Generative Transformer for Auto Computer-aided Design

Thao Nguyen Phuong Nam Nguyen Xuan Hidetomo Sakaino
AI-Image Group, Data Solution Dept., (FCJ.ABC) FPT Consulting Japan, FPT Software
{ThaoNP47, NamNX21, HidetomoS}@fpt.com

Abstract

Reconstructing the 2D or 3D form of an object from various 2D sketches is essential in Computer-Aided Design (CAD). Despite many improvements that have been made to deep neural networks, generating isometric images automatically from three orthographic view line drawings using deep learning has yet to be solved. Current image-to-image translation methods typically transform only one input image to another image domain. In this paper, we introduce a new approach using a GAN model, namely IsoTGAN, which incorporates Transformer in its generator. This method takes three images of object's front, side, and top view as input, then analyzes the spatial and geometrical relation between each view by an encoder, the output vector is fed to the Transformer in the generator along with image features to enable long-range interactions across three orthographic view input images and finally generates the corresponding isometric view image of the object. We also propose a novel Gaussian Enhanced Euclidean attention mechanism and a geometry-constrained loss function for improved isometric contour reconstruction. Extensive experiments on the SPARE3D dataset show IsoTGAN's promising capabilities in generating isometric view task, demonstrating its effectiveness.

1. Introduction

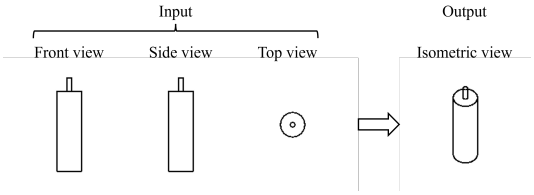

Figure 1. Example of a pair of three-view line drawings and their corresponding isometric image. The task is to generate isometric view image given front, side, and top view image.

In today's digital era, CAD software is essential for creating 3D shapes in various industrial sectors, including automotive, aerospace, manufacturing, and architectural design. The most commonly used technique for depicting 3D objects on a 2D surface is orthographic projection, also known as orthogonal projection or analemma. This method projects lines from the object perpendicular to the projection plane, resulting in a parallel representation. Designers often use 2D orthographic drawings to showcase their concepts in the early stages of design. Nevertheless, to further explore and understand the object, a 3D model is necessary. Isometric view images are crucial for the 3D reconstruction process as they retain a significant amount of information about the 3D object. Therefore, there is a strong demand for a method that can convert three-view contour drawings into isometric view images, which would greatly facilitate the design process and increase overall efficiency.

In recent years, numerous research papers have introduced deep learning models for 2D/3D CAD modeling [11, 15, 19, 22, 32, 36, 44, 46, 49]. However, these studies typically treat the task as a sequence-to-sequence issue, outputting sequences of CAD commands for creating 2D drawings or reconstructing 3D objects. Only one study has explored the generation of isometric images from three-view orthographic line drawings [19], and this approach similarly generates CAD commands.

Isometric view images are crucial for reconstructing 3D objects. Generating these images from drawings of three orthographic views can be approached as a multi-images-to-image translation problem. Unlike common image-to-image translation tasks, there is a strong spatial and geometrical relation between each orthographic view of the object, as shown in Fig. 1. To generate a comprehensive isometric view image, it is necessary to understand the relation between each input view and isometric view. Previous image-to-image translation methods exploit GAN [14, 34], Autoregressive models [41], Variational AutoEncoders (VAEs) [26], Normalizing Flows [7], and recently, Diffusion models [37]. Nevertheless, these methods only take one image or text as input and map it to the target domain without the need to consider the relation between multiple input images. Although recent Diffusion methods [12, 18] can take many

input images to generate 3D scenes, they deal with natural 2D images, which are very different from images containing only flatted contour lines. Since many different objects have the same single-view image, these Diffusion frameworks [28, 29, 50] that produce 3D scenes from only one image may not generate correct isometric view images.

A number of variants of GAN models with attention mechanism [3, 5, 9, 40, 54] have been proposed. However, they are weak at generating locally detailed texture images without sufficient constraints. Moreover, they do not have any mechanism to tackle the important relation between each view of the object. To address this problem, we propose IsoTGAN, a GAN-based model, which directly transforms three orthographic view contour drawings to the corresponding isometric view contour image by incorporating an encoder to embed the spatial and geometrical relation between each view into a representation vector. To enhance the ability of IsoTGAN to reconstruct local details of isometric view images, the Transformer blocks are equipped in the generator, which take both the vector output by the above encoder and image feature as input and model the attention between them, ensuring long-range interactions across three orthographic views and constraint guidance between each input image patch.

Moreover, although there are many researches about spatial and channel attention, they cannot deal with geometrical matching enhancement. Therefore, the novel Gaussian Enhanced Euclidean norm (GEE) attention is proposed to focus more on important view relation features, which takes responsibility for efficiently mapping flattened projection to isometric projection. Here, we use a Gaussian function which takes the Euclidean norm of channel or spatial dimension as input to refine features output by the channel and spatial attention mechanism. The loss function is adjusted by a geometry constraint to enhance the proposed IsoTGAN capability in contour-to-contour transformation. In summary, the contributions of this paper are as follows:

- To the best of our knowledge, we are the first to investigate the spatial and geometrical relation between each orthographic view contour drawing of an object on raster images using an encoder network.
- We propose IsoTGAN, a GAN-based model whose generator is employed with Transformer that takes two types of input to capture long-range features of object's views. Compared to other SOTA GAN-based methods, many small parts and features of objects in isometric view have been generated notably better.
- An attention mechanism, GEE, is presented for effectively recognizing each object view's relation. A modification of the GAN loss function by a geometry constraint is also proposed to generate the isometric view image for facilitating the 3D reconstruction process.
- Extensive experiments on SPARE3D [16] dataset show superior results on isometric view generation task compared to previous methods.

