Neural Computing and Applications (2024) 36:17245–17260
https://doi.org/10.1007/s00521-023-09021-x

ORIGINAL ARTICLE

Swinv2-Imagen: hierarchical vision transformer diffusion models for text-to-image generation

Ruijun Li[1] · Weihua Li[1] · Yi Yang[2] · Hanyu Wei[3] · Jianhua Jiang[4] · Quan Bai[3]

Received: 7 December 2022 / Accepted: 5 September 2023 / Published online: 6 October 2023

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, using T5 alone to obtain text

Yi Yang, Hanyu Wei, Jianhua Jiang and Quan Bai have contributed equally to this work.

✉ Ruijun Li
zjc0233@autuni.ac.nz

✉ Weihua Li
weihua.li@aut.ac.nz

Yi Yang
yyang@hfut.edu.cn

Hanyu Wei
hanyu.wei@utas.edu.au

Jianhua Jiang
jjh@jlufe.edu.cn

Quan Bai
quan.bai@utas.edu.au

[1] Auckland University of Technology, Auckland 1010, New Zealand

[2] Hefei University of Technology, Hefei 230601, China

[3] University of Tasmania, Hobart 7005, Australia

[4] Jilin University of Finance and Economics, Changchun, China

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 focuses on the consistency of these low-level features under transformations, such as translation [10] and rotation [11]. This is also the main reason why CNNs are widely used in object detection [12]. In other words, while the convolutional filters are good at detecting key points, object boundaries and other basic units that constitute the visual elements, it fails to extract features efficiently in terms of global and layout. For text-to-image synthesis tasks, it is significant to consider how to accurately extract the complex relationships between objects from the limited text. The Transformer is more natural and efficient than CNN in processing this demand. This is mainly because the attention in the Transformer can effectively mine the relationships between text features, allowing the model not only focuses on local information but also has a diffusion mechanism to find expressions from the local to global layout [13, 14].

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 [15] to obtain the Scene Graph from the text. Subsequently, the entity and relationship embeddings are extracted using a frozen Graph Convolution Network (GCN) [15]. 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 [16], which allows the model to learn features from local to global. 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. The model leverages attention to explore the relationship between features, allowing the diffusion model to focus on different granularities of features at different moments, from local to global;
3. We fuse the scene graph with the diffusion model, and the experimental results demonstrate that the resulting images not only generate the objects specified in the text, but also additional objects based on specific words (e.g. kitchen);
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 Sect. 2, related works are reviewed. In Sect. 3, we elaborate on the proposed Swinv2-Imagen model. In Sect. 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 Sect. 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 [17–22] and the diffusion-based models [23–27]. 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, 28, 29].

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, 30]. 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 [25], 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 [15]. 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 [31]. Previous studies reveal that the performance of multimodal models, such as text-to-image synthesis, is significantly dependent on mining visual relationships [32]. Scene graphs can provide a high level of understanding regarding scene information [15]. 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 [33]. Recently, the scene graphs have been used extensively for tasks such as text-based image retrieval [34, 35], semantic segmentation [36, 37], visual question answering [38], image captioning [39–42] and image generation [15, 31, 43, 44]. The recently proposed dynamic scene graph generation also demonstrates the prospects of scene graphs in video monitoring, autonomous driving, and other fields related to video processing and generation [45].

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 [46]. 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 [46]. Node2vec [47] is a typical representative model of the former. It is based on Skip-Gram [48] theory to learn the embedding of nodes on a graph and optimises the sampling method. It is proposed in related studies [49] 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 [50] comprises 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 [15], 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 Convolutional Neural Networks (GCNs) have gained significant attention in recent studies as powerful tools for analysing graph-structured data, such as social networks and database tables [51]. For the text-to-image synthesis task, generation models leverage GCNs to capture semantic relationships between textual descriptions and visual features, enabling more accurate image generation. Recently, many studies have been conducted to enhance graph-based neural networks. Specifically, a novel augmentation method called GraphENS has been proposed to address the issue of overfitting to neighbour sets of minor class nodes [52, 53]. They proposed a saliency-based node mixing method to leverage the abundant class-generic attributes of other nodes while preventing the injection of class-specific features. To mitigate the negative impact of oversampling on message passing, they restricted the message passing only to the incoming edges of the oversampled nodes.

2.3 UNet

UNet is an encoder–decoder architecture, which is scalable in structure [54]. 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) [55], 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 upsampling also allows the segmentation map to recover information such as edges more finely. However, UNet also has some shortcomings. For example, UNet++ [56, 57] 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+ [58] 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 [59] and DenseUNet [60] are inspired by Residual and Dense connections, respectively; each sub-module of the UNet is replaced with a form having a Residual connection and a Dense connection. There are variants, e.g. MultiResU-Net [61] and R2 UNet [62]. 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 [63]. 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 [16] 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 Fig. 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 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 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 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 [64–66], BERT [67] 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.

