

# SWINV2-IMAGEN: HIERARCHICAL VISION TRANSFORMER DIFFUSION MODELS FOR TEXT-TO-IMAGE GENERATION

**Ruijun Li**

Auckland University of Technology  
Auckland, New Zealand  
zjc0233@autuni.ac.nz

**Weihua Li**

Auckland University of Technology  
Auckland, New Zealand  
weihua.li@aut.ac.nz

**Yi Yang**

Hefei University of Technology  
Hefei, China  
yyang@hfut.edu.cn

**Hanyu Wei**

University of Tasmania  
Hobart, Australia  
hanyu.wei@utas.edu.au

**Jianhua Jiang**

Jilin University of Finance and Economics  
Changchun, China  
jjh@jlu.edu.cn

**Quan Bai**

University of Tasmania  
Hobart, Australia  
quan.bai@utas.edu.au

## ABSTRACT

Recently, diffusion models have been proven to perform remarkably well in text-to-image synthesis tasks in a number of studies, immediately presenting new study opportunities for image generation. Google’s Imagen follows this research trend and outperforms DALLE2 as the best model for text-to-image generation. However, Imagen merely uses a T5 language model for text processing, which cannot ensure learning the semantic information of the text. Furthermore, the Efficient UNet leveraged by Imagen is not the best choice in image processing. To address these issues, we propose the Swinv2-Imagen, a novel text-to-image diffusion model based on a Hierarchical Visual Transformer and a Scene Graph incorporating a semantic layout. In the proposed model, the feature vectors of entities and relationships are extracted and involved in the diffusion model, effectively improving the quality of generated images. On top of that, we also introduce a Swin-Transformer-based UNet architecture, called Swinv2-Unet, which can address the problems stemming from the CNN convolution operations. Extensive experiments are conducted to evaluate the performance of the proposed model by using three real-world datasets, i.e., MSCOCO, CUB and MM-CelebA-HQ. The experimental results show that the proposed Swinv2-Imagen model outperforms several popular state-of-the-art methods.

**Keywords** Text-to-image synthesis · Diffusion models · Scene Graph · Graph Neural Network · UNet

## 1 Introduction

People tend to describe rich and detailed pictures of scenes through language, and the ability to generate images from these descriptions can facilitate creative applications in various life contexts, including art design and multimedia content creation [1, 2]. This fact has inspired researchers to design models of text to image comparative learning to assist people with making decisions quickly in specific scenarios, such as presentation and advertising design [3, 4]. In recent years, diffusion models have attracted the attention of many scholars due to their promising performance in image generation. Within this framework, DALL-E 2 [5] and Imagen [6] have become successful generative models for image generation.

Imagen is currently one of the greatest image generation models. Its most significant distinguishing feature is its immensity, which is reflected, in particular, by its utilisation of a large text encoder, i.e., T5 [7]. T5 is pre-trained on a sizable plain text corpus. It turns out that T5 is very effective for enhancing image fidelity and image-text alignment [6]. However, T5 is Transformer-based, and using T5 alone to obtain text embeddings cannot guarantee that the model learns important text features, such as semantic layout. Besides visual elements, the semantic layout is recognised as an important factor in guiding text-to-image synthesis [8]. Our experimental results provide evidence for this claim.

Furthermore, very few research works are dedicated to addressing the UNet issue of Imagen. The diffusion model of Imagen relies on the Efficient-UNet, which suffers from the limitations of CNN convolution operations. CNN are good at extracting the low-level features and elements of visual structure, such as colour, contour, texture and shape [9]. However, CNN just focuses on the consistency of these low-level features under transformations, such as translation and rotation, i.e., Translation Invariance [10] and Rotation Invariance [11]. It fails to extract features efficiently in terms of global and layout. In other words, the convolutional filter detects key points, object boundaries and other basic units that constitute the visual elements. The elements should be transformed simultaneously under translation and rotation transformations. Therefore, CNN is one of the choices for addressing the consistency problem, though it cannot represent high-level visual semantic information regarding the connections between objects. For text-to-image synthesis tasks, it is significant to consider how such basic elements can be organised as objects and how they can be constructed into a scene using spatial location relationships. The Transformer is more natural and efficient than CNN in processing this high-level semantic information. This is mainly because the Transformer focuses on local information and has a diffusion mechanism to find expressions from the local to global layout [12, 13].

To solve the aforementioned drawbacks of Imagen, in this paper, we propose a diffusion text-to-image generation model called Swinv2-Imagen. The proposed model is based on a Hierarchical Visual Transformer and Scene Graph incorporating layout information. Specifically, the semantic layout is generated via semantic scene graphs, enabling Swinv2-Imagen to parse the layout information in the text description effectively. In this paper, we adopt Stanford Scene Graph Parser [14] to obtain the Scene Graph from the text. Subsequently, the entity and relationship embeddings are extracted using a frozen Graph Convolution Network (GCN) [14]. The image generation process appears conditional on text, object and relationship embeddings. The layout representation with global semantic information ensures the realism of the generated images. In addition, the diffusion models are developed based on Swinv2-UNet, a variant of Swin Transformer v2 [15], which implements a full Transformer text to image generation model for a more accurate understanding of text semantics. Finally, we evaluate our model on the MSCOCO, CUB and MM-CelebA-HQ datasets. The results show that the proposed model outperforms the current best generative model, Imagen, on MSCOCO. The ablation experiments reveal that the addition of semantic layouts is effective in improving the semantic understanding of the model.

The key contributions of this paper are summarised below.

1. We leverage scene graphs to extract entity and relational embeddings to improve local and layout information representation of text for a more accurate understanding of the text and realistic image generation;
2. We propose Swinv2-UNet as a novel diffusion model architecture to implement a full Transformer text generation image model;
3. We combine the scene graph with the Transformer to improve the effectiveness of the model;
4. We achieve a new state-of-the-art FID result, (FID=7.21), on the MSCOCO dataset compared to the latest generative models. Better results are also obtained on both the CUB (FID=9.78) and MM CelebA-HQ (FID=10.31).

The rest of the paper is organised as follows. In Section 2, related works are reviewed. In Section 3, we elaborate on the proposed Swinv2-Imagen model. In Section 4, we conduct extensive experiments to evaluate the performance of the proposed model and perform an ablation study to evaluate the contributions of each key component of our model. Finally, we conclude this paper in Section 5, and discuss future research directions.

## 2 Related work

### 2.1 Diffusion models

Text-to-Image synthesis is a typical application of multimodal and cross-modal comparative learning. In the field of image generation, most models mainly fall into two categories, i.e., the GAN-based generation models [16, 17, 18, 19, 20, 21] and the diffusion-based models [22, 23, 24, 25, 26]. The former has been developed over the last few years and widely used in many scenarios, such as medical and image restoration. The latter has demonstrated outstanding performance over the GAN models, acknowledged as state-of-the-art deep generative models [6, 27, 28].