2. Related Works

2.1. Transformer for Image Generation

Originally developed for natural language processing (NLP), the Transformer architecture [42] uses multi-head self-attention and stacked feed-forward MLP layers to capture the long-term relationships between words. Dosovitskiy *et al.* [8] have adapted this architecture for highly competitive ImageNet classification by viewing an image as a sequence of 16×16 visual words. This approach boasts robust representational abilities without the human-defined inductive biases. In contrast, CNNs display a pronounced bias toward local features and spatial invariance, which is achieved by sharing filter weights across different locations. Recently, many papers have incorporated Transformer into GAN architecture for image generation. [20] proposes GANformer, in which the Bipartite Transformer with simplex and duplex attention is applied between the image features and latent variables in the generator. In [23], a Transformer-based generator and discriminator with grid self-attention for efficient computation are presented, and the authors did not use any convolution in model architecture. Another work [53] employs double attention in Swin Transformer [30] which additionally attends to the style tokens. Nevertheless, there is no other work that incorporates Transformer to reconstruct one image from multiple images.

2.2. Image-to-image Translation

Numerous methods have utilized GANs for image-to-image translation tasks. This process involves converting an input image from one domain to another, using either paired or unpaired datasets for training. The initial pix2pix framework [21] employs image-conditional GAN [31] for various paired image-to-image translation tasks, such as turning semantic labels into synthetic images, given groundtruth image. Additionally, various techniques have been developed for unpaired translation [55], unsupervised cross-domain generation [39], multi-domain translation [4], and few-shot translation [27]. More recently, Diffusion models [37] have shown outstanding performance in areas like image generation [6, 17], image super-resolution [13, 51], image restoration [10, 47], unpaired image-to-image translation [38, 48], and image editing [2, 24], surpassing GANs in these applications. However, Diffusion models need a lot of computation resources for training and inferencing, which hinders their ability in the industrial field. Moreover, typical image-to-image translation methods primarily aim to convert an image to one or many target domains without involving multiple input images and the relation between them.

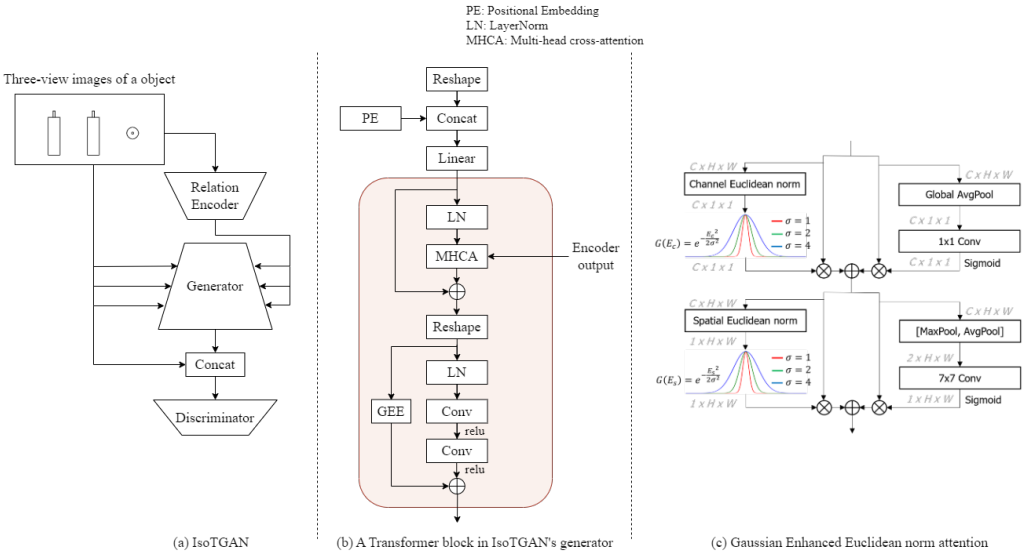

Figure 2. The overall architecture of the proposed IsoTGAN. $\otimes$ represents broadcast element-wise multiplication and $\oplus$ denotes element-wise addition. $\mathbf{E_c}$ and $\mathbf{E_s}$ represent Channel Euclidean norm and Spatial Euclidean norm, respectively.

While many of these models focus on generating realistic photos or animations, the generation of contour line drawings has been less explored. Contrary to these approaches, our framework addresses the challenge of converting three orthographic views contour drawings into their respective isometric view image, necessitating an understanding of the relation among multiple input images and the ability to produce contour images.