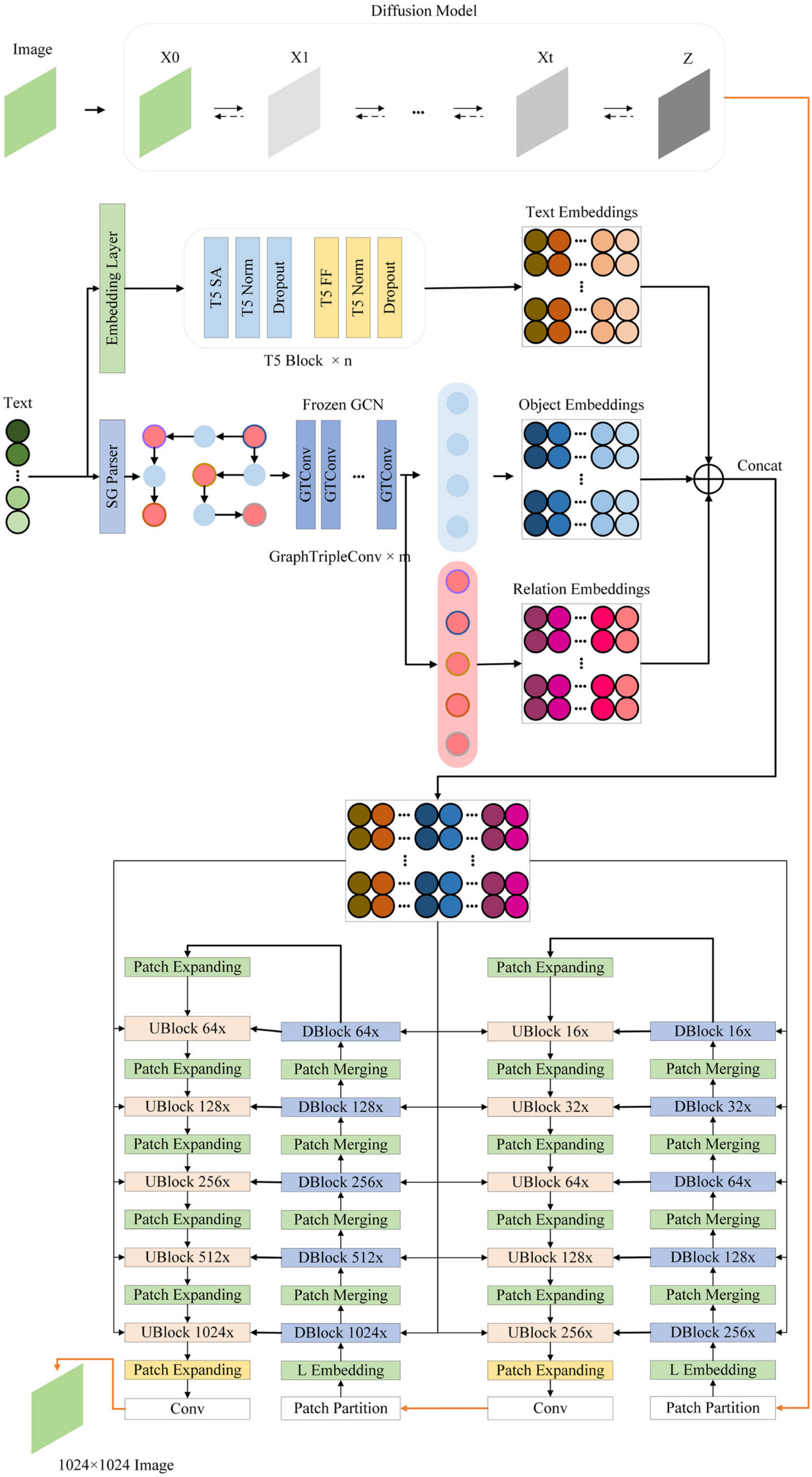


Fig. 1 Overall architecture of Swinv2-Imagen. The text is passed through both a frozen T5 Encoder and a scene graph. The scene graph mines complex entity relationships explicitly, ensuring that the model understands the text semantics accurately

3.2 Scene graph and frozen graph convolutional neural network

This sub-module aims to extract entity and relationship features 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 [15] have been proven to be highly effective in extracting object relationships from the text. As shown in Fig. 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 follows: (O, E), where $O = (o_1, o_2, o_3, \cdots, o_n)$ denotes each 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. 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}, i.e. $\mathbf{v}_i, \mathbf{v}_r \in \mathbb{R}^{D_{\text{in}}}$. In the graph convolution sub-module, these vectors are adopted to compute output vectors with dimension D_{out} for each node and edge, i.e. $\mathbf{v}'_i, \mathbf{v}'_r \in \mathbb{R}^{D_{\text{out}}}$. Three functions, g_s, g_o and g_p are used to calculate the object features vectors and relation vectors of output. They take a triplet as input, i.e. $(\mathbf{v}_i, \mathbf{v}_r, \mathbf{v}_j)$. In the scene graph, given an edge $\mathbf{v}_r$, the two associated objects, $\mathbf{v}_i$ and $\mathbf{v}_j$, are determined. Thus, the output relationship vector $\mathbf{v}'_r$ can be simply expressed as:

$$\mathbf{v}'_r = g_p(\mathbf{v}_i, \mathbf{v}_r, \mathbf{v}_j) \tag{1}$$

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 Eq. (2) is used to compute all vectors starting at node o_i and function g_o in Eq. (3) is used to compute all vectors ending at node o_i. Afterwards, these vectors are collected into lists V_i^s and V_i^o.

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

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

Then, the output vector $\mathbf{v}'_\mathbf{i}$ for the entity o_i is expressed as follows:

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

where h denotes a function that pools all vectors in lists V_i^s and V_i^o to a single output vector [15].

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.

The diffusion model can be described as an Encoder–Decoder architecture. It first adds Gaussian noise (ϵ) to the original image $(x_0 \sim q(x_0))$ in an iterative manner, the number of iterations being T (T is timestep, usually $T = 1000$). When T tends to infinity, i.e. $(T \to \infty)$, the image is nearly a random Gaussian noise distribution x_T.

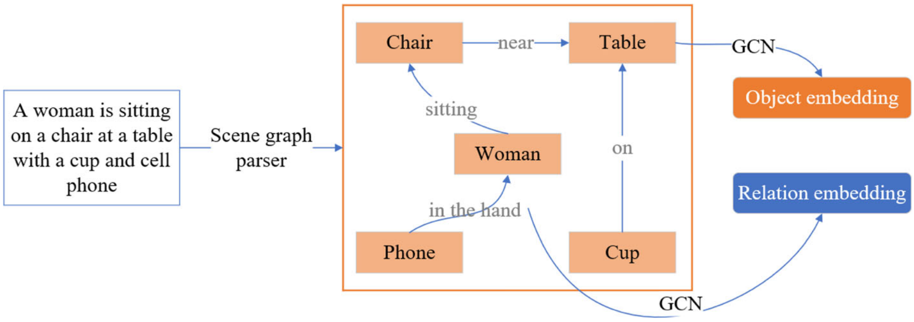


Fig. 2 Process of Object and Relation embeddings extraction. The input to the model is a sentence, which is first parsed by the scene graph parser into a graph structure (scene graph). Each node represents an object in the text and each edge represents a relationship between objects. Finally, all the nodes and edges are parsed by the graph neural network into an object embedding and a relation embedding, respectively

This process is called forward diffusion and can be thought of as an encoder. The model then learns how to recover the noise distribution (x_T) to the original image ($x_0 \sim q(x_0)$) by gradually removing the noise from x_T. The process is called reverse diffusion and can be thought of as a decoder [28, 29].

