3D-TOGO: Towards Text-Guided Cross-Category 3D Object Generation

Zutao Jiang[1, 6 *], **Guansong Lu** [2 *], **Xiaodan Liang** [3, 4], **Jihua Zhu** [1 †], **Wei Zhang** [2], **Xiaojun Chang** [5], **Hang Xu** [2 †]

[1] School of Software Engineering, Xi'an Jiaotong University
[2] Huawei Noah's Ark Lab
[3] Sun Yat-sen University
[4] MBZUAI
[5] ReLER, AAII, University of Technology Sydney
[6] PengCheng Laboratory

taozujiang@gmail.com, luguansong@huawei.com, xdliang328@gmail.com, zhujh@xjtu.edu.cn, wz.zhang@huawei.com, xiaojun.chang@uts.edu.au, chromexbjxh@gmail.com

Abstract

Text-guided 3D object generation aims to generate 3D objects described by user-defined captions, which paves a flexible way to visualize what we imagined. Although some works have been devoted to solving this challenging task, these works either utilize some explicit 3D representations (e.g., mesh), which lack texture and require post-processing for rendering photo-realistic views; or require individual time-consuming optimization for every single case. Here, we make the first attempt to achieve generic text-guided cross-category 3D object generation via a new 3D-TOGO model, which integrates a text-to-views generation module and a views-to-3D generation module. The text-to-views generation module is designed to generate different views of the target 3D object given an input caption. *prior*-guidance, caption-guidance and view contrastive learning are proposed for achieving better view-consistency and caption similarity. Meanwhile, a pixelNeRF model is adopted for the views-to-3D generation module to obtain the implicit 3D neural representation from the previously-generated views. Our 3D-TOGO model generates 3D objects in the form of the neural radiance field with good texture and requires no time-cost optimization for every single caption. Besides, 3D-TOGO can control the category, color and shape of generated 3D objects with the input caption. Extensive experiments on the largest 3D object dataset (i.e., ABO) are conducted to verify that 3D-TOGO can better generate high-quality 3D objects according to the input captions across **98** different categories, in terms of PSNR, SSIM, LPIPS and CLIP-score, compared with text-NeRF and Dreamfields.

Introduction

Automatic 3D object generation has significant application values for many practical application scenarios, including games, movies, virtual reality, etc. In this paper, we study a challenging yet interesting and valuable task, called text-guided 3D object generation. With a text-guided 3D object generation model, one can give a textual description of their wanted 3D object, and leverage such a model to generate the corresponding 3D object, providing a flexible path for visualizing what we imagined.

Along with the success of image generation models (Goodfellow et al. 2014; Vaswani et al. 2017; Ho, Jain, and Abbeel 2020), plenty of works have been devoted for text-to-image generation (Reed et al. 2016; Zhang et al. 2017, 2018a; Xu et al. 2018; Dong et al. 2017; Ramesh et al. 2021; Ding et al. 2021, 2022; Nichol et al. 2021; Ramesh et al. 2022), which shows appealing text-guided generation results. However, there are still few works for text-guided 3D object generation. Some prior works generate 3D shapes from natural language descriptions in the form of meshes (Michel et al. 2022), voxels (Chen et al. 2018), point clouds (Zhou, Du, and Wu 2021), and implicit functions (Liu et al. 2022). While they provide promising results, the issue is that they require tedious post-processing steps, e.g. unwrapping a UV map in Blender, due to the lack of texture when used for multimedia applications. Recently, Neural Radiance Field (NeRF (Mildenhall et al. 2020)) has been successfully applied to the novel view synthesis task. Compared with other 3D representations, neural radiance fields can be sampled at high spatial resolutions and is easy to optimize. Empowered with the visual-language alignment capability of the pre-trained CLIP model, Dreamfields (Jain et al. 2022) leverages a given input text to guide the training of neural radiance fields. The shortcoming of Dreamfields is that it requires individually optimizing a network for each input text, which is time-consuming and computation expensive. Built on the disentangled conditional NeRF (Schwarz et al. 2020) and CLIP model, CLIP-NeRF (Wang et al. 2022) designs two code mappers to edit the shape and color of existing 3D objects with a text or image prompt. However, it only allows editing objects in the same category.

In this paper, we make the first attempt to achieve the generic text-guided **cross-category** 3D object generation and propose our 3D-TOGO model, standing on the progress of text-to-image generation models and Neural Radiance Fields. Our 3D-TOGO model consists of two modules: a) a view-consistent text-to-views generation module that

[*] These authors contributed equally.
[†] Corresponding authors.

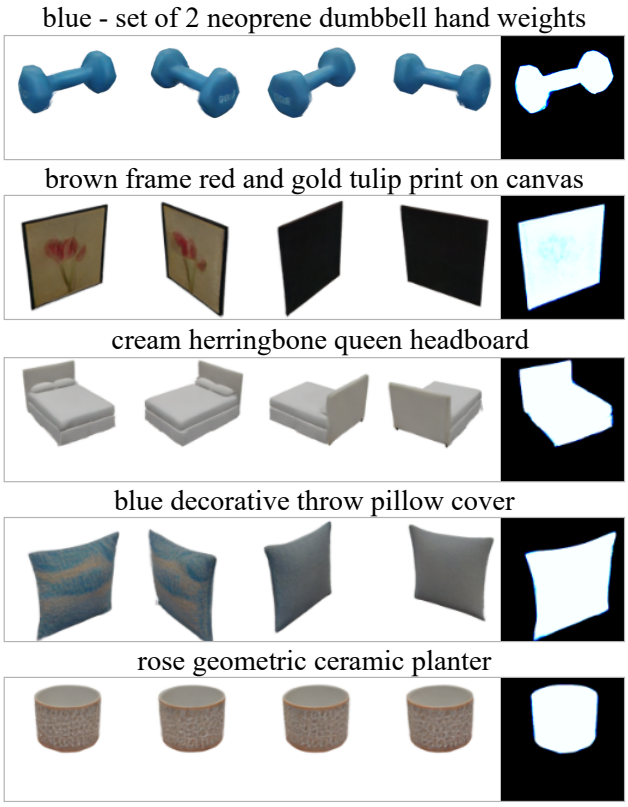

Figure 1: Generation results of our proposed 3D-TOGO model. For each case, we show the input caption, 4 rendered novel views of the generated 3D object and the transmittance from the first view. Transmittance represents how visible a point is from a particular view.