2.3. Deep Generative Models For CAD

Recently, several researches have focused on generating CAD commands either as 2D contour drawings [11, 32, 36, 44] or 3D objects [15, 19, 22, 46, 49]. All of these methods employ Transformer-based encoder-decoder architecture, using 2D sketches to create CAD modeling programs. Among these, the work by Hu *et al.* (2023) [19] is most closely related to our study. They use a generative model to develop a method for reconstructing 3D CAD models from three orthographic views. However, this method deals with sequences rather than images and generates a sequence of CAD programs.

2.4. Isometric View Image Generation

Han *et al.* [16] apply the pix2pix framework [21] for generating isometric view images from three orthographic views, proposing a baseline that does not account for any relation between different views of object during training and generating. In this paper, we introduce a new approach that incorporates that relation into a vector, which is then used to generate the isometric view image.

3. Proposed Method

This section introduces the IsoTGAN framework, which is proposed to automatically generate isometric view image from three orthographic view contour drawings. Initially, an encoder is utilized to capture the spatial and geometric relation between each orthographic view of the object and embed these relations into a vector. This vector is then input into IsoTGAN along with orthographic view images at three different scales. The Transformer blocks are employed, which are able to model long-range interactions and constraints across three orthographic views of the object, to produce an isometric image. Additionally, the novel GEE attention mechanism and a modification to the GAN loss function are detailed. The overall architecture of the proposed method is illustrated in Fig. 2.

3.1. Three-view Contour Drawings Encoder

Our framework enhances the generation of isometric view image by using a three-view relation vector as input for the generator. This setup involves using an encoder to convert a triplet of front, side, and top view images into a vector, which is then provided to the generator. The image encoder comprises 6 stride-2 convolutional layers followed by two linear layers that calculate the mean and variance of the output distribution. The encoder and generator together form a VAE. In this framework, the encoder aims to capture the spatial and geometrical relation of the three views, while the generator combines the encoded vector with information from three-view images to reconstruct the isometric view image. During testing, the encoder acts as a relation guidance network to direct the generator in accurately con-

structing the isometric view image. We implement a KL-Divergence loss term [26] to facilitate training as follows:

$$\mathcal{L}_{\mathrm{KLD}} = \mathcal{D}_{\mathrm{KL}}(q(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z})), \tag{1}$$

where the prior distribution $p(\mathbf{z})$ is a standard Gaussian distribution and the variational distribution q is fully determined by a mean vector and a variance vector [26].

3.2. IsoTGAN's Generator

3.2.1. Transformer Block

The conventional Transformer block consists of a multi-head self-attention and a feed-forward layer. In the self-attention layer, all pairwise relationships between input elements are considered, with each element being updated through attention to all others. In the task of generating isometric view image from three orthographic view contour drawings, only attending to image features is not adequate because the strong spatial and geometrical relation between each view are neglected. Therefore, we perform cross-attention between input image features and the vector output by the encoder to enhance reconstruction result. Formally, let $X^{n\times d}$ denote an input set of n image patch vectors of dimension d, and $Y^{m\times d}$ is a set of m spatial and geometrical vectors output by the relation encoder, the cross-attention in one head is then computed as follow:

$$\begin{aligned} Q &= XW_q \\ K &= YW_k \\ V &= YW_v \\ Attention(Q,K,V) &= softmax\left(\frac{QK^\top}{\sqrt{d}}\right)V, \end{aligned} \tag{2}$$

where Q, K, V represent query, key, value matrices, all keeping the dimension d, and W denote weight matrices. The positional embedding is also added to provide the distinct spatial position of each image patch.

After the multi-head cross-attention (MHCA) layer, two 3×3 convolution layers are utilized, then ReLU [1] activation function is applied for nonlinearity. We use a sinusoidal positional encoding across horizontal and vertical dimensions for the image features X to ensure features location inside the image. The output of MHCA layer goes through the GEE attention (detailed in section 3.2.2), then is added to the output of the convolution layer so that the model can focus more on important image features.

IsoTGAN is our adaptation of SPADE [33]. The network architecture follows the design of SPADE, except that SPADE blocks are replaced by Transformer blocks. With cross-attention, instead of intensively modeling interactions between all pixel pairs in an image, it enables adaptive long-range interactions between distant areas in an adequate manner, through a global spatial and geometrical vector. This vector selectively collects information from the entire input and distributes it to related regions, controlling the way the model generates isometric image, in which the output image must follow projection physics rules. Intuitively, information can move in both directions, from the local pixel level to the global high-level representation and back again. Given that soft-attention is inclined to group elements by proximity and content similarity, the transformer architecture allows the model to precisely modulate local semantic regions, avoiding discontinuity in generating the contour of object shape.

3.2.2. Gaussian Enhanced Euclidean norm Attention

The relation representation vector in three-view drawings can carry extraneous information from axis contours, which traverse the object's center, but not the object outlines. To enhance understanding of each object view's relation as well as minimizing the influence of axes contour in isometric view image generation, we introduce the Gaussian Enhanced Euclidean norm (GEE) attention. This approach hypothesizes that smaller attention activations correlate with global contexts of higher absolute values [35]. The GEE framework includes both channel and spatial attention modules, each split into two branches, as depicted in Fig. 2. Unlike the Gaussian Context Transformer approach [35], the left branch of GEE directly uses the Euclidean norm of feature maps as input to a Gaussian function without the normalization operation of global average pooling (GAP). The intuition behind is that the Euclidean norm quantifies the magnitude of a vector or matrix, with larger Euclidean norm indicating greater deviation from the vector or matrix to its origin, so constraining the Euclidean norm with Gaussian function will improve model generalizability. Additionally, we argue that this principle applies to spatial dimensions as well, thus extending it to capture global spatial information. The right branch is quite similar to the CBAM architecture [45], except that we utilize only GAP followed by a 1×1 convolution in its channel attention module to reduce model parameters.

