TeMO: Towards Text-Driven 3D Stylization for Multi-Object Meshes

Xuying Zhang[1], Bo-Wen Yin[1], Yuming Chen[1], Zheng Lin[1], Yunheng Li[1], Qibin Hou[2,1]*, Ming-Ming Cheng[2,1]

[1]VCIP, College of Computer Science, Nankai University
[2]NKIARI, Shenzhen Futian

https://github.com/zhangxuying1004/TeMO

Abstract

Recent progress in the text-driven 3D stylization of a single object has been considerably promoted by CLIP-based methods. However, the stylization of multi-object 3D scenes is still impeded in that the image-text pairs used for pre-training CLIP mostly consist of an object. Meanwhile, the local details of multiple objects may be susceptible to omission due to the existing supervision manner primarily relying on coarse-grained contrast of image-text pairs. To overcome these challenges, we present a novel framework, dubbed ***TeMO****, to parse multi-object 3D scenes and edit their styles under the contrast supervision at multiple levels. We first propose a Decoupled Graph Attention (****DGA****) module to distinguishably reinforce the features of 3D surface points. Particularly, a cross-modal graph is constructed to align the object points accurately and noun phrases decoupled from the 3D mesh and textual description. Then, we develop a Cross-Grained Contrast (****CGC****) supervision system, where a fine-grained loss between the words in the textual description and the randomly rendered images are constructed to complement the coarse-grained loss. Extensive experiments show that our method can synthesize high-quality stylized content and outperform the existing methods over a wide range of multi-object 3D meshes.*

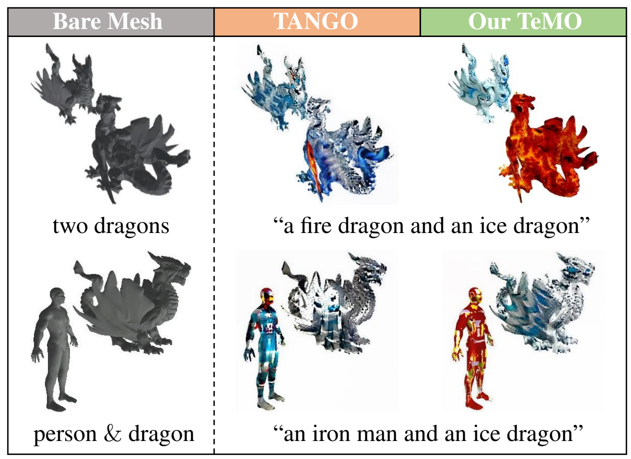


Figure 1. Visual comparisons between the existing 3D stylization methods (*e.g.,* TANGO [15]) and our TeMO in multi-object scenes. For a scene with multiple objects of the same/different categories, existing methods are prone to interference between different properties of the objects, while our TeMO is able to accurately synthesize the desired stylized content for each object.

1. Introduction

3D asset creation through stylization aims to synthesize stylized content on the bare meshes to conform to the given text descriptions [15, 19, 25], referring images [38, 50], or 3D shapes [34, 44]. This research plays an important role in a wide spectrum of applications, *e.g.,* virtual/augmented reality [5, 9], gaming industries [49], and robotics [13]. Moreover, it also presents considerable potential and has attracted increasing attention in computer vision and graphics communities. Considering the ready availability and expressiveness of text prompts as well as the popularity of large-scale Contrastive Language-Image Pre-training (CLIP) [28] model, we choose to work with text-driven 3D stylization.

Recent years have witnessed the emergence of a series of impressive works [8, 15, 23, 25], aiming to drive the advancement of text-driven 3D stylization. Existing methods usually adopt multi-layer perceptrons (MLPs) to predict the location attribute displacements of the bare mesh under the supervision of the contrastive loss in CLIP. We observe that these works focus on the stylization of a single 3D object and perform poorly on multiple objects, as shown in the second column of Fig. 1. We argue that two inherent characteristics of CLIP result in this issue: **i)** CLIP is mainly pre-trained with image-text pairs mostly consisting of a single object; **ii)** CLIP loss employs global representation vectors from images and text to coarsely match these two modalities, which inevitably causes the loss of local details. More-

*Qibin Hou is the corresponding author.

over, the key to synthesizing desired styles for multiple 3D objects lies in the parsing of such 3D scenes and the multi-grained supervision for details refinement.