In the forward diffusion, the result at timestep t is mainly related to the outcome at moment $t-1$ and the added noise ϵ_t, i.e.

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t \quad q(x_t|x_{t-1}) \sim \mathcal{N}(\sqrt{1-\beta_t}, \beta_t), \tag{5}$$

$$q(x_{1:T}|x_0) = \prod_{i=1}^{T} q(x_t|x_{t-1}) \tag{6}$$

where β_t I prefer to understand as a linear weight value. At different timestep, x_{t-1} and ϵ_t have different effects on the result. When T is small, e.g. $t=1$, x_{t-1} has a greater impact on the result and adds little noise. Conversely, when t is large, e.g. $t=900$, more noise is added and the contribution to the result is larger than x_{t-1}.

The distribution of the noise added at each timestep in the forward process is identical, i.e. $\epsilon_1, \epsilon_2, \ldots\ldots \sim \mathcal{N}(0, \mathbf{I})$. Thus, we can compute the result at any timestep x_t directly from x_0, i.e.

$$x_t = \sqrt{\bar{\alpha}_t}x_o + \sqrt{1-\bar{\alpha}_t}\epsilon_t \tag{7}$$

where $\alpha = 1-\beta$, $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$.

Reverse diffusion is an image generation process. The Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ will be taken as input to infer and reconstruct the true sample by sampling from distribution $q(x_{t-1}|x_t)$, i.e.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\right)f_\theta(x_t, t) \tag{8}$$

where $f_\theta(x_t, t)$ is a function used to predict the noise ϵ added in the forward diffusion. This is mainly because it is difficult to infer the true distribution of the image directly from the random noise x_T. In other words, the objective of the diffusion generation model is to evaluate the difference between the predicted noise data and the true added noise data, i.e.

$$p(x_{t-1}|x_t) = ||\epsilon - f_\theta(x_t, t)||. \tag{9}$$

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 [63], the complete structure of which is shown in Fig. 3.

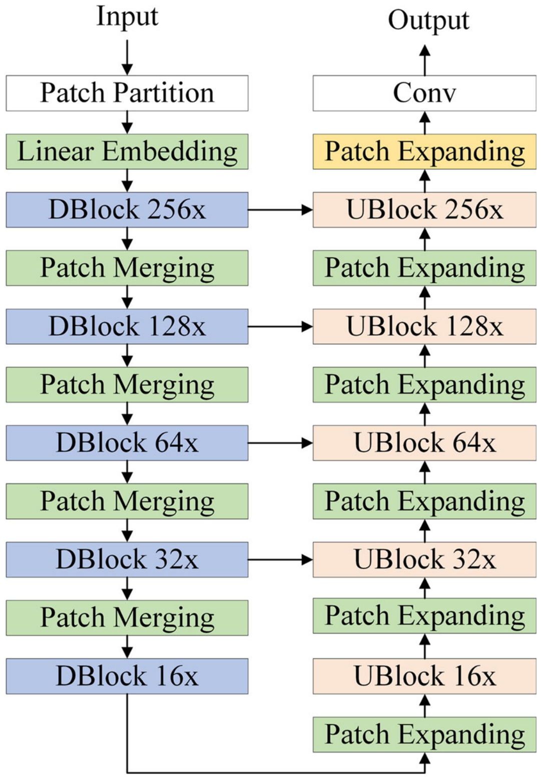


Fig. 3 UNet architecture of the super-resolution sub-module. The architecture includes an encoder(downsampling), bottleneck, and decoder(upsampling). Skip connections are used between the encoder and decoder. All components are built based on the Swin Transformer v2 block

A distinctive feature of Swinv2-Unet compared to Swin-Unet is the replacement of the $dot(\mathbf{K}, \mathbf{Q})$ operation with cosine normalisation [68] in the attention part, which makes the attention output more stable. Given two vectors, $\mathbf{Q}$ and $\mathbf{K}$, the cosine normalisation could be expressed as follows:

$$Cosine(\mathbf{Q}, \mathbf{K}) = \frac{\sum_i (q^i k^i)}{\sqrt{\sum_i (q^i)^2}\sqrt{\sum_i (k^i)^2}}. \tag{10}$$

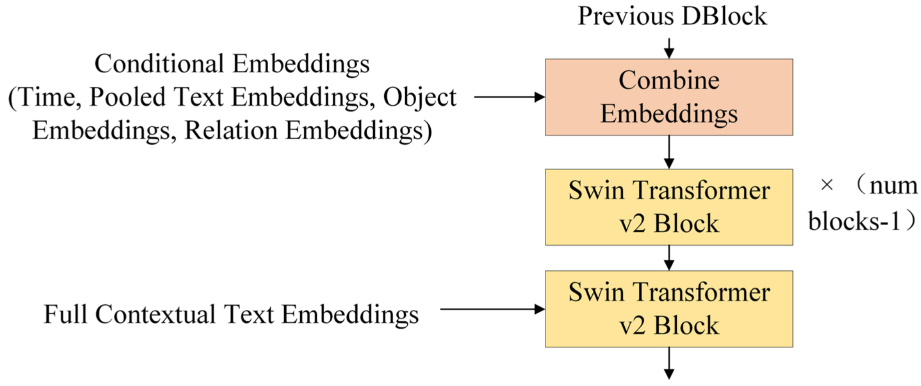
The DBlock and UBlock of Swinv2-UNet consist of the Swin Transformer v2 block, which comprises LayerNorm (LN) layers, multi-headed self-attention modules, Residual connections and a 2-layer MLP with GELU nonlinearity. The Swin Transformer v2 block could be represented as follows:

$$\hat{z}^{l+1} = \mathrm{LN}(\mathrm{Attn}(\mathrm{z}^{\mathrm{l}})) + \mathrm{z}^{\mathrm{l}} \tag{11}$$

$$z^{l+1} = \mathrm{MLP}(\mathrm{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1}, \tag{12}$$

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.