Concretely, given a feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ as input, C denotes the number of channels and H, W are spatial dimension, GEE computes attention as follows.

$$\begin{aligned} \mathbf{F}' &= \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{F} \oplus \mathbf{G}(\mathbf{E_c}) \otimes \mathbf{F} \\ \mathbf{F}'' &= \mathbf{M_s}(\mathbf{F}') \otimes \mathbf{F}' \oplus \mathbf{G}(\mathbf{E_s}) \otimes \mathbf{F}', \end{aligned} \tag{3}$$

where $\otimes$ represents broadcast element-wise multiplication and $\oplus$ denotes element-wise addition. $\mathbf{F}''$ is the final output. $\mathbf{E_c} \in \mathbb{R}^{C\times 1\times 1}$ and $\mathbf{E_s} \in \mathbb{R}^{1\times H\times W}$ represents Channel Euclidean norm and Spatial Euclidean norm, respectively, and is formulated as follows.

$$\mathbf{E_c} = \left\{ e_{ck} = \sqrt{\textstyle\sum_{i=1}^{W}\sum_{j=1}^{H} \mathbf{F}_k(i,j)^2} : k \in \{1,\ldots,C\} \right\}, \tag{4}$$

$$\mathbf{E_s} = \left\{ e_{sij} = \sqrt{\sum_{k=1}^{C} \mathbf{F}'_{ij}(k)^2} : i \in \{1, \ldots, W\}, j \in \{1, \ldots, H\} \right\}. \tag{5}$$

A Gaussian function $\mathbf{G}(x) = \exp(-\frac{x^2}{2\sigma^2})$ processes the input x with its maximum value at 1, a mean of 0, and a standard deviation σ, aligning with the hypothesis about the relationship between global contexts and attention activations. A larger σ leads to a more uniform distribution among attention activations.

$\mathbf{M_c} \in \mathbb{R}^{C\times1\times1}$ and $\mathbf{M_s} \in \mathbb{R}^{1\times H\times W}$ is channel attention map and spatial attention map, respectively, and computed as

$$\mathbf{M_c}(\mathbf{F}) = \text{Sigmoid}(f^{1\times1}(\text{GAP}(\mathbf{F}))), \tag{6}$$

$$\mathbf{M_s}(\mathbf{F}) = \text{Sigmoid}(f^{7\times7}([\text{MaxPool}(\mathbf{F}); \text{AvgPool}(\mathbf{F})])), \tag{7}$$

where $f^{k\times k}$ represents a convolution operation with kernel size of $k \times k$.

The left branch enhances the output from the right branch by focusing more on significant spatial and geometrical relationships between each view of the object. The output from the GEE block is then added element-wise to the output of the convolution layer.

3.2.3. Loss Function

The conventional conditional GAN framework for image-to-image translation consists of a generator G and a discriminator D. In this paper, the generator G is responsible for converting three orthographic views drawings into isometric view image, whereas the role of discriminator D is to differentiate between real isometric images and those produced by the generator. This framework operates under a supervised learning manner, where the training dataset consists of three views-isometric view image pairs $\{(\mathbf{x_i}, \mathbf{y_i})\}$, with $\mathbf{x_i}$ representing a triplet of object's front, side, and top view image and $\mathbf{y_i}$ being the corresponding isometric view image. Conditional GANs work by modeling the conditional distribution of real images based on the input images through the following minimax game:

$$\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D). \tag{8}$$

For our task, the generator G takes the relation vector output by the encoder E as an extra input, hence, the objective function $\mathcal{L}_{\text{GAN}}(G, D)$ is computed as

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})}[\log D(\mathbf{x}, \mathbf{y})] + \mathbb{E}_{\mathbf{x}}[\log(1 - D(\mathbf{x}, G(\mathbf{x}, E(\mathbf{x}))))]. \tag{9}$$

The above objective function does not include the geometry constraint of isometric view image, which is crucial for reasonable generation, i.e. the generated image must follow projection physics rules. To incorporate the geometric constraints of the isometric view image, we use a geometric transformation function $f(\cdot)$. We input both the combined three-view images $\mathbf{x}$ and their transformed versions $\tilde{\mathbf{x}} = f(\mathbf{x})$ into the generator G. In this study, we apply vertical and horizontal flipping. The aim is to minimize the discrepancy between the generated isometric image $\mathbf{g} = G(\mathbf{x}, \mathbf{z})$ and its flipped version $\tilde{\mathbf{g}} = f(G(f(\mathbf{x}), \mathbf{z}))$, where $\mathbf{z} = E(\mathbf{x})$ is the output vector from the encoder. The geometry loss is calculated as follows:

$$\mathcal{L}_{geo}(G) = \mathbb{E}_{\mathbf{x},\mathbf{z}} \left[\|G(\mathbf{x}, \mathbf{z}) - f(G(f(\mathbf{x}), \mathbf{z}))\|_1 \right]. \tag{10}$$