To simultaneously generate stylized content for multiple 3D objects, the primary step is to achieve accurate alignment between the objects in the 3D mesh and the target text. However, existing methods employ global semantics of the text to stylize a single object, which inevitably produces noises when stylizing the objects in multi-object scenes. To overcome this challenge, we propose to parse the 3D scene by introducing a Decoupled Graph Attention (DGA) module. Specifically, all noun phrases are decoupled from the text prompt and the mesh surface points of the current view are divided into several clusters as well. Then, a cross-modal graph is constructed to establish connections of the noun phrases to their corresponding object points while distancing them from the irrelevant ones. This graph enables the accurate interaction between two interrelated modalities. Finally, the surface point features of 3D objects can be reinforced by independent cross-attention fusions with their neighboring word nodes in the graph architecture.

Furthermore, we also design a Cross-Grained Contrast (CGC) loss to perform comprehensive cross-modal supervision for the stylization of multiple objects. The goal is to guide the network to generate more stylization details for multiple 3D objects to match the target text. Our loss consists of two parts, *i.e.,* coarse-grained contrast and fine-grained contrast. In the former part, the text prompt is regarded as sentence-level supervision, which calculates the similarity between the 2D views rendered from the stylized 3D mesh and the text prompt using the global feature vectors from the CLIP model. In the latter part, we see the text prompt from the word level, and consider the similarities between each word of the sentence and the rendered images of the view sets. To be specific, we produce the word representations of the text prompt by taking the hidden states from the text encoder of CLIP. Motivated by the recent process in video-text retrieval [22], we calculate fine-grained loss via the weighted summation of the element in similarity vectors based on the importance of each word or image.

Based on the well-designed DGA module and CGC loss, we propose a novel framework towards **Te**xt-Driven 3D stylization for **M**ulti-**O**bject Meshes, called TeMO. To validate the effectiveness of our TeMO, extensive experiments are conducted on various multi-object 3D scenes, as shown in the 3^{rd} column of Fig. 1. The experimental results demonstrate our TeMO is less susceptible to interference from multiple objects and can generate superior stylized assets compared with the existing 3D stylization methods.

Our contributions can be summarized as follows:

- We present a new 3D stylization framework, called TeMO. To the best of our knowledge, it is the first attempt to parse the objects in the text and 3D meshes and generate stylizations for multi-object scenes.
- We propose a Decoupled Graph Attention (DGA) module, which constructs a graph structure to align the surface points in the multi-object mesh and the noun phrases in the text prompt.
- We design a Cross-Grained Contrast (CGC) loss, in which the text is contrasted with the rendered images from sentence and word levels.

2. Related Work

2.1. Text-Driven 3D Manipulation

Generating or editing 3D content according to a given prompt is a long-standing objective in computer vision and graphics [2, 39, 42]. Among all forms of the prompt, the text has garnered the most conspicuous gaze due to three reasons: i) Text descriptions are readily accessible from the existing corpus; ii) Text descriptions are particularly user-friendly since they are easily modifiable and can effectively express complex concepts related to stylizations; iii) The popularity of large-scale multimodal models [16, 28] has made achieving visual-language supervision possible.

Text2Mesh [25] proposes a neural-style field network to predict the color and displacement of mesh vertices. TANGO [15] proposes to disentangle the appearance style as the spatially varying bidirectional reflectance distribution, the local geometric variation, and the lighting condition. Then, X-Mesh [23] integrates the target text guidance by utilizing text-relevant spatial and channel-wise attention during vertex feature extraction. Motivated by the remarkable progress in text-driven 2D generation [29, 31], TEXTure [30] and Text2Tex [7] incorporate a pre-trained depth-aware image diffusion model to synthesize high-resolution partial textures from multiple viewpoints progressively.

To make full use of the priors in the pre-trained 2D text-to-image diffusion model, DreamFusion [27] introduces a Score Distillation Sample (SDS) loss to perform text-to-3D synthesis. With the help of SDS loss, Latent-NeRF [24] and Fantasia3D [8] can generate 3D shapes and appearances for 3D objects. Despite achieving impressive results, these methods focus on the stylization of a single 3D object and rarely explore multi-object scenes. CLIP-Mesh [26] attempts to generate multiple 3D objects for target text. Nevertheless, the resulting content is not satisfactory. In this paper, we parse the objects described in rendered images and text prompts, aligned by two well-designed strategies.