Fig. 4 Swinv2-Unet DBlock


Previous DBlock
Conditional Embeddings (Time, Pooled Text Embeddings, Object Embeddings, Relation Embeddings)
Combine Embeddings
Swin Transformer v2 Block
× （num blocks-1）
Full Contextual Text Embeddings
Swin Transformer v2 Block

The attention of Swinv2 is expressed as follows:

$$Attn(\mathbf{Q},\mathbf{K},\mathbf{V}) = SoftMax\left(\frac{Cosine(Q,K)}{\tau} + B\right) \quad , \qquad (13)$$

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

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 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 [57]. Subsequently, the conditional embedding inputs are also introduced to the cross-attention layer. Similar to the DBlock, 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× 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) [69], the Caltech-UCSD Birds-200-2011 (CUB-200-2011) [70] and MM CelebA-HQ [20] 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

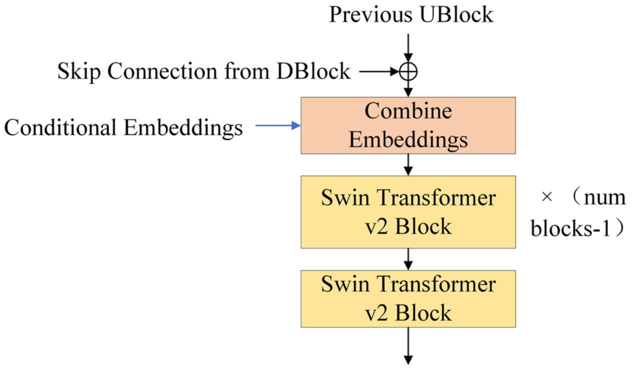

Previous UBlock
Skip Connection from DBlock
Conditional Embeddings
Combine Embeddings
Swin Transformer v2 Block
× （num blocks-1）
Swin Transformer v2 Block

Fig. 5 Swinv2-Unet UBlcok

generative models such as CogView and Craiyon, produce distorted and less realistic faces.

- MSCOCO[1] 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[2] 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[3] 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) [71] 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 [72] (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 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 [17] (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 [73] (Semantics Disentangling for Text-to-Image Generation) consists of two modules, i.e. Siamese and semantic conditioned batch normalisation, to extract high-level and low-level semantic features respectively.
- CogView [74] 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 [25] 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 priori 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 priori model.
- LAFITE [75] 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 1000. The loss function is Mean Squared Error (MSE), formulated as follows.

[1] https://cocodataset.org/.

[2] https://deepai.org/dataset/cub-200-2011.

[3] https://github.com/weihaox/Multi-Modal-CelebA-HQ-Dataset.

$$MSE(I,K) = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [I(i,j) - K(i,j)]^2, \qquad (14)$$

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. It can be seen from Fig. 6 that our model has achieved the best result in terms of FID. However, our model, in IS metric, is lower than SDGAN and LAFITE. One possible reason for this result is that IS is not very robust in evaluating classes that differ significantly from the ImageNet [78] and is more sensitive to data perturbations. This is also the main reason why this metric is not widely used in most diffusion-based generation methods, such as DALLE2 and Imagen.

4.2.2 Qualitative analysis

Figure 7 shows examples of images generated by our proposed model on MSCOCO, CUB and MM CelebA-HQ. 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.

Figure 8 illustrates the qualitative comparison of the proposed model and the GAN-based, diffusion-based generative models, i.e. DM-GAN [17], DF-GAN [79], VQ-Diffusion [80]. Compared to diffusion-based models, the GAN-based models lose many detailed features in the generated results. For example, the bird’s eyes are very blurred in the third image in the first row and the second image in the second row. One possible reason for this result is that the diffusion model improves the generalisation ability of the model by iterating over the image several times, with each iteration perturbing the image slightly (by adding randomly noisy data to the image). GANs, on the other hand, usually rely on continuous optimisation over large amounts of data in order to generate high-quality images. Compared to VQ-Diffusion, which is a diffusion-based model, our results are more realistic and contain more fine-grained features. Particularly, the blue birds in the third column generated by our method are better than that generated by VQ-Diffusion. One possible reason for this result is that VQ-diffusion is a typical two-stage generative model [80]. First is the vector quantisation stage,

Table 1 Experimental results of varied models for Text-To-Image synthesis

Model	MSCOCO		CUB		MM CelebA-HQ
	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓
PCCM-GAN [72]	33.59	26.52	22.15	4.65	14.52
DM-GAN [17]	32.64	30.49	16.09	4.75	131.05
SDGAN [73]	29.35	35.69	29.3	4.64	15.1
DALL-E [76]	27.5	17.9	56.1	2.65	12.54
CogView [74]	27.1	18.2	N/A	N/A	N/A
GLIDE [25]	12.24	30.47	N/A	N/A	9.69
DALL-E 2 [5]	10.39	N/A	N/A	N/A	N/A
LAFITE [75]	8.12	32.34	10.48	5.97	12.54
Make-A-Scene [77]	7.55	N/A	22.5	N/A	N/A
Imagen [6]	7.27	N/A	N/A	N/A	N/A
Swinv2-Imagen	**7.21**	**31.46**	**9.78**	**8.44**	**10.31**

The bold values indicate the best results

Symbols ↑ and ↓ indicate the higher the best and the lower the best, respectively. N/A means that the indicator is not used in the article