In contrast to other image-to-image translation tasks, an isometric view image consists solely of the contour lines of the object. Consequently, a pixel can only assume one of two values, 0 or 255, where 0 indicates that the pixel is part of the object's contour, and 255 denotes that it belongs to the background. The total loss is added by both Cross-Entropy loss and geometry loss. Additionally, MoNCE [52] is incorporated to enhance versatility, leading to the following modification of the total loss function:

$$\mathcal{L} = \mathcal{L}_{\text{GAN}}(G, D) + MoNCE + \mathcal{L}_{geo}(G) + \ell\left(\hat{p}_i, p_i\right), \tag{11}$$

where $\ell(\cdot, \cdot)$ denotes the standard Cross-Entropy. $\hat{p}_i$ is the model's output probability whereas there is a contour point at pixel i, p_i is the ground truth contour point of the isometric image.

4. Experiments

4.1. Datasets

We conducted our experiments using the SPARE3D [16] dataset, which comprises 5000 pairs of three-view and isometric images. Typically, represented hidden lines in objects are not displayed in actual CAD drawings; thus, we preprocessed the images to remove all hidden lines. We concatenated three-view images along channel dimension so the input $C = 9$. The dataset was divided into 4000 training pairs and 1000 testing pairs.

4.2. Experiment Setup

We carried out all experiments using a single NVIDIA RTX 4090 GPU with 24GB of VRAM. Our model was implemented using the PyTorch framework. The learning rates were set to 0.001 for the generator and 0.003 for the discriminator. We utilized the Adam optimizer [25] with parameters $\beta_1 \coloneqq 0.001$ and $\beta_2 \coloneqq 0.9$. The evaluation metrics employed in our experiments include:

- Frechet Inception Distance (FID) is employed to measure the distance between the distribution of generated results and the distribution of real images.
- Structural Similarity Index (SSIM) is used to measure similarity of generated results and real images.
- L1 and L2 calculates the absolute difference and square of the difference between generated and real images, respectively.

4.3. Experimental Results

4.3.1. Quantitative Results

Table 1. Quantitative generation performance comparisons with baseline model.

Method	FID ↓	SSIM ↑	L1 ↓	L2 ↓
pix2pix [16] (baseline)	43.82	0.436	29.34	28.73
SPADE [33]	40.39	0.458	29.17	28.55
MoNCE [52]	41.29	0.477	27.11	26.20
DINO [43]	38.91	0.506	24.60	24.63
IsoTGAN (ours)	**21.18**	**0.723**	**15.44**	**14.37**

Table 1 shows the quantitative comparison results of the proposed IsoTGAN with the baseline model pix2pix [16], SPADE [33], MoNCE [52], and DINO [43]. Overall, our model outperforms other methods with a large margin in all metrics scores. Compared to the runner-up model, a reduction of about 45% in FID and an increase of approximately 42% in SSIM are obtained. Standard reconstruction metrics L1 and L2 also significantly decrease by about 37% and 41%, respectively, proving the efficiency of our proposed IsoTGAN when modeling the spatial and geometrical relation between each view of the object and guiding the generator to reconstruct the corresponding isometric view image. While other GAN-based methods, without Transformer blocks, are only strong at decomposing global attributes of the entire image, the cross-attention layer and GEE in the Transformer block help IsoTGAN enhance the ability to reconstruct local details of objects in isometric view. The encoder of IsoTGAN provides comprehensive spatial and geometrical relation vector for the generator to understand these relations; therefore, it is capable of generating reasonable isometric view images.

4.3.2. Qualitative Results

Fig. 3 shows the qualitative results of IsoTGAN and other paired image-to-image translation models. It can be clearly seen that our proposed method generates more similar isometric view images compared to ground truth images. In other methods' generated results, the contour lines of isometric images are discontinued or incompleted, showing that typical GAN loss function is not adequate in reconstructing images containing only contour lines from images of the same type, even though their structure looks more simple compared to natural images and other generative models are good at converting drawing edges into color images. The pix2pix [16] baseline generates isometric images that lack a lot of details and are incomprehensive. SAPDE [33] and MoNCE [52] perform better as they could capture the overall structure of objects, but small details are not reconstructed successfully. Especially, MoNCE's output of example seventh looks very similar to the groundtruth. Generated results of DINO [43] are discontinued at some parts in objects contour lines, and in example fifth, one small circle on top of the object is missing. Our proposed IsoTGAN successfully generates small details of objects, and the discontinuity in objects contour lines is minimized.

Some objects have the same front view and side view projection, other objects have top view projection similar to the rotation operation of front view or side view projection. Therefore, IsoTGAN has to distinguish between three views and discern their relations and constraints to produce comprehensive output. The modified loss function helps IsoTGAN predict contour lines better. By incorporating geometry loss, the proposed model has learned to reconstruct an isometric view image given its vertical flipped three-view image input; therefore, IsoTGAN is able to understand how projection translation affects an object's image on a 2D plane. The combination of shapes in the isometric view is also understood well by the model.