2.2. Attention Mechanism

The concept of the attention mechanism was initially introduced in neural machine translation [1], where the weighted summation of the candidate vectors is calculated according to their importance scores. This technology has been extended to a myriad of tasks, *e.g.,* natural language pro-

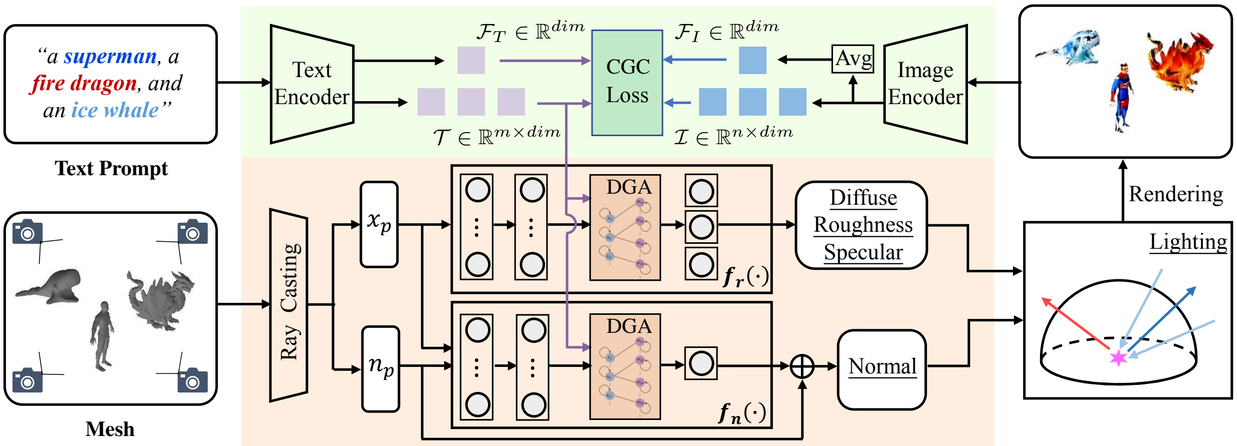


Figure 2. The overall architecture of the proposed TeMO framework. We first specify several cameras to cast rays toward the objects in the 3D mesh scene. Then, a surface point x_p and normal n_p can be attained from each ray intersected with the objects. These points and normals are fed to the attribute prediction network where the features of 3D objects are parsed and interacted with the decoupled text features via our proposed DGA module. Meanwhile, we employ a series of spherical Gaussians to represent the lighting. Finally, a differentiable SG render is adopted to render images, which are utilized to contrast with the text prompt by our designed CGC loss.

cessing [10, 21, 37], computer vision [14, 17, 47, 54], and multi-modal learning [20, 41, 48, 53]. For instance, Transformer [37] employs the self-attention operation to establish connections between words within a sentence and utilize the cross-attention mechanism to align source and target sentences. Non-local network [40] takes the lead in introducing self-attention to computer vision and achieves great success in video understanding and object detection. ViT [11] treats an image as a sequence of patches and employs a Transformer encoder based on self-attention to perform image classification. Swin Transformer [17] introduces shifted windows to enhance the local perception ability of self-attention. More recently, X-Mesh [23] designs a text-guided dynamic attention mechanism for vertex feature extraction of a 3D object. However, this guidance only relies on a text feature vector without considering the parsing of text and 3D scenes. In this paper, the multiple objects decoupled from the target text and 3D mesh are aligned via a cross-modal graph to achieve precise guidance.

2.3. Multi-modal Contrastive Learning

Contrastive learning has become an increasingly popular research topic in the multi-modal community due to its ability to align different modal representations. Based on this strategy, CLIP [28] is pre-trained on an abundance of image-text pairs, achieving great success in cross-modal supervision. TACo [45] presents a token-aware cascade contrastive learning based on the syntactic classes of words to achieve fine-grained semantic alignment in text-video retrieval. Concurrently, FILIP [46] proposes comparing the image patches with the words in the sentence. Regarding the text-driven 3D stylization, the CLIP loss, which calculates the similarity between the image and text vectors in the embedding space of CLIP, is adopted by the vast majority of methods. Although achieving impressive results in stylizing a single object, these methods cannot be well adapted to scenes with multiple 3D objects. We argue an important reason for this issue is the loss of local details caused by such coarse-grained supervision. In this paper, we propose a cross-grained supervision strategy, which considers fine-grained and coarse-grained contrasts to achieve a more precise semantic alignment between rendered image and text.

3. Methodology

3.1. Overall Architecture

Fig. 2 shows the end-to-end architecture of our TeMO framework. Given a bare mesh and a text prompt containing multiple objects, the TeMO aims to synthesize stylization on the mesh to match the text descriptions. We employ a set of vertices $V \in \mathbb{R}^{e \times 3}$ and faces $F \in \{1, ..., e\}^{u \times 1}$ to explicitly define the input triangle mesh, which is fixed throughout the training. Following TANGO [15], we disentangle the appearance style as the spatially varying bidirectional reflectance distribution function [4, 51, 52] (including diffuse, roughness, specular), the local geometric variation (normal map), and the lighting condition.