Diffusion models and GAN generative models are essentially comparable, both being a process of gradually removing noise. However, in contrast to GAN, the diffusion models do not suffer from training instability and model collapse. The diffusion model transforms the data distribution into random noise and reconstructs data samples with the same distribution [6, 29]. The diffusion model demonstrates outstanding performance for a number of tasks, such as multimodal modelling. Many contemporary text-to-image synthesis models, e.g., DALL-E 2 [5], Imagen [6] and GLID [24], are constructed based on the diffusion model. They cascade multiple diffusion models to improve the image

generation quality step by step. DALL-E 2 uses a priori diffusion model and CLIP Latents to process the text. In contrast, Imagen discards the priori model and replaces it with a large pre-trained text encoder, i.e., T5. Although the T5 model leveraged in the Imagen model improves the understanding of the text, it does not ensure that the model understands the semantic layout of the text, especially in complex sentences containing multiple objects and relationships. As a result, the model will not be able to reproduce some entities or will lose some entity relationships. Therefore, we attempt to model the global semantic layout by adding a scene graph in the text processing. Furthermore, Imagen builds its diffusion model based on Efficient-Unet. Efficient-Unet is not the best choice in image generation tasks, because it contains multiple CNN blocks and leads to a limited view within the CNN kernel window.

## 2.2 Scene graph and Graph representation learning

A sentence's nature is a linear data structure, where one word follows another [14]. Usually, when a sentence is complex with multiple objects, it is time-consuming to analyse the sentence directly, and the accuracy of the text-image alignment is not guaranteed. Complex sentences often incorporate rich scene information. Mapping this information into a scene graph can provide an intuitive understanding of the relationships between objects in a sentence [30]. Previous studies reveal that the performance of multimodal models, such as text-to-image synthesis, is significantly dependent on mining visual relationships [31]. Scene graphs can provide a high level of understanding regarding scene information [14]. Therefore, the scene graph is recognised as a useful representation of images and text. Specifically, each node in a scene graph represents an object, such as a person or an event, and each object has multiple attributes, such as shape. The relationships between objects are denoted by the edges between nodes, which can be an action or a position [32]. Recently, the scene graphs have been used extensively for tasks such as text-based image retrieval [33, 34], semantic segmentation [35, 36], visual question answering [37], image captioning [38, 39, 40, 41] and image generation [14, 30, 42, 43].

In addition, there is no way for an image generation model to manipulate graph-like data such as scene graphs directly, so scene graphs are usually used in conjunction with graph representation learning [44]. The main objective of graph representation learning is to extract node and edge contexts from the scene graph and map them to a set of embeddings. Graph representation learning methods can currently be classified into two types, i.e., machine learning based on Random-Walk and deep learning Graph Convolution-based methods [44]. Node2vec [45] is a typical representative model of the former. It is based on SkipGram [46] theory to learn the embedding of nodes on a graph and optimises the sampling method. It is proposed in related studies [47] that two sampling methods, Breadth-First Search (BFS) and Depth First Search (DFS), are mainly included when sampling neighbouring nodes in a graph. BFS requires that each sampled node is a direct neighbour of that node. This sampling method results in a graph representation that is more concerned with local information. In contrast, DFS, where each node is sampled to increase the distance to the initial node as much as possible, produces a graph representation that focuses more on global information. Random-Walk-based representation learning [48] is composed of multiple stages, each with different optimisation goals, which is a typical non-end-to-end model. Graph convolution-based methods, e.g. graph convolution neural networks [14], are able to learn both node feature information and structural information via an end-to-end way. It focuses on both local information and global structural features. Graph convolution is extremely applicable to nodes and graphs of topology. It is currently the best choice for graph data learning.

## 2.3 UNet

UNet is an encoder-decoder architecture, which is scalable in structure [49]. The encoding stage of the UNet consists of four downsamples. Symmetrically, its decoding stage is also upsampled four times, restoring the result of the encoder to the resolution of the original image. In contrast to Fully Convolutional Networks (FCN) [50], UNet upsamples four times and uses a jump connection in the encoder and decoder of the corresponding convolution blocks. The jump connection ensures that the final recovered feature map incorporates more low-level semantic features and features at different scales are well fused, allowing for multi-scale prediction. In addition, the four times up-sampling also allows the segmentation map to recover information such as edges more finely. However, UNet also has some shortcomings. For example, UNet++ [51, 52] argues that it is inappropriate to directly combine the shallow features from the encoder with the deeper features from the decoder in UNet. Direct fusion would potentially lead to semantic gaps. Furthermore, UNet 3+ [53] maximises the scope of model information fusion and circulation. Each decoder layer in the UNet 3+ fuses small-scale and same-scale feature maps from the encoder with larger-scale feature maps from the decoder, which capture both fine-grained and coarse-grained semantics at full scale.

Many researchers develop a set of UNet variants by improving and optimising the original UNet. For example, ResUNet [54] and DenseUNet [55] are inspired by Residual and Dense connections, respectively; each sub-module of the U-Net is replaced with a form having a Residual connection and a Dense connection. There are variants, e.g., MultiResU-Net [56] and R2 U-Net [57]. All of these models are constructed using multiple convolutional blocks. With the advent

of the Transformer, researchers begin to develop the UNet base on the Transformer, such as Swin-UNet [58]. While Swin-UNet mitigates the limitations of CNN convolutional operations, it is likely to suffer from training instability due to the use of the Swin-Transformer block. Swin-Transformer v2 [15] is an improvement on Swin-Transformer, which is effective in avoiding training instability and is easier to scale.

Inspired by these research works, we propose a Swinv2-Imagen model that leverages scene graphs as auxiliary modules to help the model understand the text semantics more comprehensively. In addition, Swinv2-Unet is applied to build the diffusion models so that our model is based on the full Transformer implementation. As a result, it effectively addresses the limitations of CNN convolution operations, theoretically enabling the synthesis of images better than baselines.

### 3 Swinv2-Imagen

The overall architecture of the proposed Swinv2-Imagen model is shown in Figure 1. It takes text descriptions as input and uses scene graphs to guide downstream image generation more accurately and efficiently. The upstream comprises two sub-modules: the text encoder, which maps the text input to a text embedding sequence and the scene graph generator sub-module. The downstream consists of a set of conditional diffusion models, integrating the intermediate embeddings in the upstream and generating high-fidelity images step by step. The scene graph generator includes a Scene Graph parser and a frozen Graph Neural Network, which aims to represent objects and relationships in a text with a graph structure.