generates views of the target 3D object given an input caption; b) a generic views-to-3D generation module for 3D object generation based on the previously-generated views. Specifically, we adopt the Transformer-based auto-regressive model (Vaswani et al. 2017) for our text-to-views generation module, because of its excellent cross-modal fusion capability. To complement the original token-level cross-entropy loss, we introduce the fine-grained pixel-level supervision signals for better view fidelity. We also incorporate caption-guidance by leveraging the visual-language alignment ability from CLIP (Radford et al. 2021) for better caption similarity. Furthermore, our text-to-views generation module achieves better view-consistency by conditioning on $prior$ view and adopting a novel view contrastive learning method. For the generic views-to-3D generation module, we follow the diagram of the previously-proposed pixelNeRF (Yu et al. 2021). As pixelNeRF optimizes neural radiance fields in the view space of the input image, so we can process each generated view independently and obtain an individual latent intermediate representation for each generated view, which can be aggregated across different views and generate the desired 3D neural representation matched with the input text. Besides, our views-to-3D generation module aims to learn scene prior instead of remembering the training dataset, allowing it to be used for generating objects across different categories.

We perform extensive experiments on the largest 3D object dataset ABO (Collins et al. 2022). Quantitative and qualitative comparisons against baseline methods, including text-NeRF and Dreamfields (Jain et al. 2022), show that our proposed 3D-TOGO model can better generate high-quality 3D objects according to the input captions across different object categories. Compared to baseline methods, the average CLIP-score of our model surpasses **4.4** on randomly selected text inputs, indicating better semantic consistency between the input captions and the generated 3D objects of our model. Besides, results from our 3D-TOGO model show that text-guided 3D object generation allows for flexible control over categories, colors and shapes. Our main contributions are summarized as follows:

- We make the first attempt to resolve the new text-guided cross-category 3D object generation problem and propose 3D-TOGO model, which has an efficient generation process requiring no inference time optimization.
- We propose a text-to-views generation module to generate consistent views given the input captions. We design $prior$-guidance to improve the consistency between adjacent views of a 3D object and introduce view contrastive learning to improve the consistency between different views of a 3D object. Caption-guidance is proposed for better caption similarity. Fine-grained pixel-level supervision is designed for better view fidelity.
- Our 3D-TOGO model can generate high-quality 3D objects across **98** categories. Besides, our 3D-TOGO model is empowered with the ability to control the **category**, **color** and **shape** according to the input caption.

Related Work

Text-to-Image Generation. Text-to-image generation focuses on generating images described by input captions. Based on the progresses on generative models, including generative adversarial networks (GANs (Goodfellow et al. 2014)), auto-regressive model (Vaswani et al. 2017) and diffusion model (Ho, Jain, and Abbeel 2020), there are numbers of works for text-to-image generation. Among them, many GAN-based models are proposed for better visual fidelity and caption similarity (Reed et al. 2016; Zhang et al. 2017, 2018a; Xu et al. 2018; Li et al. 2019a; Dong et al. 2017; Zhu et al. 2019; Tao et al. 2020; Ye et al. 2021). However, GANs suffer from the well-known problem of mode-collapse and unstable training process. Besides GANs, another line of works explore applying Transformer-based auto-regressive model for text-to-image generation (Ramesh et al. 2021; Ding et al. 2021; Esser et al. 2021; Ding et al. 2022; Zhang et al. 2021; Lee et al. 2022). Recent works adopt diffusion model for text-to-image generation (Nichol et al. 2021; Ramesh et al. 2022). However, as the diffusion model predicts the added noise instead of the target images, it is complicated to apply constraints on the generated images. We adopt the architecture of Transformer (Vaswani

et al. 2017) for our view-consistent text-to-views generation module, due to its high cross-modality fusion capability proven in the domain of multi-modal pre-training (Li et al. 2019b; Chen et al. 2020; Wang et al. 2021; Tan and Bansal 2019; Ni et al. 2021; Zhou et al. 2021; Zhuge et al. 2021) and generation mentioned above. Furthermore, we adopt the auto-regressive generation paradigm due to the aforementioned drawbacks of GANs and diffusion model.