4.3.3. Ablation Studies

Ablation experiments were conducted on SPARE3D dataset to investigate the contribution of different components in the proposed method. First, we utilized only the encoder in our framework, while Transformer and GEE were replaced with conventional convolution, and the loss function contained only $\mathcal{L}_{GAN}$ and MoNCE. Then, we gradually added Transformer, GEE, $\mathcal{L}_{geo}$, and ℓ. In the next step, the encoder was discarded while other components were kept. When the encoder was not used, MCHA in Transformer was substituted by self-attention that performs attention on image patches. We also performed experiments where only Transformer and GEE were applied, and then only $\mathcal{L}_{geo}$ and ℓ were added.

Overall, it can be seen clearly that the encoder plays a crucial role. As Table 2 illustrates, without the encoder, despite keeping all other components, the generation metrics are significantly worse compared to when the encoder is equipped. Only utilizing the encoder yields better FID and SSIM than discarding it while keeping all other four components. Without the encoder, IsoTGAN becomes a conventional GAN model, so the model cannot understand the spatial and geometrical relation between each orthogonal view of the object. The Transformer with MHCA layers also contributes immensely to the effectiveness of the proposed method by incorporating a spatial and geometrical vector to selectively distribute information from the entire input to appropriate regions, enhancing the generation of local parts of objects. The GEE attention mechanism in conjunction with Transformer provides extra information about noise contexts for a more accurate generation of isometric view image. Additionally, the modification of loss function gives more geometry constraint for the model to ensure that the generated output follows objects' physical rules.

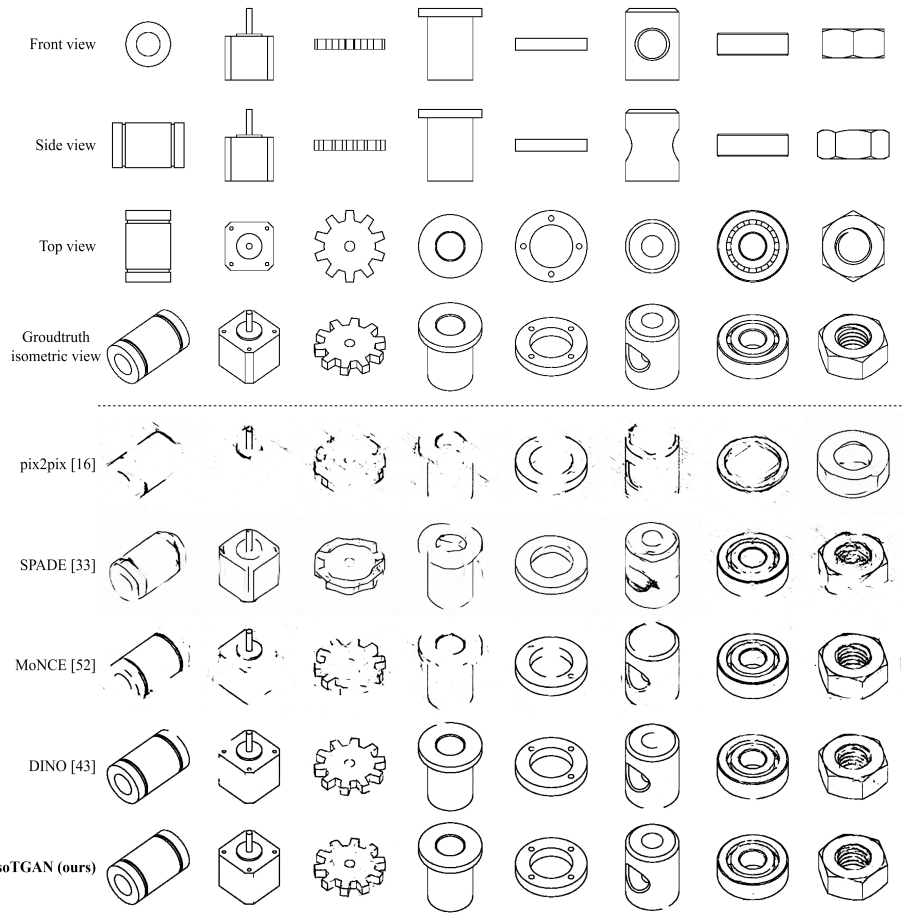

Figure 3. Qualitative comparisons of isometric view generation performance with baseline model.

Table 2. Ablation studies on SPARE3D dataset of the proposed method.

Method	Encoder	Transformer	GEE	$\mathcal{L}_{geo}$	ℓ	FID ↓	SSIM ↑	L1 ↓	L2 ↓
(a)	✓					28.18	0.527	21.33	21.05
(b)		✓				31.15	0.508	23.86	22.56
(c)		✓	✓			30.12	0.512	23.77	22.66
(d)		✓	✓	✓	✓	29.47	0.520	21.02	20.53
(e)				✓	✓	31.21	0.480	24.50	23.21
(f)		✓		✓	✓	29.30	0.498	21.10	20.77
(g)	✓			✓	✓	24.20	0.605	17.89	16.37
(h)	✓	✓		✓	✓	21.82	0.695	15.66	14.49
(i)	✓	✓				22.89	0.587	16.70	15.52
(j)	✓	✓	✓			21.98	0.686	15.85	14.43
(k)	✓	✓	✓	✓		21.37	0.711	15.54	14.62
(l)	✓	✓	✓	✓	✓	**21.18**	**0.723**	**15.44**	**14.37**