The input of the model is a text-picture pair. Firstly, the text is encoded by T5 tokenizers and input to the embedding layer to get the initial text embedding. Next, it goes through the T5 encoder (n-layer T5 Block) to obtain Text Embeddings. Meanwhile, the scene graph parser extracts the scene graph from the text, and the frozen GCN (m-layer Graph Triple Convolution) obtains the corresponding Object and Relation embeddings. Finally, the Conditional embeddings are obtained by concatenating the Text embeddings, Object embeddings and Relation embeddings in this order. The Conditional embeddings are used as conditional input for subsequent super-resolution image generation. In the following subsections, we describe the main components of Swinv2-Imagen in detail.

#### 3.1 Pre-trained frozen text encoders

It is widely acknowledged that a robust semantic text encoder is essential for text-to-image synthesis models and plays a crucial role in analysing the complexity and composition of textual input [6]. Previously, language models were mainly built on RNN architectures. However, since the emergence of the Transformer, a number of transformer-based pre-trained language models have been developed, such as GPT [59, 60, 61], BERT [62] and T5 [7]. The traditional Imagen model is compared against popular text encoders, BERT, CLIP and T5-XXX, by freezing parameters. The existing research results prove the promising performance of T5-XXX in terms of both image-text alignment and image fidelity [6]. Therefore, we adopt the T5 large language model for text encoding in the proposed model.

#### 3.2 Scene Graph and Frozen Graph Convolutional Neural Network

This sub-module aims to extract entity relationships from the text to enhance the text understanding of the model. We adopt a Scene Graph parser to represent text as a scene graph, followed by a frozen GCN to extract the entity and relational embeddings for the image generation with diffusion models. Scene graphs with graph neural networks [14] have been proven to be highly effective in extracting object relationships from the text. As shown in Figure 2, Swinv2-Imagen constructs a scene graph for the text and is followed by a graph neural network to extract the entities and relationships from the scene graph. For any given text description, the corresponding scene graph is represented as  $(O, E)$ , where  $O = (o_1, o_2, o_3, \dots, o_n)$  denotes each entity object in the sentence i.e., subject and object, and  $E$  is a collection of edges of the form  $(o_i, r, o_j)$ , where  $r \in \mathcal{R}$ ,  $\mathcal{R}$  refers to a collection of relationships.  $E$  denotes the relationship between the two entities, e.g., action and position. In the end, object and relation embeddings are constructed, which are used to assist the T5 model in analysing and understanding the text more comprehensively.

The input to the graph convolution is a scene graph, having each node and edge represented as a vector with dimension  $D^{in}$ . In the graph convolution model, these vectors are newly computed as new vectors with dimension  $D^{out}$  for each node and edge. In addition, two functions,  $g_s$  and  $g_o$ , are used to calculate the entity features vectors of output. They use a triplet of an edge,  $(\mathbf{v}_i, \mathbf{v}_r, \mathbf{v}_j)$  as input and obtain two vectors representing the subject and object, respectively. Function,  $g_p$ , is used to calculate the relation features vectors of output. It uses an edge vector and two associated object vectors as inputs and outputs a predicted relationship  $r$ . In the scene graph, given an edge  $\mathbf{v}_r$ , the two associated objects,  $\mathbf{v}_i$  and  $\mathbf{v}_j$ , can be determined. Thus, the output relationship vector  $\mathbf{v}'_r$  can be simply expressed as:

$$\mathbf{v}'_r = g_p(\mathbf{v}_i, \mathbf{v}_r, \mathbf{v}_j) : (o_i, r, o_j) \in E \quad (1)$$

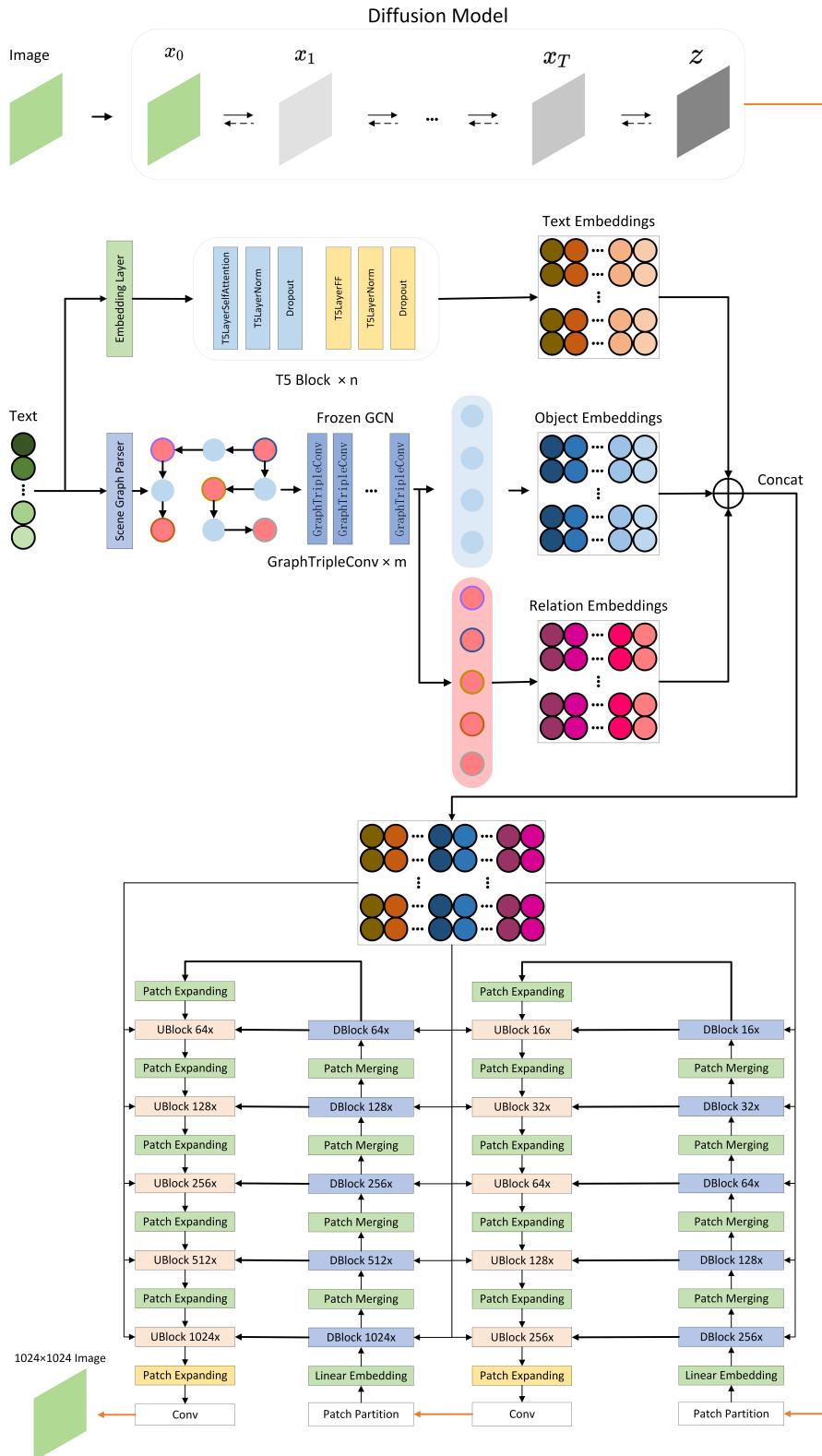


Figure 1: The overall architecture of Swinv2-Imagen.