We start by normalizing the vertex coordinates to lie inside a unit sphere. Then, we randomly sample points around the mesh using Gaussian distribution as camera positions to render images. Next, we can obtain a camera ray $R_p = \{c + t\nu_p\}$ from the sampled camera position c and a pixel p in rendered images, where ν_p is the direction of the ray. Further, ray casting [32] is used to seek out the ray and mesh's first intersection point and intersection face. Moreover, the normal $n_p \in \mathbb{R}^3$ of the intersection face is

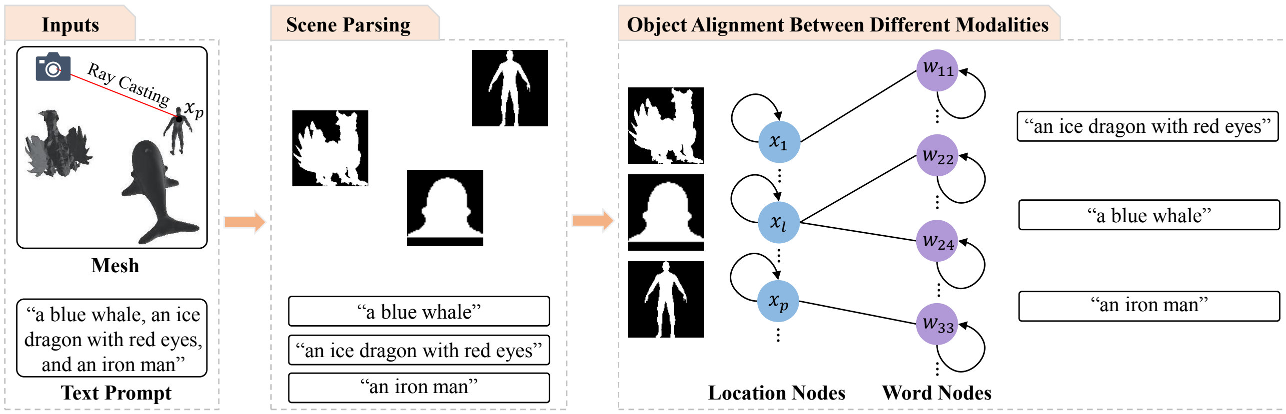


Figure 3. Construction pipeline of the cross-modal graph architecture in our DGA module. Note that x_p, the surface point of the 3D objects, and w_{ij}, the j-th word in the i-th noun phrase, are connected together only if they correspond to the same object.

employed as surface normal at the point $x_p \in \mathbb{R}^3$.

To achieve multiview-consistent features, our TeMO is restricted to predicting the normal displacement as a function of the location, while allowing the color materials to be predicted as a function of both location and viewing direction. Therefore, our TeMO represented as MLPs includes two branches, *i.e.,* normal branch $f_n(\cdot)$ and reflectance branch $f_r(\cdot)$. Specifically, the former is utilized as the prediction of normal offset on the point x_p, and the latter is designed to predict surface reflectance coefficients of the material at the location x_p, *i.e.,* diffuse, roughness, and specular. To synthesize high-frequency details, we also apply the Fourier positional encoding [35] to every input. In addition, the spherical Gaussian is employed to represent each light intensity $L_i(\cdot)$ due to its closed-form nature and analytical solution. Based on the attained geometric and appearance components, each pixel color in the rendered image can be calculated by a hemisphere renderer [15]:

$$L_p(\nu_p, x_p, n_p) = \int_{\Omega} L_i(w_i) f_r(\nu_p, w_i, x_p)(w_i \cdot \hat{n}_p) dw_i, \quad (1)$$

$$\hat{n}_p = n_p + f_n(x_p, n_p), \quad (2)$$

where $\Omega = \{w_i : w_i \cdot \hat{n}_p \geq 0\}$ represents the hemisphere, w_i is the incident light direction, and $\hat{n}_p$ is the estimated normal on surface point x_p.

3.2. Decoupled Graph Attention

To achieve text-drive stylization for multiple 3D objects, the key issue that needs to be solved is the accurate alignment between the objects described in the text and those in the meshes. X-Mesh [23] has incorporated text-guided dynamic linear layers, in which the global representation vector of the target object in the text is utilized as guidance to acquire text-aware vertex features. Nevertheless, the global vector containing information about multiple objects is prone to mutual interference and produces semantic noises during guidance for multi-object scenes.