4.3.4. Impact of Standard Deviation in GEE

In this section, we explore how the standard deviation σ in the Gaussian function $\mathbf{G}(x) = \exp(-\frac{x^2}{2\sigma^2})$ influences the performance result of IsoTGAN in generating isometric view image from three orthographic view contour drawings. The findings are presented in Table 3. It is observed that as σ increases, the network performance first improves and then decreases. The optimal performance is achieved when σ is set at 4. This behavior makes sense because a very high

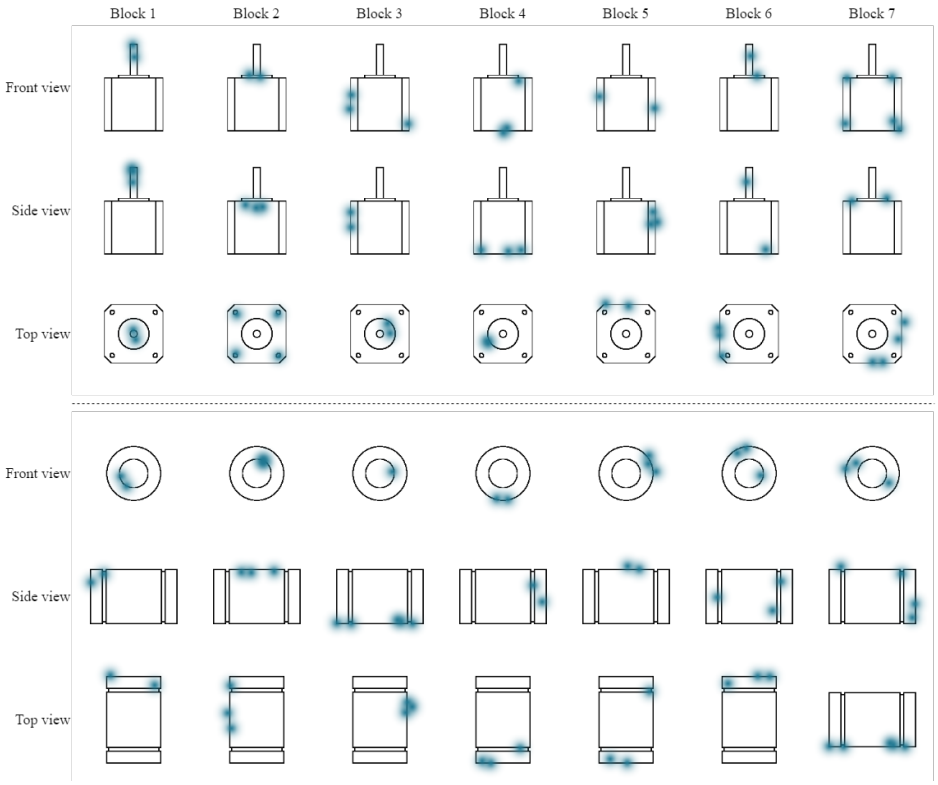

Figure 4. Attention heatmap generated by attention weights from Transformer blocks in IsoTGAN's generator.

Table 3. Generation result of IsoTGAN on SPARE3D dataset with different standard deviation σ in GEE.

σ	FID ↓	SSIM ↑	L1 ↓	L2 ↓
1	22.56	0.648	16.13	15.31
2	22.51	0.674	15.98	15.22
4	**21.18**	**0.723**	**15.44**	**14.37**
6	21.89	0.705	16.07	15.16

variance can reduce the differences in attention activations across channel and spatial dimension, which interferes with the effective reduction of global noise contexts. Conversely, a too low variance might limit the significance of other important features and incorrectly emphasize noise contexts.

4.3.5. Attention Visualization

To better understand the interpretability of the IsoTGAN's generator, we analyze the attention matrices generated by the Transformer blocks. We examine the attention weights output by the MHCA layer. Each element (i, j) in these matrices shows the degree of attention that token i gives to token j. Since the model employs multi-head attention, multiple attention matrices are produced, one per head. For clarity, we calculate the average of these weights across all heads and target tokens for each block. Considering an input image size of 256×256, this results in an average attention vector with dimensions $w \times h$ (16×16). Each entry j-th in this vector indicates the average attention token j receives. By overlaying the attention heatmap on the input images, shown in Fig. 4, it can be seen clearly that each block pays attention to a specific part inside three-view images. It focuses on contour lines of the shape and by that, strengthens the generation of local details of the object.