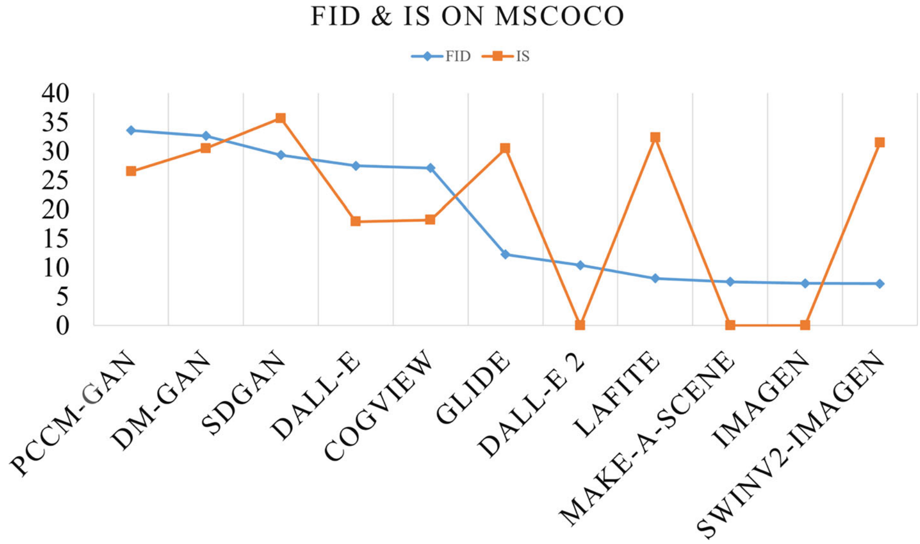


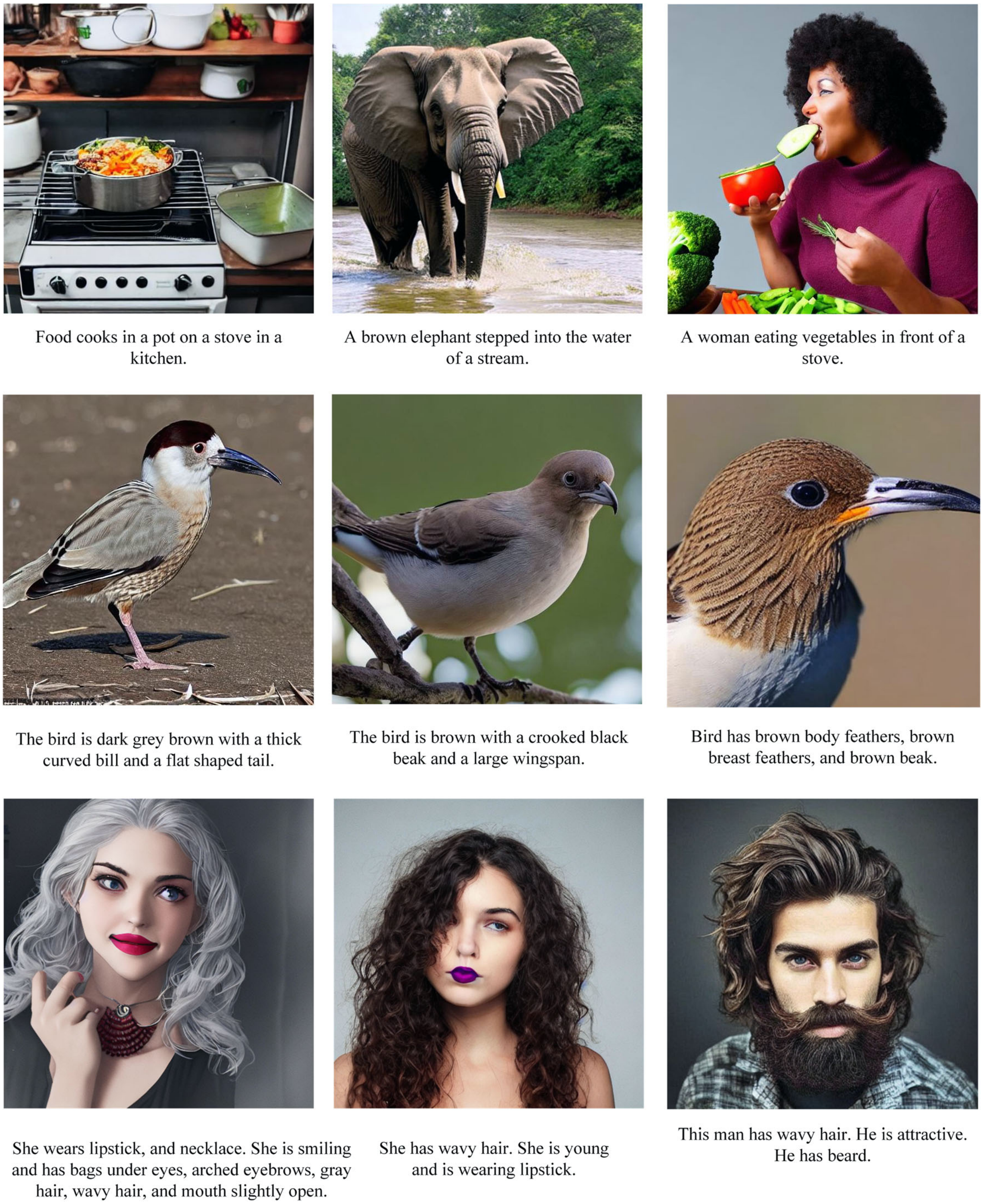
Fig. 6 FID and IS on MSCOCO. Smaller FID is better, larger IS is better. 0 means that the model does not use this evaluation metric

where the original image is represented as a set of high-dimensional vectors and mapped into a discretised space using a vector quantisation model. Second is the diffusion computation stage, where a discretised sequence of vectors is used as the initial state, which is iterated over several times using a diffusion model. Finally, the resulting discretised vector sequence is transformed into an image. The vectorisation results of the first stage will directly affect the quality of the generation results. By comparison, our proposed model is designed as an end-to-end architecture that optimises the entire generation process holistically. Our model eliminates the need for intermediate stages, facilitating better optimisation and faster convergence. In addition, our model also outperforms other generation models in terms of text-image alignment. The text description of the first column requires the bird's breast to be white, but this feature seems to be grey in the results of other models, especially DM-GAN. In summary, by comparing with other GAN-based and diffusion-based generation models, it can be seen that our model synthesises fine-grained and detailed images on CUB.

Figure 9 presents the qualitative comparison between our model and LAFITE [75] on MSCOCO. Intuitively, our results are more colourful and saturated. For example, in the first and fourth columns, our bus and city street include more colours and the images are brighter. Furthermore, our model is also better for text understanding. In the third column, the room should include two colours, white and beige, however, in the LAFITE result, there are just white walls and a white cupboard. There is not any trace of the beige features. In contrast, our generated room contains the two colours required by the text, and the overall layout is more realistic. Finally, our model is also better regarding image quality. The tops of the bus and room generated by LAFITE are distorted and the results are generally blurred. Our model has a significant advantage over LAFITE in generating objects such as buildings, buses, trees, etc. Although the two models are very close in terms of FID and IS in the quantitative analysis in Table 1, our model is superior in terms of the quality of the generated images.

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-



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

A brown elephant stepped into the water of a stream.

A woman eating vegetables in front of a stove.

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

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

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

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.