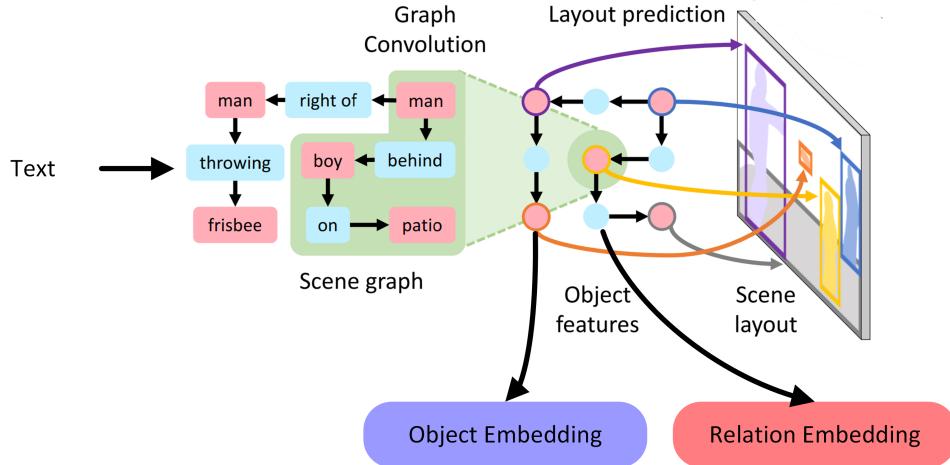


Figure 2: The process of Object and Relation embeddings extraction.

In contrast, the calculation of output object vectors  $\mathbf{v}'_i$ , is more complicated. Generally, an object is associated with two or more relations. Therefore, the output vector of an entity  $o_i$  is calculated by considering all the vectors directly connected to the object, i.e.,  $\mathbf{v}_j$ , and the corresponding relationship vectors,  $\mathbf{v}_r$ . The function  $g_s$  in Equation 2 is used to compute all vectors starting at node  $o_i$  and function  $g_o$  in Equation 3 is used to compute all vectors ending at node  $o_i$ . Afterwards, these vectors are collected into  $V_i^s$  and  $V_i^o$ .

$$V_i^s = \{g_s(\mathbf{v}_i, \mathbf{v}_r, \mathbf{v}_j) : (o_i, r, o_j) \in E\} \quad (2)$$

$$V_i^o = \{g_o(\mathbf{v}_j, \mathbf{v}_r, \mathbf{v}_i) : (o_j, r, o_i) \in E\} \quad (3)$$

Then the output vector  $\mathbf{v}'_i$  for the entity  $o_i$  is expressed as.

$$\mathbf{v}'_i = h(V_i^s \cup V_i^o), \quad (4)$$

where  $h$  denotes a function that pools all vectors in  $\mathbf{V}_i^s$  and  $\mathbf{V}_i^o$  to a single output vector [14].

### 3.3 Image generator

The image generator is composed of three diffusion models located downstream. In the diffusion model, a hidden variable  $z$  is obtained by adding noise to the image for  $T$  times. After forward and backward diffusion, a basic  $64 * 64$  image can be learned. The basic image is input to the first Swinv2-Unet to generate a  $256 * 256$  image. Finally, the image goes to the second Swinv2-Unet super-resolution generation, producing a  $1024 * 1024$  high-definition image. In contrast to Imagen, we focus on improving super-resolution diffusion models. We introduce a new UNet variant to our super-resolution diffusion model, called Swinv2-UNet. The Swin Transformer Block is replaced with the Swin Transformer v2 Block based on the original Swin-UNet [58], the complete structure of which is shown in Figure 3.

The DBlock and UBlock of Swinv2-UNet consist of the Swin Transformer v2 block, which comprises LayerNorm (LN) layers, multi-headed self-attention modules, Residual connections and a 2-layer MLP with GELU non-linearity. The Swin Transformer v2 block could be represented as:

$$\hat{z}^{l+1} = LN(Attn(z^l)) + z^l \quad (5)$$

$$z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1}, \quad (6)$$

where  $z^l$  and  $z^{l+1}$  denote the input and output of the Transformer v2 block, respectively.  $\hat{z}^{l+1}$  is an intermediate variable.  $+$  denotes the residual connection or skip connection.

The attention of Swin-v2 is expressed as:

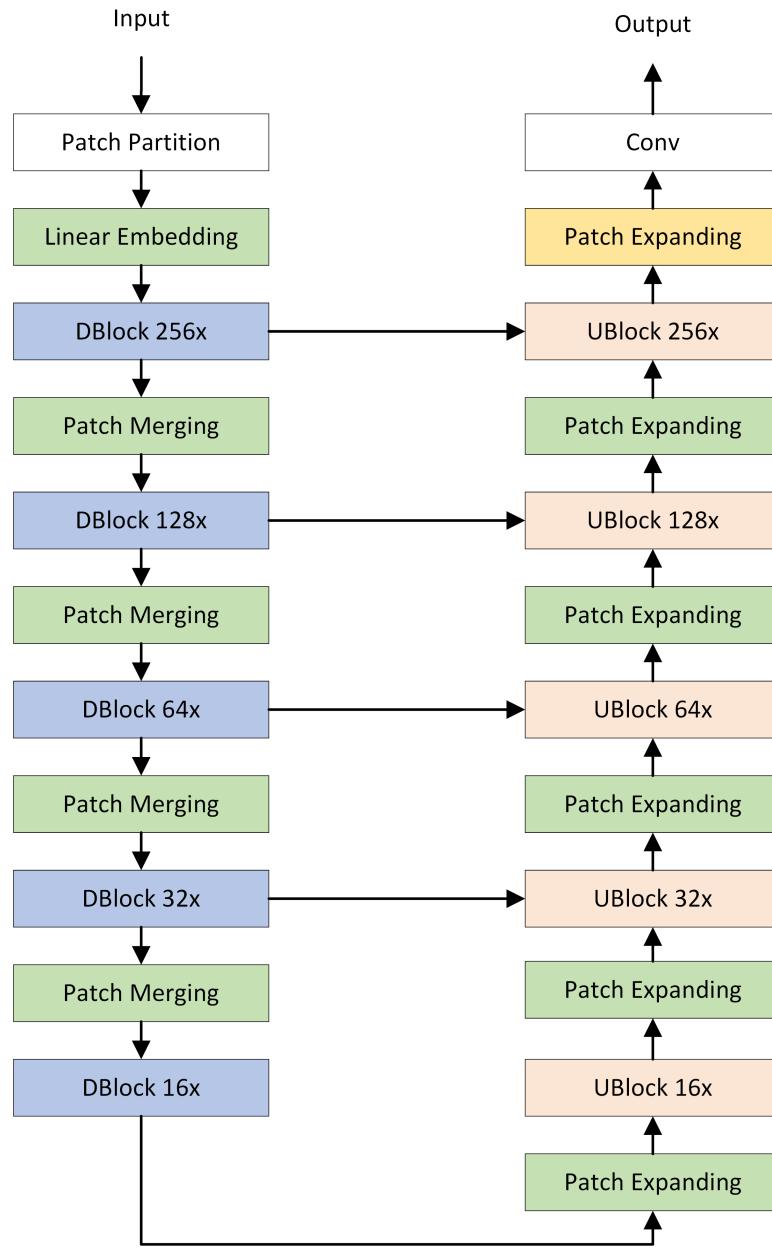


Figure 3: The UNet architecture of the super-resolution submodule.

$$Attn(Q, K, V) = SoftMax \left( \frac{Cosine(Q, K)}{\tau} + B \right) \quad , \quad (7)$$

where  $Q$ ,  $K$ , and  $V$  denote the matrix of query, key and value, respectively.  $\text{Cosine}()$  refers to a function that calculates the scaled cosine similarity of  $Q$  and  $K$ .  $\tau$  denotes a learnable scalar, usually greater than 0.01.  $B$  is a matrix of relative position bias.

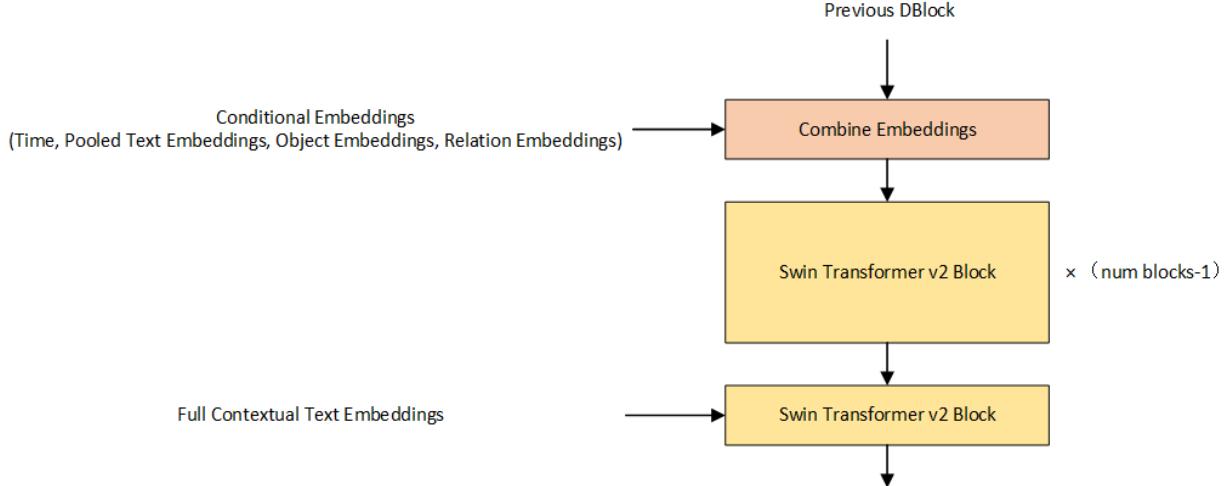


Figure 4: Swinv2-Unet DBlock

Figure 4 illustrates the network structure of the Swinv2-Unet DBlock, which is the basic component of the downsampling path under the encoding-decoding structure of UNet. Firstly, the DBlock combines the pooled text embeddings, object embeddings and relation embeddings into a conditional embedding input to the cross-attention layer. Next, it is followed by the Swinv2-Transformer v2 blocks for (num\_block-1) times feature extraction.

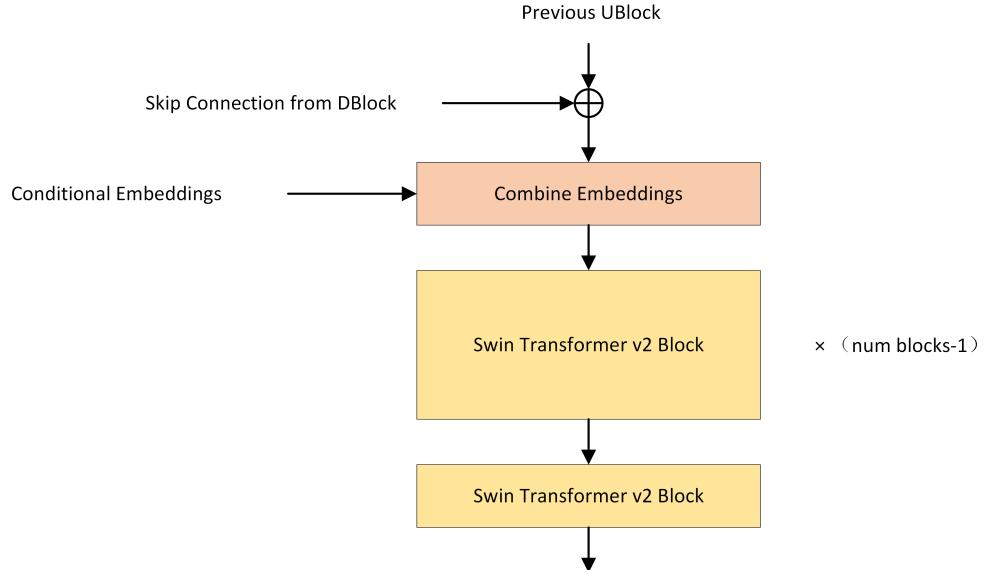


Figure 5: Swinv2-Unet UB1cok

Figure 5 shows the network structure of the Swinv2-UNet UBlock, which is the basic component of the upsampling path on the UNet encoder-decoder. The inputs to the UBlock include the output of the previous UBlock layer and the corresponding DBlock. The DBlock and UBlock are connected using skip connections [52]. Subsequently, the

conditional embedding inputs are also introduced to the cross-attention layer. Similar to the DBBlock, this layer is followed by the Swinv2-Transformer v2 blocks for (num\_block-1) times feature extraction.