Text-Guided 3D Object Generation. Compared with text-to-image generation, it is more challenging to generate 3D objects from the given text description. Some early works generate or edit 3D objects with a pre-trained CLIP model (Radford et al. 2021). Text2Mesh(Michel et al. 2022) edits the style of a 3D mesh by conforming the rendered images to a target text prompt with a CLIP-based semantic loss. It is tailored to a single mesh and could not generate 3D objects from scratch given a text prompt. (Khalid et al. 2022) facilitates zero-shot text-driven mesh generation by deforming from a template mesh guided by CLIP. Text2shape (Chen et al. 2018) generates the voxelized objects using text-conditional Wasserstein GAN (Arjovsky, Chintala, and Bottou 2017), but only allows the 3D object generation of individual category and the performance is limited by the low-resolution 3D representation. CLIP-Forge (Sanghi et al. 2022) models the distribution of shape embeddings conditioned on the image features using a normalizing flow network during the training stage, and then conditions the normalizing flow network with text features to generate a shape embedding during the inference stage, which can be converted into a 3D shape via the shape decoder. However, their generated 3D objects lack color and texture and the quality of generated 3D objects is still limited, which is crucial for practical applications. Recently, (Liu et al. 2022) represents 3D shape with the implicit occupancy representation, which can be used to predict an occupancy field. They design a cyclic loss to encourage the consistency between the generated 3D shape and the input text. However, it cannot generate realistic 3D objects with high fidelity. There are also some works focus on 3D avatar generation and animation from text(Hong et al. 2022; Hu et al. 2021; Canfes et al. 2022), while our work focuses on 3D object generation from text. Compared with the above approaches, our 3D-TOGO model can generate high-quality 3D objects with color and texture across categories and requires no inference time optimization.

Method

Figure 2 shows the framework of our proposed 3D-TOGO model for text-guided 3D object generation. In this section, we will first introduce our view-consistent text-to-views generation module, which takes captions of 3D objects and different camera poses as input, enabling the multi-view image generation. Then we will introduce how to obtain the 3D implicit neural representation of the objects from the previously-generated views.

View-Consistent Text-to-Views Generation Module

Given an input caption t, our text-to-views generation module aims to generate 2D images $\{\hat{x}_i\}_{i=1}^{N}$ of different camera poses $\{\mathbf{P}_i\}_{i=1}^{N}$ for the corresponding 3D object described by caption t, where N is the number of generated views for a 3D object. $\{x_i\}_{i=1}^{N}$ denotes the ground truth images in dataset. The generated images $\hat{x}_i$ need to be consistent with its corresponding input caption t and camera pose $\mathbf{P}_i$. Besides, different views of the same input caption need to be view-consistent, i.e., images of different poses need to be consistent with each other as they are rendered from the same 3D object. For brevity, we omit the subscript of x, $\hat{x}$, and $\mathbf{P}$ in the rest of this section.

Base Text-to-Views Generation Module To generate images of different camera poses given an input caption, we start with designing our base generation module to generate image $\hat{x}$ conditioned on caption t and camera pose $\mathbf{P}$. We adopt the architecture of Transformer (Vaswani et al. 2017) due to its high cross-modality fusion capability (Li et al. 2019b; Wang et al. 2021; Ramesh et al. 2021; Ding et al. 2021; Lee et al. 2022). Following previous works, we transform camera pose $\mathbf{P}$, caption t and image x into sequences of tokens, and train a decoder-only Transformer model with a causal attention mask to predict the sequence of image tokens autoregressively conditioned on camera pose and caption tokens.

Specifically, our Transformer-based image generation module consists of an VQGAN (Esser, Rombach, and Ommer 2021) model, serving as an image tokenizer for quantizing the input image as discrete tokens and recovering the origin image from these discrete tokens, and a Transformer model for fitting the joint distribution of camera pose, caption and image tokens. The autoencoder model consists of an encoder E, a decoder G and a codebook $Z \in \mathbb{R}^{K \times n_z}$ containing K n_z-dimensional codes. Given an image $x \in \mathbb{R}^{H \times W \times 3}$, E first encodes the image into a two-dimensional feature map $\mathbf{F} \in \mathbb{R}^{h \times w \times n_z}$, and then the feature map $\mathbf{F}$ is quantized by replacing each pixel embedding with its closest code within the codebook element-wisely: $\hat{\mathbf{F}}_{ij} = \arg\min_{\mathbf{z}_k} \parallel \mathbf{F}_{ij} - \mathbf{z}_k \parallel^2$. The decoder G is for taking the quantized feature map $\hat{\mathbf{F}}$ as input and reconstructing an pixel-level image $\hat{x}$ close to the original image x, i.e., $\hat{x} \approx x$. With the aforementioned image tokenizer, image x can be tokenized as a sequence of discrete tokens $\{\mathbf{I}_i\}_{i=1}^{N_\mathbf{I}}$. Meanwhile, caption t is encoded into sequence of discrete tokens $\{\mathbf{T}_i\}_{i=1}^{N_\mathbf{T}}$ with a Byte-Pair Encoding (BPE (Sennrich, Haddow, and Birch 2015)) tokenizer. $N_\mathbf{I}$ and $N_\mathbf{T}$ denotes the length of image token sequence and caption token sequence respectively. For camera pose $\mathbf{P}$, we select $N_\mathbf{P}$ poses across different objects so that each camera pose is correspond to an unique token denoted as $\mathbf{V}$. The Transformer is trained to predict the sequence of $[\mathbf{V}, \{\mathbf{T}_i\}_{i=1}^{N_\mathbf{T}}, \{\mathbf{I}_i\}_{i=1}^{N_\mathbf{I}}]$ auto-regressively, which minimizes cross entropy losses applied to the predicted tokens of camera pose, text and image, respectively as follows: $\mathcal{L}_{pose} = CE(\hat{\mathbf{V}}, \mathbf{V})$, $\mathcal{L}_{txt} = \mathbb{E}_i[CE(\hat{\mathbf{T}}_i, \mathbf{T}_i)]$, $\mathcal{L}_{img} = \mathbb{E}_i[CE(\hat{\mathbf{I}}_i, \mathbf{I}_i)]$, where $\hat{\mathbf{V}}$, $\hat{\mathbf{T}}_i$ and $\hat{\mathbf{I}}_i$ are the predicted tokens of camera pose, caption and image respectively; $\mathbb{E}_i[\cdot]$ denotes the expectation and CE represents cross-entropy loss.