5. Conclusion

This paper has proposed a novel IsoTGAN framework for effectively automatic isometric view image generation from three orthographic views contour drawings. It employs an encoder to capture and convert the spatial and geometric relation between each view of the object into a vector. This vector serves as additional input to the generator. The Transformer with cross-attention equipped in IsoTGAN's generator enables long-range interactions and constraint guidance to promote comprehensive isometric view image generation. An attention mechanism called GEE which utilizes Gaussian function and a geometry-constraint loss function are also proposed to refine the generation results. Extensive experimental results demonstrate that our proposed method shows a strong capability in isometric view image generation, achieving state-of-the-art performance. Physical constraints for better refining the reconstruction of difficult features of objects will be considered.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 4

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[3] Yu Cheng, Zhe Gan, Yitong Li, Jingjing Liu, and Jianfeng Gao. Sequential attention gan for interactive image editing. In *Proceedings of the 28th ACM international conference on multimedia*, pages 4383–4391, 2020. 2

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[5] Giannis Daras, Augustus Odena, Han Zhang, and Alexandros G Dimakis. Your local gan: Designing two dimensional local attention mechanisms for generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14531–14539, 2020. 2

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 1

[8] 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. 2

[9] Hajar Emami, Majid Moradi Aliabadi, Ming Dong, and Ratna Babu Chinnam. Spa-gan: Spatial attention gan for image-to-image translation. *IEEE Transactions on Multimedia*, 23:391–401, 2020. 2

[10] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9935–9946, 2023. 2

[11] Yaroslav Ganin, Sergey Bartunov, Yujia Li, Ethan Keller, and Stefano Saliceti. Computer-aided design as language. *Advances in Neural Information Processing Systems*, 34:5885–5897, 2021. 1, 3

[12] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 1

[13] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10021–10030, 2023. 2

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[15] Haoxiang Guo, Shilin Liu, Hao Pan, Yang Liu, Xin Tong, and Baining Guo. Complexgen: Cad reconstruction by b-rep chain complex generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–18, 2022. 1, 3

[16] Wenyu Han, Siyuan Xiang, Chenhui Liu, Ruoyu Wang, and Chen Feng. Spare3d: A dataset for spatial reasoning on three-view line drawings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14690–14699, 2020. 2, 3, 5, 6

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[18] Lukas Höllein, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, Matthias Nießner, et al. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5043–5052, 2024. 1

[19] Wentao Hu, Jia Zheng, Zixin Zhang, Xiaojun Yuan, Jian Yin, and Zihan Zhou. Plankassembly: Robust 3d reconstruction from three orthographic views with learnt shape programs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18495–18505, 2023. 1, 3

[20] Drew A Hudson and Larry Zitnick. Generative adversarial transformers. In *International conference on machine learning*, pages 4487–4499. PMLR, 2021. 2

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2, 3

[22] Pradeep Kumar Jayaraman, Joseph George Lambourne, Nishkrit Desai, Karl Willis, Aditya Sanghi, and Nigel JW Morris. Solidgen: An autoregressive model for direct b-rep synthesis. *Transactions on Machine Learning Research*, 2022. 1, 3

[23] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34:14745–14758, 2021. 2

[24] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 4

[27] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10551–10560, 2019. 2

[28] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[29] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations*. 2

[30] 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*, pages 10012–10022, 2021. 2

[31] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[32] Wamiq Para, Shariq Bhat, Paul Guerrero, Tom Kelly, Niloy Mitra, Leonidas J Guibas, and Peter Wonka. Sketchgen: Generating constrained cad sketches. *Advances in Neural Information Processing Systems*, 34:5077–5088, 2021. 1, 3

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 4, 6

[34] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[35] Dongsheng Ruan, Daiyin Wang, Yuan Zheng, Nenggan Zheng, and Min Zheng. Gaussian context transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15129–15138, 2021. 4

[36] Ari Seff, Wenda Zhou, Nick Richardson, and Ryan P Adams. Vitruvion: A generative model of parametric cad sketches. In *International Conference on Learning Representations*, 2021. 1, 3

[37] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1, 2

[38] Shikun Sun, Longhui Wei, Junliang Xing, Jia Jia, and Qi Tian. Sddm: score-decomposed diffusion models on manifolds for unpaired image-to-image translation. In *International Conference on Machine Learning*, pages 33115–33134. PMLR, 2023. 2

[39] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *International Conference on Learning Representations*, 2016. 2

[40] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE transactions on neural networks and learning systems*, 34(4):1972–1987, 2021. 2

[41] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 1

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[43] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Dino: A conditional energy-based gan for domain translation. In *International Conference on Learning Representations*. 6

[44] Karl DD Willis, Pradeep Kumar Jayaraman, Joseph G Lambourne, Hang Chu, and Yewen Pu. Engineering sketch generation for computer-aided design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2105–2114, 2021. 1, 3

[45] 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. 4

[46] Rundi Wu, Chang Xiao, and Changxi Zheng. Deepcad: A deep generative network for computer-aided design models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6772–6782, 2021. 1, 3

[47] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13095–13105, 2023. 2

[48] Sihan Xu, Ziqiao Ma, Yidong Huang, Honglak Lee, and Joyce Chai. Cyclenet: Rethinking cycle consistency in text-guided diffusion for image manipulation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[49] Xiang Xu, Karl DD Willis, Joseph G Lambourne, Chin-Yi Cheng, Pradeep Kumar Jayaraman, and Yasutaka Furukawa. Skexgen: Autoregressive generation of cad construction sequences with disentangled codebooks. In *International Conference on Machine Learning*, pages 24698–24724. PMLR, 2022. 1, 3

[50] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7079–7088, 2024. 2

[51] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[52] Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Rongliang Wu, and Shijian Lu. Modulated contrast for versatile image synthesis. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, pages 18280–18290, 2022. 5, 6

[53] 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*, pages 11304–11314, 2022. 2

[54] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 2

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2