The encoder is presented as a stacking of DBlocks and Patch Merging. In the encoder, images are fed into five consecutive DBlocks for learning, where the feature dimension and resolution are maintained. Meanwhile, Patch Merging performs Token Merging and increases the feature dimension to four times the original dimension. Next, we apply a linear layer to standardise the feature dimension to twice the original dimension. The process is repeated four times in the encoder.

Similar to UNet, skip connections are used to integrate the multi-scale features of the encoder with the upsampled features. We connect shallow and deep features to minimise the loss of spatial information due to downsampling. The next layer is a linear layer where the dimensionality of the connected features is kept the same dimensionality as that of the upsampled features.

The decoder is a symmetric decoder corresponding to the encoder. For this reason, unlike the Patch Merging used in the encoder, we use Patch Expanding in the decoder to upsample the extracted features. The Patch Expanding reshapes the feature maps of adjacent dimensions into a higher resolution feature map ( $2 \times$  upsampling) and accordingly reduces the number of feature dimensions to half the original dimensionality.

## 4 Experiments

In this section, we perform extensive experiments to evaluate the proposed Swinv2-Imagen model by using the MSCOCO, CUB and Multi-modalCelebA-HQ (MM CelebA-HQ) datasets. Firstly, a brief description of the datasets is given. Secondly, we compare the performance of the Swinv2-Imagen model with state-of-the-art generative models. Finally, we conduct ablation experiments to compare the contributions of each module.

### 4.1 Setup

#### 4.1.1 Datasets

The Microsoft Common Objects in Context 2014 (MS COCO-2014) [63], the Caltech-UCSD Birds-200-2011 (CUB-200-2011) [64] and MM CelebA-HQ [19] datasets are utilised in this research. Three datasets cover both simple (CUB) and complex (MSCOCO) datasets. The use of the MM CelebA-HQ dataset is mainly because most generative models such as Cogview and Craiyon, produce distorted and less realistic faces.

- MSCOCO <sup>1</sup> was released in 2014. It is a collection of 164K images, which have been partitioned into the training set (82K), validation set (41K) and testing set (41K). The dataset is complex because most of the images possess at least two objects.
- CUB <sup>2</sup> contains 12K bird images of 200 subcategories, 6K for training and 6K for testing. It is a simple dataset, having only one object per image.
- MM CelebA-HQ <sup>3</sup> is a large-scale face image dataset. It is a collection of 30K high-resolution face images. The dataset is used widely to train and evaluate algorithms for text-image generation and text-guided image manipulation.

#### 4.1.2 Evaluation metrics

We adopt Fréchet Inception Distance (FID) [65] and Inception Score (IS) [8] as evaluation metrics. Both are acknowledged as standard metrics for evaluating the image generation model. Specifically, IS examines both the clarity and diversity of the resulting images. The higher the IS, the better the quality of the generated images. FID calculates the difference between the generated image and the original image. The smaller the difference, the better the generated image is.

#### 4.1.3 Baselines

- PCCM-GAN [66] (Photographic Text-to-Image Generation with Pyramid Contrastive Consistency Model) is a typical multi-stage generative model. Its main innovations include the introduction of stack attention and the

<sup>1</sup><https://paperswithcode.com/dataset/coco>

<sup>2</sup><https://paperswithcode.com/dataset/cub-200-2011>

<sup>3</sup><https://paperswithcode.com/dataset/multi-modal-celeba-hq-1>

lateral connection of the PCCM. The two modules enhance the generative model to simultaneously extract semantic information from both global and local aspects, ensuring that the generated images are semantically consistent.

- DM-GAN [16] (Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis) is also a multi-stage generative model. It uses a memory module and a gate mechanism in the image refinement process. The aim is to re-extract important information from the image as an aid when the generated image is not as good as expected.
- SDGAN [67] (Semantics Disentangling for Text-to-Image Generation) consists of two modules, i.e., Siamese and semantic conditioned batch normalization, to extract high-level and low-level semantic features respectively.
- CogView [68] is based on the Transformer architecture. Its input is a text-image pair. The text and image features are combined and passed to the GPT language model for autoregressive training.
- GLIDE [24] is a large-scale image generation model based on diffusion models with 3.5 billion model parameters.
- DALL-E 2 [5] is also based on diffusion models. One of its highlights is the use of a priori model built on the diffusion models. Its inputs are also text and corresponding images. The text is first passed through the prior model and a corresponding image vector is generated. The image is passed through the CLIP module which also generates an image vector to supervise the result of the prior model.
- LAFITE [69] is a variant of generative adversarial networks. It leverages the CLIP model to extract features from images and text, ensuring text-image consistency.
- Imagen [6] is a text-to-image synthesising model based on the diffusion model. It passes text through a large pre-trained T5 language model and generates high-fidelity images through cascading diffusion model blocks.

#### 4.1.4 Training parameters

We apply an Imagen-like training strategy, i.e., training the base model and then the super-resolution model twice. The Adam optimiser is adopted, having a learning rate of 1e-4. We give 10,000 linear warm-up steps with a batch size of 8 and training epochs of 1,000. The loss function is Mean Squared Error (MSE), formulated as follows.

$$MSE(I, K) = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [I(i, j) - K(i, j)]^2 , \quad (8)$$

where  $M$  and  $N$  denote the total number of pixels in the real image  $I$  and the generated image  $K$ , respectively. A smaller MSE implies that the generated image is closer to the real image.

## 4.2 Experimental Results

In this subsection, we evaluate the proposed model by comparing it against a few state-of-the-art generative models.

### 4.2.1 Performance evaluation

Table 1 demonstrates the results of the quantitative comparison. The proposed model is compared against 10 popular generative models, including GAN and diffusion models. It is evident that the proposed Swinv2-Imagen model outperforms the baselines on all three datasets. Particularly, on the MSCOCO dataset, Swinv2-Imagen significantly outperforms the GAN-based generative model and slightly surpasses the Imagen, achieving an FID of 7.21.

### 4.2.2 Qualitative analysis

Figures 6, 7 and 8 show examples of images generated by our proposed model on MSCOCO, CUB and MM CelebA-HQ, respectively. It can be seen that our model understands the text very well. For example, given the text input, ‘Food cooks in a pot on a stove in a kitchen’, the resulting picture not only contains the food, the stove and the pot, but also places these objects to the exact location. More importantly, based on the word ‘kitchen’, the model also generates other common kitchen objects, such as spoons and storage shelves. This shows that our model understands the text accurately and comprehensively.