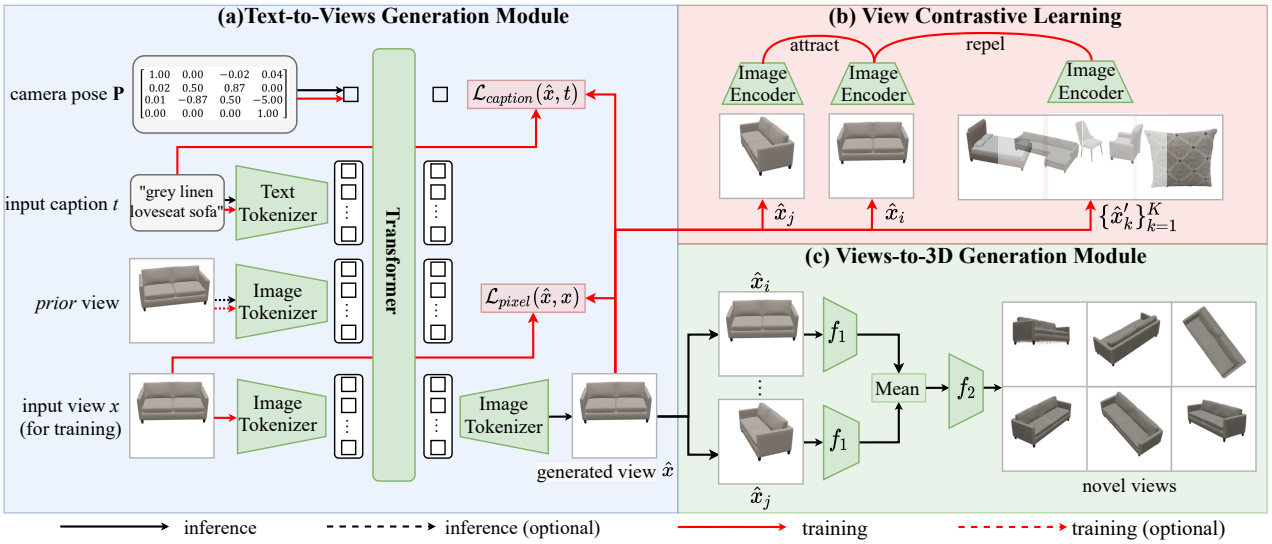

Figure 2: The framework of our 3D-TOGO model for text-guided 3D object generation. (a) Text-to-views generation module. Fine-grained pixel-level supervision signal $\mathcal{L}_{pixel}$ and caption-guidance loss $\mathcal{L}_{caption}$ are for better view fidelity and caption similarity. (b) View contrastive learning for better view-consistency. (c) Views-to-3D generation module. It takes the previously-generated views as input and generates the implicit 3D neural representation, from which novel views can be obtained.

Pixel-Level Supervision and Caption-Guidance One shortage of the aforementioned base text-to-views generation module is that the training loss is applied on image tokens, lacking fine-grained pixel-level supervision signals and leading to low visual quality. To this end, to complement such token-level loss, we explore some pixel-level supervision signals applied on the generated image $\hat{x}$ (decoded from the generated image tokens $\{\hat{\mathbf{I}}_i\}_{i=1}^{N_\mathbf{I}}$ with the image tokenizer) for more fine-grained supervision signals and better image fidelity. We first explore training losses between the original image x and generated image $\hat{x}$. In our preliminary experiments, we tried $L1$ loss and Perceptual loss (Johnson, Alahi, and Fei-Fei 2016). Similar results are observed so that we use the simpler $L1$ loss as: $\mathcal{L}_{pixel} = L1(\hat{x}, x)$. Gradient back-propagation from the generated image $\hat{x}$ to the generated image tokens $\{\hat{\mathbf{I}}_i\}_{i=1}^{N_\mathbf{I}}$ is implemented with straight-through estimator (Bengio, Léonard, and Courville 2013).

Besides, we explore some supervision signals applied between the generated image $\hat{x}$ and input caption t, called caption-guidance, for better caption similarity, i.e., the generated images better match the semantics of the input captions. To this end, we explore leveraging the power of the CLIP model (Radford et al. 2021), which is pre-trained with 400 million image-text pairs collected from the Web and shows excellent zero-shot visual-language alignment capability. Specifically, we utilize the pre-trained ViT-B/32 CLIP model to calculate the similarity score between the generated image $\hat{x}$ and input caption t and apply a caption-guidance loss as: $\mathcal{L}_{caption} = -Sim_{CLIP}(\hat{x}, t)$, to enforce the generation module to generate images that are more semantically similar to the input caption.

***prior*-Guidance and View Contrastive Learning** Until now, it is still challenging for the text-to-views generation module to generate view-consistent images among different camera poses, as it is (1) trained to generate images conditioned on only camera pose and caption, without information from images of other poses, (2) without any view-consistency supervision signal during training.

To improve the consistency between adjacent views, we first propose to condition the generation module on a piece of extra information: an image of another camera pose, which we call *prior* view. To this end, we specify a fixed order of different camera poses and condition image generation of the current camera pose on the previous one. During training, such *prior* view is masked half of the time so that the generation module is able to perform image generation with and without *prior* view. During inference, given an input caption t, the first view is generated without *prior* view, while the others are generated using the previously generated one as *prior* view one by one in order. Tokens of input *prior* view and predicted *prior* view are denoted as $\{\mathbf{I}_i^{prior}\}_{i=1}^{N_\mathbf{I}}$ and $\{\hat{\mathbf{I}}_i^{prior}\}_{i=1}^{N_\mathbf{I}}$ respectively. A reconstruction loss is also applied on the predicted *prior* view tokens as: $\mathcal{L}_{prior} = \mathbb{E}_i\, CE(\hat{\mathbf{I}}_i^{prior}, \mathbf{I}_i^{prior})$.