Fig. 7 Generated examples by proposed model on COCO, CUB and MM CelebA-HQ. The resulting images generate not only the objects requested in the sentence but also additional objects based on special words(e.g. kitchen). For example, in the first image, the resulting image includes food, a pot, and a stove (required in the text) and some spoons and rice cookers (common kitchen items)

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.

5 Conclusion and future work

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

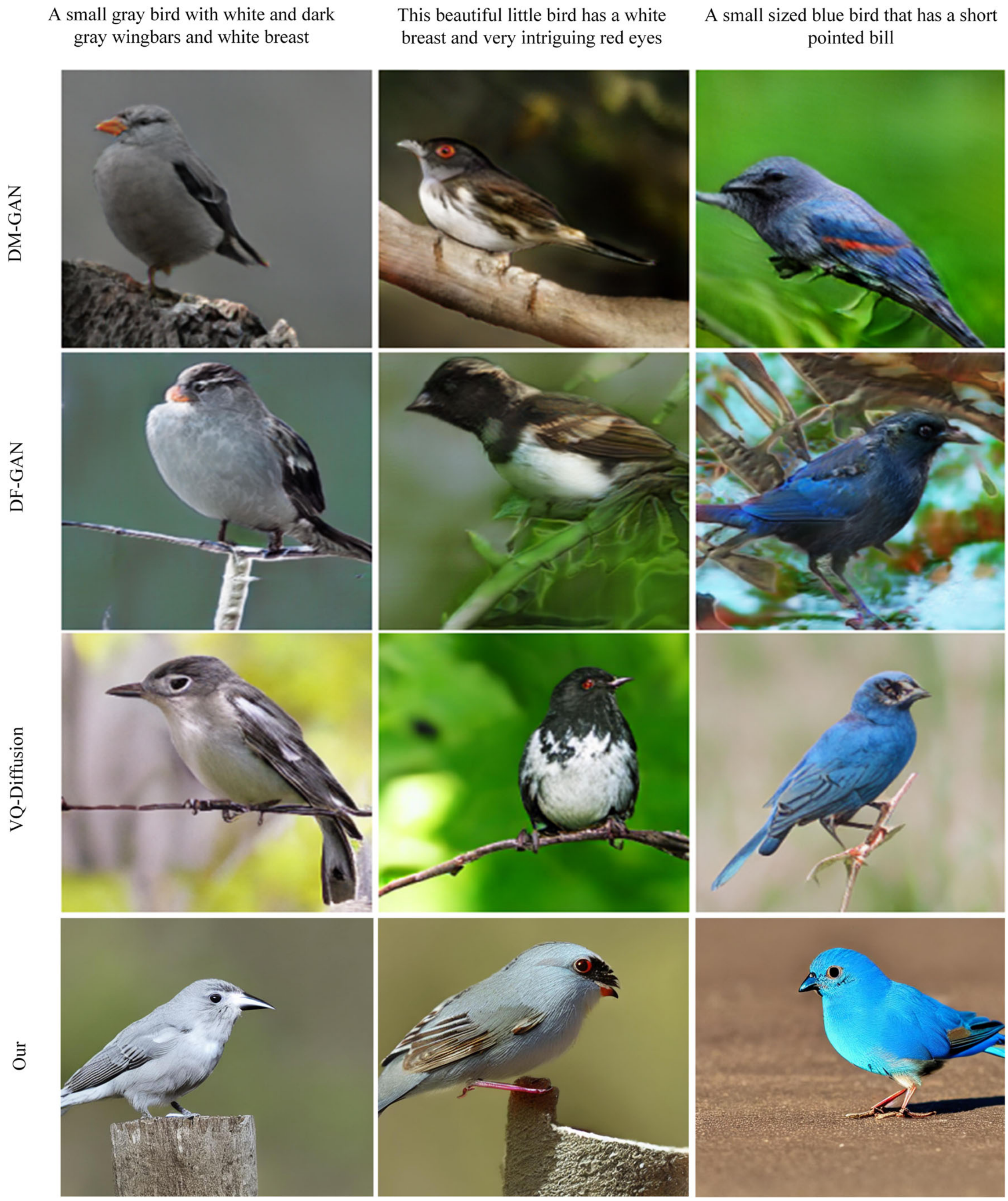

A small gray bird with white and dark gray wingbars and white breast
This beautiful little bird has a white breast and very intriguing red eyes
A small sized blue bird that has a short pointed bill
DM-GAN
DF-GAN
VQ-Diffusion
Our

Fig. 8 Comparison with GAN-based and diffusion models on CUB-200 dataset. For each method, we present three captions and the corresponding generated images. Our resulting images are more detailed in colour and higher in quality than the popular GAN models

(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 × 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.

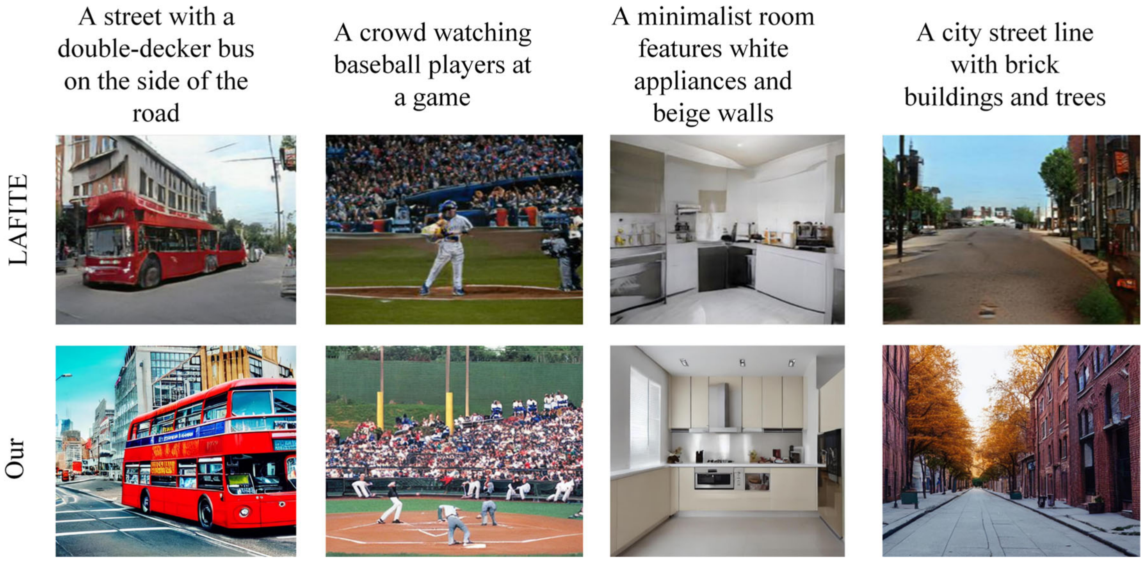

A street with a double-decker bus on the side of the road
A crowd watching baseball players at a game
A minimalist room features white appliances and beige walls
A city street line with brick buildings and trees
LAFITE
Our

