] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 10347–10357. PMLR, 2021. 1

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 1, 2, 3, 5

[48] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 3, 6, 7, 8

[49] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao.

Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 568–578, 2021. 1, 3, 4

[50] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 5, 6

[51] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2256–2265, 2021. 3, 6, 7, 8

[52] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1324, 2018. 1, 2, 4, 5, 6, 7, 8

[53] Tao Yang, Haokui Zhang, Wenze Hu, Changwen Chen, and Xiaoyu Wang. Fast-parc: Position aware global kernel for convnets and vits. *arXiv preprint arXiv:2210.04020*, 2022. 1

[54] Senmao Ye, Fei Liu, and Minkui Tan. Recurrent affine transformation for text-to-image synthesis. *arXiv preprint arXiv:2204.10482*, 2022. 2, 4, 6

[55] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4709–4717, 2017. 2, 5

[56] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, 2021. 1, 3

[57] 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 (CVPR)*, pages 11304–11314, 2022. 1, 2, 3, 4, 7, 8

[58] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–842, 2021. 6

[59] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5907–5915, 2017. 2

[60] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, 2018. 2, 6, 7

[61] Long Zhao, Zizhao Zhang, Ting Chen, Dimitris Metaxas, and Han Zhang. Improved transformer for high-resolution gans. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 18367–18380, 2021. 1, 2, 7, 8

[62] 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 (CVPR)*, pages 17907–17917, 2022. 2, 5, 6, 7, 8

[63] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5810, 2019. 6