A kitchen is shown with a variety of items on the counters.



Food cooks in a pot on a stove in a kitchen.



A brown elephant stepped into the water of a stream.



A full view of an open kitchen and dining area.



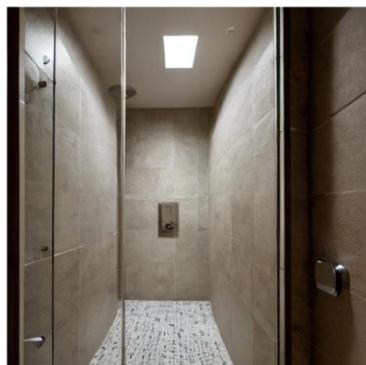
A view of a very large bathroom with mirrored walls.



A woman eating vegetables in front of a stove.



A man on a bicycle riding next to a train.



A shower stall with interesting tile is the focal point.



Two dogs are looking up while they stand near the toilet in the bathroom.

Figure 6: Generated examples on MSCOCO



The bird is dark grey brown with a thick curved bill and a flat shaped tail.



This large black bird has a long wingspan and webbed feet.



This bird is brown in color, with a large brown beak.



The bird is brown with a crooked black beak and a large wingspan.



This grey bird has a large wingspan with a white ring of feathers around its bill.



This bird has dark wings, black, white, and grey, short longs and long neck.



Bird has brown body feathers, brown breast feathers, and brown beak.



A medium sized bird that has tones of dark brown with a pointed bill.



This is a black and brown bird with black wings and a dark pointy beak.

Figure 7: Generated examples on CUB



This person has big lips, blond hair, rosy cheeks, and wavy hair and is wearing lipstick.



The woman is young and has bangs, black hair, high cheekbones, and bushy eyebrows.



This woman has pointy nose, big lips, and wavy hair. She is attractive and is wearing heavy makeup, and lipstick.



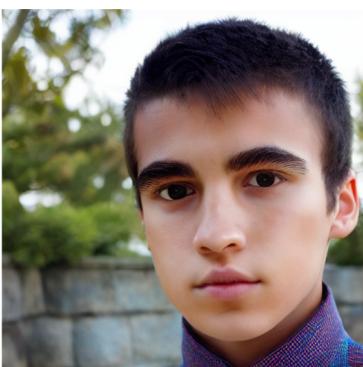
She wears lipstick, and necklace. She is smiling and has bags under eyes, arched eyebrows, gray hair, wavy hair, and mouth slightly open.



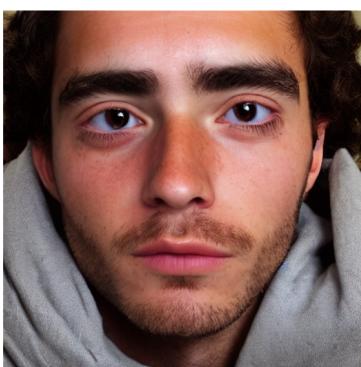
She has wavy hair. She is young and is wearing lipstick.



This man has wavy hair. He is attractive. He has beard.



This young man has oval face, and bushy eyebrows.



The man is young and has bushy eyebrows, and bags under eyes.



This man has bangs, sideburns, brown hair, and bushy eyebrows. He is young.

Figure 8: Generated examples on MM CelebA-HQ

Model	MSCOCO		CUB		MM CelebA-HQ
	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓
PCCM-GAN [66]	33.59	26.52	22.15	4.65	-
DM-GAN [16]	32.64	30.49	16.09	4.75	131.05
SDGAN [67]	29.35	35.69	-	4.64	-
DALL-E [67]	27.5	17.9	56.1	-	12.54
CogView [68]	13.9	18.2	-	-	-
GLIDE [24]	12.24	-	-	-	-
DALL-E 2 [5]	10.39	-	-	-	-
LAFITE [69]	8.12	32.34	10.48	5.97	12.54
Make-A-Scene [70]	7.55	-	-	-	-
Imagen [6]	7.27	-	-	-	-
<b>Swinv2-Imagen</b>	<b>7.21</b>	<b>31.46</b>	<b>9.78</b>	<b>8.44</b>	<b>10.31</b>

Table 1: Experimental results of varied models for Text-To-Image synthesis. Symbols ↑ and ↓ indicate the higher the best and the lower the best, respectively. – means that the indicator is not used in the article.

### 4.3 Ablation study

In order to improve the performance of the generation models, we introduce two new modules to Imagen, i.e., scene graph and Swinv2-Unet. These are the main innovations of the article. In this subsection, two ablation experiments are conducted on MSCOCO to investigate the contributions of the scene graph module and Swinv2-Unet, respectively. The choice to experiment on MSCOCO is based on two considerations. Firstly, each image in MSCOCO contains multiple objects, which is more complex than CUB dataset. Theoretically, it allows a better evaluation on the effect of each module. Secondly, the main baseline we referenced, Imagen, is only experimented on MSCOCO. The aim of Experiment 1 is to evaluate the contribution of the scene graph module. We add only the scene graph, and the diffusion model is still built using Efficient-Unet, which is called Imagen\_sg. Experiment 2 is designed to evaluate the performance of the Swinv2-Unet. We constructed a new diffusion model using our improved Swinv2-Unet and replace Imagen’s super-resolution diffusion models with it, which is called Swinv2-Imagen\_su. The result of Experiment 1 supports our conjecture that merely using a T5 encoder does not sufficiently learn the semantic information of the text, as mentioned in the introduction. Experiment 2 shows that the diffusion model constructed with the Transformer outperforms the CNN-constructed diffusion model in the image generation task. It also can be seen from Table 2 that the FIDs of the Imagen\_sg and Swinv2-Image\_su are very close. This intuitively reveals that the two submodules almost contribute equally to the FID.

Model	Scene Graph	Swinv2-UNet	FID
Imagen			7.27
Imagen_sg	YES		7.24
Swinv2-Imagen_su		YES	7.23
Swinv2-Imagen	YES	YES	7.21

Table 2: Ablation study of Swinv2-Imagen model

## 5 Conclusion

In this paper, we propose a novel text-to-image synthesis model based on Imagen, called the Swinv2-Imagen, which integrates the Transformer and Scene Graph. The improved sliding window-based hierarchical visual Transformer (Swin Transformer v2) avoids the local view of CNN convolution operations. It improves the efficiency and effectiveness of the Transformer applied to image generation. In addition, we introduce a Scene Graph in the text processing stage. Feature vectors of entities and relationships are extracted from the Scene Graph and incorporated into the diffusion model. These additional feature vectors improve the quality of generated images. Swinv2-Imagen produces  $1024 \times 1024$  samples with unprecedented fidelity with these novel components.