Furthermore, we incorporate the concept of contrastive learning for better view-consistency. In our case, as we can see, different views of the corresponding object of an input caption should be closer to each other, than to views of the corresponding object of a different input caption. This is the same as the objective of contrastive learning. To this end, we propose view contrastive learning, where views of the same object are treated as positive samples of each other, while views of different objects are treated as negative samples of each other. During training, we generate two different views $\hat{x}_i, \hat{x}_j, i \neq j$ of the same object, and a set of K views $X = \{\hat{x}'_1, \hat{x}'_2, ..., \hat{x}'_K\}$ of different objects. Besides, we learn an image encoder f_{enc} for extracting view representations $f_{enc}(x)$. Then the objective function of view con-

trastive learning can be formulated as follows:

$$\mathcal{L}_{contrastive} = -\log \frac{\exp(sim(f_{enc}(\hat{x}_i), f_{enc}(\hat{x}_j))/\tau)}{\sum_{x \in X} \exp(sim(f_{enc}(\hat{x}_i), f_{enc}(x))/\tau)}, \tag{1}$$

where $sim(\cdot, \cdot)$ denotes cosine similarity and τ denotes a temperature parameter.

Finally, the overall objective function of our view-consistent text-to-views generation module is as follows:

$$\begin{aligned} \mathcal{L} =& \lambda_{pose}\mathcal{L}_{pose} + \lambda_{txt}\mathcal{L}_{txt} \\ &+ \lambda_{prior}\mathcal{L}_{prior} + \lambda_{img}\mathcal{L}_{img} + \lambda_{pixel}\mathcal{L}_{pixel} \\ &+ \lambda_{caption}\mathcal{L}_{caption} + \lambda_{contrastive}\mathcal{L}_{contrastive}, \end{aligned} \tag{2}$$

where λ_{pose}, λ_{txt}, λ_{prior}, λ_{img}, λ_{pixel}, $\lambda_{caption}$ and $\lambda_{contrastive}$ are the balancing coefficients.

Views-to-3D Generation Module

Given images $\hat{\mathscr{I}} = \{\hat{x}_i\}_{i=1}^{N}$ generated by the text-to-views module, the aim of views-to-3D module is to obtain the implicit neural representation of the generated 3D object, where N is the number of generated images. In experiments, we find that NeRF (Mildenhall et al. 2020) fails to obtain high-quality novel view synthesis results in some cases, if we naively optimize NeRF with the generated images. This is because that there are still some small inconsistent contents among the generated images. Therefore, we introduce pixelNeRF (Yu et al. 2021) to firstly learn scene prior from the ground-truth images $\mathscr{I} = \{x_i\}_{i=1}^{M}$ across objects in the training data, where $M (M \geq N)$ is the number of rendered images of 3D objects. Please refer to (Yu et al. 2021) regarding the network architecture details of pixelNeRF model.

Once obtaining the scene prior, we can encode the generated 3D object as a continuous volumetric radiance field f of color c and density σ by using the generated multi-view images $\hat{\mathscr{I}}$. Similar to pixelNeRF, we use the view space of the generated images instead of the canonical space. Specially, for a 3D query point y in the neural radiance fields, we first retrieve the corresponding image features from $\hat{\mathscr{I}}$ by $\{w_i\}_{i=1}^{N} = \{\mathbf{W}_i(\pi_i(y))\}_{i=1}^{N}$, where $\mathbf{W}_i = E(\hat{x}_i)$ is the feature volume extracted from the generated image $\hat{x}_i$, $\pi_i(y)$ denotes the corresponding image coordinate on the image plane of the generated image $\hat{x}_i$, and $\mathrm{W}_i(\pi_i(y))$ represents the image feature extracted from feature volume W_i for the 3D query point y.

Then, we need to obtain the intermediate representation $U = \{U_i\}_{i=1}^{N}$ in each view space of the generated images $\hat{\mathscr{I}}$ for query point y with view direction d as follows:

$$U_i = f_1(\gamma(\mathbf{H}(y)_i), d_i; w_i), \tag{3}$$

where $\mathbf{H}(y)_i = \mathbf{P}_i y = \mathbf{R}_i y + h_i$ denotes that transforming the query point y into the coordinate system of the generated image $\hat{x}_i$, $\mathbf{P}_i$ is the world to camera transformation matrix, $\mathbf{R}_i$ is the rotation matrix, h_i is the translation vector, $\gamma(\cdot)$ represents a positional encoding on the transformed query point $\mathbf{H}(y)_i$ with 6 exponentially increasing frequencies (Mildenhall et al. 2020); $d_i = \mathbf{R}_i d$ denotes transforming the view direction d into the coordinate system of the generate image $\hat{x}_i$; w_i is the corresponding image feature extracted from the generate image $\hat{x}_i$; $f_1(\cdot)$ represents the layers of the pixelNeRF to process transformed query point, transformed view direction and the corresponding extracted image feature in the view space of the generated images $\hat{\mathscr{I}}$ independently, which has been trained on the ABO training dataset.