Fig. 9 Comparison with LAFITE on MSCOCO dataset

Table 2 Ablation study of Swinv2-Imagen model

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

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability The data that support the findings of this study are openly available in [Microsoft COCO] at (https://cocodataset.org/.); [CUB-200-2011] at (https://deepai.org/dataset/cub-200-2011.) and [Multi-Modal-CelebA-HQ] at (https://github.com/weihaox/Multi-Modal-CelebA-HQ-Dataset.).

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Kim D, Joo D, Kim J (2020) Tivgan: text to image to video generation with step-by-step evolutionary generator. IEEE Access 8:153113–153122
2. Li R, Wang N, Feng F, Zhang G, Wang X (2020) Exploring global and local linguistic representations for text-to-image synthesis. IEEE Trans Multimed 22(12):3075–3087
3. Mathesul S, Bhutkar G, Rambhad A (2021) Attngan: realistic text-to-image synthesis with attentional generative adversarial networks. In: IFIP conference on human-computer interaction, pp 397–403. Springer
4. Park DH, Azadi S, Liu X, Darrell T, Rohrbach A (2021) Benchmark for compositional text-to-image synthesis. In: NeurIPS datasets and benchmarks
5. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with clip latents. ArXiv arXiv:2204.06125
6. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour SKS, Ayan BK, Mahdavi SS, Lopes RG, Salimans T, Ho J, Fleet DJ, Norouzi M (2022) Photorealistic text-to-image diffusion models with deep language understanding. ArXiv arXiv:2205.11487
7. Raffel C, Shazeer NM, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. ArXiv arXiv:1910.10683
8. Li W, Zhang P, Zhang L, Huang Q, He X, Lyu S, Gao J (2019) Object-driven text-to-image synthesis via adversarial training. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12166–12174
9. Ganar AN, Gode C, Jambhulkar SM (2014) Enhancement of image retrieval by using colour, texture and shape features. In: 2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies, pp. 251–255. IEEE
10. Kauderer-Abrams E (2017) Quantifying translation-invariance in convolutional neural networks. arXiv preprint arXiv:1801.01450

11. Chidester B, Do MN, Ma J (2018) Rotation equivariance and invariance in convolutional neural networks. arXiv preprint arXiv:1805.12301
12. Zhao Z-Q, Zheng P, Xu S-T, Wu X (2019) Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst 30(11):3212–3232
13. Li J, Yan Y, Liao S, Yang X, Shao L (2021) Local-to-global self-attention in vision transformers. arXiv preprint arXiv:2107.04735
14. Liang C, Wang W, Zhou T, Miao J, Luo Y, Yang Y (2022) Local-global context aware transformer for language-guided video segmentation. arXiv preprint arXiv:2203.09773
15. Johnson J, Gupta A, Fei-Fei L (2018) Image generation from scene graphs. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1219–1228
16. Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, Ning J, Cao Y, Zhang Z, Dong L, Wei F, Guo B (2022) Swin transformer v2: Scaling up capacity and resolution. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), 11999–12009
17. Zhu M, Pan P, Chen W, Yang Y (2019) Dm-gan: dynamic memory generative adversarial networks for text-to-image synthesis. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), 5795–5803
18. Zhu B, Ngo C-W (2020) Cookgan: Causality based text-to-image synthesis. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5518–5526
19. Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN (2019) Stackgan++: realistic image synthesis with stacked generative adversarial networks. IEEE Trans Pattern Anal Mach Intell 41:1947–1962
20. Xia W, Yang Y, Xue J, Wu B (2021) Tedigan: text-guided diverse face image generation and manipulation. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 2256–2265
21. Crowson K, Biderman SR, Kornis D, Stander D, Hallahan E, Castricato L, Raff E (2022) Vqgan-clip: open domain image generation and editing with natural language guidance. ArXiv arXiv:2204.08583
22. Cheng J, Wu F, Tian Y, Wang L, Tao D (2020) Rifegan: rich feature generation for text-to-image synthesis from prior knowledge. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10908–10917
23. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. ArXiv arXiv:2006.11239
24. Ho J, Saharia C, Chan W, Fleet DJ, Norouzi M, Salimans T (2022) Cascaded diffusion models for high fidelity image generation. J Mach Learn Res 23:47–14733
25. Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I, Chen M (2022) Glide: towards photorealistic image generation and editing with text-guided diffusion models. In: ICML
26. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10674–10685
27. Song J, Meng C, Ermon S (2021) Denoising diffusion implicit models. ArXiv arXiv:2010.02502
28. Dhariwal P, Nichol A (2021) Diffusion models beat gans on image synthesis. ArXiv arXiv:2105.05233
29. Yang L, Zhang Z, Hong S, Xu R, Zhao Y, Shao Y, Zhang W, Yang M-H, Cui B (2022) Diffusion models: a comprehensive survey of methods and applications. ArXiv arXiv:2209.00796
30. Cao HK, Tan C, Gao Z, Chen G, Heng P-A, Li SZ (2022) A survey on generative diffusion model. ArXiv arXiv:2209.02646
31. Mittal G, Agrawal S, Agarwal A, Mehta S, Marwah T (2019) Interactive image generation using scene graphs. arXiv preprint arXiv:1905.03743
32. Zhu G, Zhang L, Jiang Y, Dang Y, Hou H, Shen P, Feng M, Zhao X, Miao Q, Shah SAA (2022) Bennamoun: scene graph generation: a comprehensive survey. ArXiv arXiv:2201.00443
33. Chang X, Ren P, Xu P, Li Z, Chen X, Hauptmann AG (2021) A comprehensive survey of scene graphs: generation and application. IEEE Trans Pattern Anal Mach Intell 45:1–26
34. Johnson J, Krishna R, Stark M, Li L-J, Shamma DA, Bernstein MS, Fei-Fei L (2015) Image retrieval using scene graphs. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 3668–3678
35. Schuster S, Krishna R, Chang AX, Fei-Fei L, Manning CD (2015) Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: VL@EMNLP
36. Taghanaki SA, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G (2020) Deep semantic segmentation of natural and medical images: a review. Artif Intell Rev 54:137–178
37. Jaritz M, Vu T-H, de Charette R, Wirbel É, Pérez P (2020) xmuda: cross-modal unsupervised domain adaptation for 3d semantic segmentation. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), 12602–12611
38. Li L, Gan Z, Cheng Y, Liu J (2019) Relation-aware graph attention network for visual question answering. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 10312–10321
39. Gao L, Wang B, Wang W (2018) Image captioning with scene-graph based semantic concepts. In: Proceedings of the 2018 10th international conference on machine learning and computing
40. Yang X, Tang K, Zhang H Cai J (2019) Auto-encoding scene graphs for image captioning. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10677–10686
41. Zhong Y, Wang L, Chen J, Yu D, Li Y (2020) Comprehensive image captioning via scene graph decomposition. ArXiv arXiv:2007.11731
42. Gu J, Joty SR, Cai J, Zhao H, Yang X, Wang G (2019) Unpaired image captioning via scene graph alignments. In: 2019 IEEE/CVF international conference on computer vision (ICCV), 10322–10331
43. Li Y, Ma T, Bai Y, Duan N, Wei S, Wang X (2019) Pastegan: a semi-parametric method to generate image from scene graph. Adv Neural Inf Process Syst **32**
44. Zhao B, Meng L, Yin W, Sigal L (2019) Image generation from layout. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8584–8593
45. Li Y, Yang X, Xu C (2022) Dynamic scene graph generation via anticipatory pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13874–13883
46. Hamilton WL (2020) Graph representation learning. Synthesis lectures on artificial intelligence and machine learning
47. Grover A, Leskovec J (2016) node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining
48. Mikolov T, Chen K, Corrado GS, Dean J (2013) Efficient estimation of word representations in vector space. In: ICLR
49. Chen F, Wang YC, Wang B, Kuo C-CJ (2020) Graph representation learning: a survey. APSIPA Trans Signal Inf Process **9**
50. Hamilton WL, Ying R, Leskovec J (2017) Representation learning on graphs: methods and applications. ArXiv arXiv:1709.05584
51. Chen J, Ye G, Zhao Y, Liu S, Deng L, Chen X, Zhou R, Zheng K (2022) Efficient join order selection learning with graph-based representation. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, pp 97–107