Furthermore, it has also recently been noted that autoregressive models can produce diverse and high-quality images from text. Thus, we plan to consider combining autoregressive and diffusion models for image generation and determine the best opportunities to combine their strengths.

## References

- [1] Doyeon Kim, Donggyu Joo, and Junmo Kim. Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access*, 8:153113–153122, 2020.
- [2] Ruifan Li, Ning Wang, Fangxiang Feng, Guangwei Zhang, and Xiaojie Wang. Exploring global and local linguistic representations for text-to-image synthesis. *IEEE Transactions on Multimedia*, 22(12):3075–3087, 2020.
- [3] Shubham Mathesu, Ganesh Bhutkar, and Ayush Rambhad. Atngan: realistic text-to-image synthesis with attentional generative adversarial networks. In *IFIP Conference on Human-Computer Interaction*, pages 397–403. Springer, 2021.
- [4] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *NeurIPS Datasets and Benchmarks*, 2021.
- [5] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.
- [6] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022.
- [7] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683, 2020.
- [8] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12166–12174, 2019.
- [9] Apurva N Ganar, CS Gode, and Sachin M Jambhulkar. Enhancement of image retrieval by using colour, texture and shape features. In *2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies*, pages 251–255. IEEE, 2014.
- [10] Eric Kauderer-Abrams. Quantifying translation-invariance in convolutional neural networks. *arXiv preprint arXiv:1801.01450*, 2017.
- [11] Benjamin Chidester, Minh N Do, and Jian Ma. Rotation equivariance and invariance in convolutional neural networks. *arXiv preprint arXiv:1805.12301*, 2018.
- [12] Jinpeng Li, Yichao Yan, Shengcai Liao, Xiaokang Yang, and Ling Shao. Local-to-global self-attention in vision transformers. *arXiv preprint arXiv:2107.04735*, 2021.
- [13] Chen Liang, Wenguan Wang, Tianfei Zhou, Jiaxu Miao, Yawei Luo, and Yi Yang. Local-global context aware transformer for language-guided video segmentation. *arXiv preprint arXiv:2203.09773*, 2022.
- [14] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, 2018.
- [15] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009, 2022.
- [16] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5795–5803, 2019.
- [17] Bin Zhu and Chong-Wah Ngo. Cookgan: Causality based text-to-image synthesis. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5518–5526, 2020.
- [18] 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*, 41:1947–1962, 2019.
- [19] Weihao Xia, Yujiu Yang, Jing Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2256–2265, 2021.
- [20] Katherine Crowson, Stella Rose Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *ArXiv*, abs/2204.08583, 2022.

- [21] Jun Cheng, Fuxiang Wu, Yuandong Tian, Lei Wang, and Dapeng Tao. Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10908–10917, 2020.
- [22] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.
- [23] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47:1–47:33, 2022.
- [24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.
- [25] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022.
- [26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2021.
- [27] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021.
- [28] Ling Yang, Zhilong Zhang, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Ming-Hsuan Yang, and Bin Cui. Diffusion models: A comprehensive survey of methods and applications. *ArXiv*, abs/2209.00796, 2022.
- [29] Han K. Cao, Cheng Tan, Zhangyang Gao, Guangyong Chen, Pheng-Ann Heng, and Stan Z. Li. A survey on generative diffusion model. *ArXiv*, abs/2209.02646, 2022.
- [30] Gaurav Mittal, Shubham Agrawal, Anuva Agarwal, Sushant Mehta, and Tanya Marwah. Interactive image generation using scene graphs. *arXiv preprint arXiv:1905.03743*, 2019.
- [31] Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, and Bennamoun. Scene graph generation: A comprehensive survey. *ArXiv*, abs/2201.00443, 2022.
- [32] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alexander G. Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2021.
- [33] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, 2015.
- [34] Sebastian Schuster, Ranjay Krishna, Angel X. Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *VL@EMNLP*, 2015.
- [35] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and G. Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54:137–178, 2020.
- [36] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Émilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12602–12611, 2020.
- [37] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10312–10321, 2019.
- [38] Lizhao Gao, Bo Wang, and Wenmin Wang. Image captioning with scene-graph based semantic concepts. *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, 2018.
- [39] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10677–10686, 2019.
- [40] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. *ArXiv*, abs/2007.11731, 2020.
- [41] Jiuxiang Gu, Shafiq R. Joty, Jianfei Cai, Handong Zhao, Xu Yang, and G. Wang. Unpaired image captioning via scene graph alignments. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10322–10331, 2019.
- [42] Yikang Li, Tao Ma, Yeqi Bai, Nan Duan, Sining Wei, and Xiaogang Wang. Pastegan: A semi-parametric method to generate image from scene graph. *Advances in Neural Information Processing Systems*, 32, 2019.

- [43] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019.
- [44] William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2020.
- [45] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
- [46] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.
- [47] Fenxiao Chen, Yun Cheng Wang, Bin Wang, and C.-C. Jay Kuo. Graph representation learning: a survey. *APSIPA Transactions on Signal and Information Processing*, 9, 2020.
- [48] William L. Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *ArXiv*, abs/1709.05584, 2017.
- [49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015.
- [50] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.
- [51] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, S...*, 11045:3–11, 2018.
- [52] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39:1856–1867, 2020.
- [53] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059, 2020.
- [54] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15:749–753, 2018.
- [55] Sijing Cai, Yunxian Tian, Harvey Lui, Haishan Zeng, Yi Wu, and Guannan Chen. Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quantitative imaging in medicine and surgery*, 10 6:1275–1285, 2020.
- [56] N. Ibtehaz and Mohammad Sohel Rahman. Multiresunet : Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks : the official journal of the International Neural Network Society*, 121:74–87, 2020.
- [57] Md. Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *ArXiv*, abs/1802.06955, 2018.
- [58] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *ArXiv*, abs/2105.05537, 2021.
- [59] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.
- [60] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- [61] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096, 2019.
- [62] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.
- [63] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, 2014.
- [64] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022.
- [65] 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 *NIPS*, 2017.

- [66] Zhongjian Qi, Jun Sun, Jinzhao Qian, Jiajia Xu, and Shu Zhan. Pccm-gan: Photographic text-to-image generation with pyramid contrastive consistency model. *Neurocomputing*, 449:330–341, 2021.
- [67] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–842, 2021.
- [68] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*, 2021.
- [69] 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. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17886–17896, 2022.
- [70] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *ArXiv*, abs/2203.13131, 2022.