After obtaining all the intermediate representation $U = \{U_i\}_{i=1}^{N}$, we use the average pooling operator η to aggregate them and then pass the layers of the pixelNeRF to process the aggregated representation. This process can be written as:

$$f(y, d) = (\sigma(y), c(y, d)) = f_2(\eta(U_i))_{i=1}^{N}, \tag{4}$$

where f_2 denotes the layers to process the aggregated representation $\eta(U_i)_{i=1}^{N}$, $\sigma(y)$ is the density of the 3D query point y which is independent of the view direction d, $c(y, d)$ represents the color of the 3D query point y in the view direction d, $f(\cdot)$ is the final continuous volumetric radiance field which representing the generated 3D object matched with the input caption t. For the photo-realistic rendering of the generated 3D object, we use the volume rendering technique proposed in (Mildenhall et al. 2020).

Experiments

Dataset. Our approach is evaluated on Amazon-Berkeley Objects (ABO) (Collins et al. 2022), a large-scale dataset containing nearly **8,000** real household objects from **98** categories with their corresponding nature language descriptions. We use ABO dataset because it contains the categories of other small datasets, such as ShapeNet(Chen et al. 2018). Benefiting from their detailed texture and non-lambertian BRDFs, the 3D models in ABO can be photo-realistically rendered. To construct multi-view images dataset with their nature language descriptions, we use Blender to render each 3D model into 256×256 RGB-alpha images from 36 cameras. Camera elevation is set as $-30°$ and camera azimuth is sampled uniformly from the range $[-180°, 180°]$. Totally, 286,308 multi-view images are rendered from 7,953 objects belong to 98 categories. We randomly split 80%, 10%, 10% objects as our training, validation, and test set, respectively.

Metrics. Following the settings of (Yu et al. 2021), we evaluate the quality of our generated 3D objects by measuring the quality of novel view synthesis. Specifically, PSNR, SSIM(Wang et al. 2004), LPIPS (Zhang et al. 2018b) are adopted as our metrics. We also compute the CLIP score (Radford et al. 2021) between the rendered novel view images and the corresponding natural language description, which can measure the semantic consistency of the generated 3D objects with the description. The ResNet-50 CLIP model is adopted. Besides, we also adopt human evaluation for comparison. Two metrics are considered: **object fidelity** (including view fidelity and consistency among different views) and **caption similarity**. For each input caption, results from different methods are shown in random order and the workers are asked to order different results in terms of these two metrics. The average rank from different workers is used as the final score. 94 human evaluation results are collected. The higher PSNR, SSIM and CLIP score, the better; the lower LPIPS, object fidelity and caption similarity,

Metric	Method	text1	text2	text3	text4	text5	text6	12 Texts Avg.
PSNR ↑	text-NeRF	18.15	19.96	0.79	22.02	18.12	0.48	14.04
	Ours	**20.12**	**23.34**	**26.41**	**24.02**	**23.80**	**26.53**	**24.98**
SSIM ↑	text-NeRF	0.856	0.898	0.001	0.876	0.863	0.001	0.636
	Ours	**0.889**	**0.920**	**0.925**	**0.898**	**0.897**	**0.864**	**0.900**
LPIPS ↓	text-NeRF	0.138	0.091	0.503	0.142	0.167	0.475	0.239
	Ours	**0.122**	**0.063**	**0.075**	**0.128**	**0.165**	**0.082**	**0.092**
CLIP- Score ↑	Dreamfields	18.92	21.34	18.13	16.31	18.08	18.73	18.40
	text-NeRF	26.65	20.28	13.90	21.12	23.16	11.84	18.38
	Ours	**27.08**	**22.67**	**20.98**	**22.74**	**24.13**	**25.08**	**22.84**
Object Fidelity ↓	Dreamfields	3.00	2.56	2.04	2.94	2.94	1.94	2.63
	text-NeRF	1.97	2.32	2.93	2.04	1.97	2.95	2.27
	Ours	**1.03**	**1.12**	**1.03**	**1.02**	**1.09**	**1.12**	**1.11**
Caption Similarity ↓	Dreamfields	2.97	2.66	2.02	2.94	2.87	2.01	2.64
	text-NeRF	1.98	2.19	2.95	2.03	1.97	2.95	2.25
	Ours	**1.05**	**1.15**	**1.03**	**1.03**	**1.16**	**1.04**	**1.11**

Table 1: Quantitative comparison against text-NeRF (short for text-to-views generation + NeRF(Mildenhall et al. 2020)) and Dreamfields (Jain et al. 2022).

the better.

Experimental Setup. We implement our algorithm with Pytorch. The hyper-parameters of λ_{pose}, λ_{txt}, λ_{prior}, λ_{img}, λ_{pixel}, $\lambda_{caption}$ and $\lambda_{contrastive}$ are set to 0.1, 0.1, 0.1, 0.6, 1, 1 and 1 respectively. For our text-to-views generation module, we use AdamW optimizer to train 20 epochs. For the views-to-3D generation module, we use Adam optimizer to train 100 epochs and randomly select 9 views during each training step. More details are provided in the Appendix.

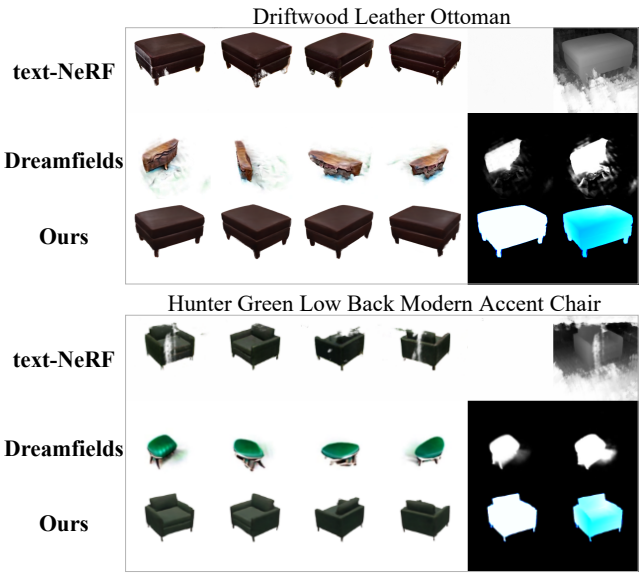

Figure 3: Visual comparison against two baseline methods. For each sub figure, the textual title is the input caption, and the first 4 images are rendered novel views of the generated 3D object while the last two images are transmittance and depth from the first view respectively.