52. Park J, Song J, Yang E (2021) Graphens: Neighbor-aware ego network synthesis for class-imbalanced node classification. In: International conference on learning representations
53. Ghorbani M, Kazi A, Baghshah MS, Rabiee HR, Navab N (2022) Ra-gcn: graph convolutional network for disease prediction problems with imbalanced data. Med Image Anal 75:102272
54. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. ArXiv arXiv:1505.04597
55. Shelhamer E, Long J, Darrell T (2015) Fully convolutional networks for semantic segmentation. 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 3431–3440
56. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2018) 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
57. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2020) Unet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging 39:1856–1867
58. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen Y-W, Wu J (2020) Unet 3+: A full-scale connected unet for medical image segmentation. ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1055–1059
59. Zhang Z, Liu Q, Wang Y (2018) Road extraction by deep residual u-net. IEEE Geosci Remote Sens Lett 15:749–753
60. Cai S, Tian Y, Lui H, Zeng H, Wu Y, Chen G (2020) Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. Quant Imaging Med Surg 10(6):1275–1285
61. Ibtehaz N, Rahman MS (2020) Multiresunet: rethinking the u-net architecture for multimodal biomedical image segmentation. Neural Netw Off J Int Neural Netw Soc 121:74–87
62. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK (2018) Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. ArXiv arXiv:1802.06955
63. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M (2021) Swin-unet: Unet-like pure transformer for medical image segmentation. ArXiv arXiv:2105.05537
64. Radford A, Narasimhan K (2018) Improving language understanding by generative pre-training
65. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners
66. Brock A, Donahue J, Simonyan K (2019) Large scale gan training for high fidelity natural image synthesis. ArXiv arXiv:1809.11096
67. Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL
68. Luo C, Zhan J, Wang L, Yang Q (2018) Cosine normalization: Using cosine similarity instead of dot product in neural networks. ArXiv arXiv:1702.05870
69. Cho K, van Merrienboer B, Çaglar Gülçehre Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: EMNLP
70. Ho J (2022) Classifier-free diffusion guidance. ArXiv arXiv:2207.12598
71. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS
72. Qi Z, Sun J, Qian J, Xu J, Zhan S (2021) Pccm-gan: photographic text-to-image generation with pyramid contrastive consistency model. Neurocomputing 449:330–341
73. Zhang H, Koh JY, Baldridge J, Lee H, Yang Y (2021) Cross-modal contrastive learning for text-to-image generation. 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 833–842
74. Ding M, Yang Z, Hong W, Zheng W, Zhou C, Yin D, Lin J, Zou X, Shao Z, Yang H, Tang J (2021) Cogview: Mastering text-to-image generation via transformers. In: NeurIPS
75. Zhou Y, Zhang R, Chen C, Li C, Tensmeyer C, Yu T, Gu J, Xu J, Sun T (2022) Towards language-free training for text-to-image generation. 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 17886–17896
76. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I (2021) Zero-shot text-to-image generation. ArXiv arXiv:2102.12092
77. Gafni O, Polyak A, Ashual O, Sheynin S, Parikh D, Taigman Y (2022) Make-a-scene: scene-based text-to-image generation with human priors. ArXiv arXiv:2203.13131
78. Barratt ST, Sharma R (2018) A note on the inception score. ArXiv arXiv:1801.01973
79. Tao M, Tang H, Wu F, Jing X-Y, Bao B-K, Xu C (2022) Df-gan: A simple and effective baseline for text-to-image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 16515–16525
80. Gu S, Chen D, Bao J, Wen F, Zhang B, Chen D, Yuan L, Guo B (2022) Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 10696–10706