Comparison Against Baselines

As our 3D-TOGO model generates objects in the form of neural radiance fields, so we select two NeRF-based text-guided 3D object generation methods as our baseline: text-to-views generation module + NeRF (Mildenhall et al. 2020) (called text-NeRF for convenient in the following) and Dreamfields (Jain et al. 2022). We use the code open-sourced by the authors. As both text-NeRF and Dreamfields require training an individual network for each given natural language description, we randomly select 12 text descriptions from our test set as the input captions. The selected text descriptions are included in the Appendix. As there is no ground truth for the generated views of Dreamfields, we do not use PSNR, SSIM and LPIPS in the comparison against Dreamfields.

Table 1 shows the quantitative comparison among different methods. Results of the first 6 descriptions and the average of 12 descriptions are shown while results of the rest 6 text descriptions are included in the Appendix. As we can see, for all cases, 3D-TOGO achieves the best results. Additionally, Figure 3 shows the qualitative comparison among different methods. Results of the first 2 descriptions are shown while results of the rest 10 descriptions are included in the Appendix. As we can see, text-NeRF generates broken objects, as there are still some small inconsistent contents among the generated images. Figure 3 shows that Dreamfields cannot generate reasonable results. This may be because Dreamfields cannot generalize to household objects or it requires attentive hyperparameter-tuning. Besides, both text-NeRF and Dreamfields require time-consuming per-case optimization, while 3D-TOGO can be used across objects.

Text-Guided 3D Object Generation

Figure 1 shows the cross-category text-guided 3D generation results of our proposed 3D-TOGO model. Our model gen-

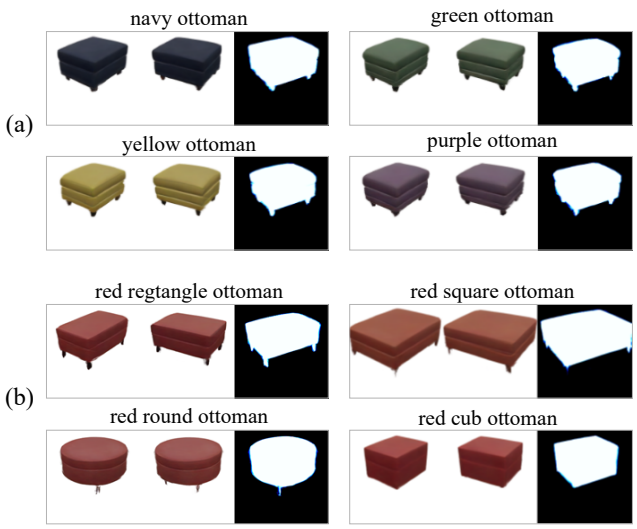

Figure 4: 3D object generation results with controlled (a) color and (b) shape. For each case, we show the input caption, 2 rendered novel views of the generated 3D object and the transmittance from the first view.

prior	contrastive	caption	L1	consistency-error ↓
				9.47
✓				8.88
	✓			9.17
		✓		9.23
			✓	9.35
✓	✓			8.61
✓		✓		8.81
✓		✓	✓	8.74
✓	✓	✓	✓	**8.56**
	GT			7.55

Table 2: Ablation study of our text-to-views generation module. '*prior*', 'contrastive', 'caption' and 'L1' indicates *prior*-guidance, view contrastive learning, caption-guidance and pixel-level L1 loss respectively.

erates high-fidelity 3D objects matching the input caption across different object categories. Besides, Figure 4 shows that our method can control the color and shape of the generated 3D objects by the input caption flexibly. More results are included in the Appendix.

Ablation Study

In this section, we study the effectiveness of different objectives on the quality of generated views from the text-to-views generation module and 3D objects from the views-to-3D generation module respectively. For quantitative comparison of the quality of generated views, we adopt metrics including FID (Heusel et al. 2017), KID (Bińkowski et al. 2018), CLIP score and consistency error. Consistency error measures the average L2 error between views of adjacent camera poses and reflects view consistency to some degree. The lower the consistency error, the better view consistency. Results of consistency error are shown in Table 2 while results of FID, KID, and CLIP-score are included in the Appendix.

prior	contrastive	caption	L1	CLIP-score ↑
				19.91
✓				20.50
	✓			19.97
		✓		20.14
			✓	19.99
✓	✓			20.54
✓		✓		20.90
✓		✓	✓	20.93
✓	✓	✓	✓	**21.01**

Table 3: Ablation study of our text-to-views generation module and the effectiveness on 3D object generation.

View Generation Quality. Table 2 shows the quantitative results of our text-to-views generation module. As we can see, *prior*-guidance improves the consistency-error from 9.47 to 8.88, and view contrastive learning further improves it to 8.61, indicating both of these two improvements contribute to improving view-consistency. Besides, our complete text-to-views generation module achieves the best consistency-error of 8.56.

3D Object Generation Quality. Table 3 shows the CLIP-scores of our views-to-3D generation module for different object categories. The detailed results for each category will be provided in the Appendix. As we can see, 3D object generation based on the results of our complete text-to-views generation module achieves the best CLIP-score in all shown categories and achieves the best average CLIP-score among all categories of the ABO test set.

Conclusion

In this paper, we propose the 3D-TOGO model for the first attempt to achieve the generic text-guided 3D object generation. Our 3D-TOGO integrates a view-consistent text-to-views generation module for generating views of the target 3D object given an input caption; and a generic cross-scene neural rendering module for 3D object generation. For the text-to-views generation module, we adopt fine-grained pixel-level supervision signals, *prior*-guidance, caption-guidance and view contrastive learning for achieving better view fidelity, view-consistency and caption similarity. A pixelNeRF model is adopted for the generic implicit 3D neural representation synthesis module. Extensive experiments on the largest 3D object dataset ABO show that our proposed 3D-TOGO model can better generate high-quality 3D objects according to the input captions across 98 different object categories both quantitatively and qualitatively, compared against text-NeRF and Dreamfields (Jain et al. 2022). Our 3D-TOGO model also allows for flexible control over categories, colors and shapes with the input caption. We describe the potential negative societal impacts and limitations of our work in the Appendix.

Acknowledgments

This work is supported in part by the Key Research and Development Program of Shaanxi Province under Grants. 2021GXLH-Z-097 and 2021GY-025, the Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant No. 22lgqb38, and the Guangdong Outstanding Youth Fund (Grant No. 2021B1515020061). We thank MindSpore for the partial support of this work, which is a new deep learning computing framework[1].

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Bengio, Y.; Léonard, N.; and Courville, A. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Canfes, Z.; Atasoy, M. F.; Dirik, A.; and Yanardag, P. 2022. Text and Image Guided 3D Avatar Generation and Manipulation. *arXiv preprint arXiv:2202.06079*.

Chen, K.; Choy, C. B.; Savva, M.; Chang, A. X.; Funkhouser, T.; and Savarese, S. 2018. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian conference on computer vision*, 100–116. Springer.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Collins, J.; Goel, S.; Deng, K.; Luthra, A.; Xu, L.; Gundogdu, E.; Zhang, X.; Vicente, T. F. Y.; Dideriksen, T.; Arora, H.; et al. 2022. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21126–21136.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835.

Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers. *arXiv preprint arXiv:2204.14217*.

Dong, H.; Yu, S.; Wu, C.; and Guo, Y. 2017. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 5706–5714.

Esser, P.; Rombach, R.; Blattmann, A.; and Ommer, B. 2021. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34: 3518–3532.

[1]https://www.mindspore.cn/

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12873–12883.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; and Liu, Z. 2022. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *ACM Transactions on Graphics (TOG)*, 41(4): 1–19.

Hu, L.; Qi, J.; Zhang, B.; Pan, P.; and Xu, Y. 2021. Text-driven 3D Avatar Animation with Emotional and Expressive Behaviors. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2816–2818.

Jain, A.; Mildenhall, B.; Barron, J. T.; Abbeel, P.; and Poole, B. 2022. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 867–876.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Khalid, N.; Xie, T.; Belilovsky, E.; and Popa, T. 2022. Text to Mesh Without 3D Supervision Using Limit Subdivision. *arXiv preprint arXiv:2203.13333*.

Lee, D.; Kim, C.; Kim, S.; Cho, M.; and Han, W.-S. 2022. Autoregressive Image Generation using Residual Quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11523–11532.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. H. 2019a. Controllable text-to-image generation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2065–2075.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019b. Visualbert: A simple and performant baseline for vision and language. Preprint arXiv:1908.03557.

Liu, Z.; Wang, Y.; Qi, X.; and Fu, C.-W. 2022. Towards Implicit Text-Guided 3D Shape Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17896–17906.

Michel, O.; Bar-On, R.; Liu, R.; Benaim, S.; and Hanocka, R. 2022. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13492–13502.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 405–421. Springer.

Ni, M.; Huang, H.; Su, L.; Cui, E.; Bharti, T.; Wang, L.; Zhang, D.; and Duan, N. 2021. M3p: Learning universal representations via multitask multilingual multimodal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3977–3986.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, 1060–1069. PMLR.

Sanghi, A.; Chu, H.; Lambourne, J. G.; Wang, Y.; Cheng, C.-Y.; Fumero, M.; and Malekshan, K. R. 2022. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18603–18613.

Schwarz, K.; Liao, Y.; Niemeyer, M.; and Geiger, A. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33: 20154–20166.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Tao, M.; Tang, H.; Wu, S.; Sebe, N.; Jing, X.-Y.; Wu, F.; and Bao, B. 2020. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2022. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3835–3844.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. Preprint arXiv:2108.10904.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.

Ye, H.; Yang, X.; Takac, M.; Sunderraman, R.; and Ji, S. 2021. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4578–4587.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 5907–5915.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2018a. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1947–1962.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018b. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Z.; Ma, J.; Zhou, C.; Men, R.; Li, Z.; Ding, M.; Tang, J.; Zhou, J.; and Yang, H. 2021. M6-UFC: Unifying Multi-Modal Controls for Conditional Image Synthesis. *arXiv preprint arXiv:2105.14211*.

Zhou, L.; Du, Y.; and Wu, J. 2021. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5826–5835.

Zhou, M.; Zhou, L.; Wang, S.; Cheng, Y.; Li, L.; Yu, Z.; and Liu, J. 2021. UC2: Universal Cross-lingual Cross-modal Vision-and-Language Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4155–4165.

Zhu, M.; Pan, P.; Chen, W.; and Yang, Y. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5802–5810.

Zhuge, M.; Gao, D.; Fan, D.-P.; Jin, L.; Chen, B.; Zhou, H.; Qiu, M.; and Shao, L. 2021. Kaleido-BERT: Vision-Language Pre-training on Fashion Domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12647–12657.