To address this challenge, we propose to parse the objects in the text and mesh. We first extract the noun phrases modified by adjectives or prepositional phrases from the text using the NLTK tools [3]. Then, we employ the Gaussian Mixture Model (GMM) [56] to cluster the intersection point set $\{x_1, ..., x_p, ...\}$ of the current ray and the mesh. Meanwhile, we can obtain a binary map of objects in the current view based on whether the ray intersects with the mesh. Further, we can decouple the objects in the binary map according to the clustered points and acquire several binary maps of individual objects. Based on the disentangled noun phrases and binary maps for multiple objects, we can match the correct pairs by their semantic similarities. As a result, the objects described in the text are aligned with their corresponding objects in the mesh, which are utilized to construct a cross-modal graph $G = (\mathcal{V}, \mathcal{E})$, as shown in Fig. 3. To be specific, all surface point features and word features are considered as independent nodes to form the node set $\mathcal{V}$. For the edge set $\mathcal{V}$, the link between the surface point node and the word node will be built if the semantic objects they belong to are the same.

Under the setting of this cross-modal graph, we can individually perform cross-attention between the surface point nodes and their neighboring word nodes, where the parsed surface point features are used as queries, and the parsed text features serve as keys and values. The enhancement of the surface point node $v_i \in \mathbb{R}^{dim}$ can be formulated as:

$$\hat{v}_i = \sum_{v_j \in \mathrm{Adj}(v_i)} \alpha_{ij} \mathrm{Linear}(v_j), \quad (3)$$

$$\alpha_{ij} = \frac{e^{\mathcal{W}_{ij}}}{\sum_{v_j \in \mathrm{Adj}(v_i)} e^{\mathcal{W}_{ij}}}, \quad (4)$$

$$\mathcal{W}_{ij} = \frac{\mathrm{Linear}(v_i)\mathrm{Linear}(v_j)^{\mathrm{T}}}{\sqrt{d_l}}, \quad (5)$$

where $\mathrm{Adj}(v_i)$ is the adjacency nodes of v_i and $\mathrm{Linear}(\cdot)$ represents a linear transformation. With this attention mechanism, the surface point features of different objects in the mesh can be distinguishably reinforced under the guidance of the word features in the parsed text.

3.3. Cross-Grained Contrast Supervision

To guide the optimization of the neural network for 3D stylization, the first step is to render the stylized 3D mesh from multiple 2D views. Most existing methods usually employ the visual encoder and text encoder of CLIP [28] to extract global feature vectors for the rendered image and target text, respectively, which are contrasted to perform cross-modal supervision via cosine similarity:

$$\mathcal{L}_{coarse} = -\frac{\mathcal{F}_I \cdot \mathcal{F}_T}{\|\mathcal{F}_I\|_2 \|\mathcal{F}_T\|_2}, \tag{6}$$

where $\mathcal{F}_I \in \mathbb{R}^{512}$ is the averaged feature vector of the images rendered from different views, $\mathcal{F}_T \in \mathbb{R}^{512}$ denotes the global feature vector of the target text, and $\|\cdot\|_2$ represents the Euclidean norm function.

Although achieving impressive results for stylizing a single 3D object, these methods still have limitations in multi-object scenes. Considering that a single feature vector still represents a sentence describing multiple objects, the object details may be lost in large amounts. Therefore, such a coarse-grained contrast supervision is insufficient to guide the neural network in synthesizing photorealistic stylized content for multiple 3D objects.

To solve this issue, we construct a fine-grained contrast supervision to complement the coarse-grained one. Specifically, we first calculate the correlation map, *i.e.,* $\mathcal{S} \in \mathbb{R}^{n\times m}$ between the word features in the text and the visual features of the rendered images, which are also extracted from the text encoder and visual encoder of the CLIP:

$$\mathcal{S} = \frac{\mathcal{I} \cdot \mathcal{T}^{\mathrm{T}}}{\|\mathcal{I}\|_2 \|\mathcal{T}\|_2}, \tag{7}$$

where $\mathcal{I} \in \mathbb{R}^{n\times 512}$ represents the features of the images rendered from n views, $\mathcal{T} \in \mathbb{R}^{m\times 512}$ indicates the features of m words in the text. Then, we normalize the correlation matrix along the image axis and the text axis, respectively, to retrieve the text of interest and visual components. This process can be formulated as:

$$\mathcal{S}_I(i) = \frac{\sum_{k=1}^{m} \mathcal{S}(i,k)}{m}, \tag{8}$$

$$\mathcal{S}_T(j) = \frac{\sum_{k=1}^{n} \mathcal{S}(k,j)}{n}. \tag{9}$$

Inspired by [22], we further calculate an image-centered fine-grained contrast score and a text-centered fine-grained contrast score by the weighted summation of the similarity vectors, which can be formulated as follows:

$$\mathcal{L}_I = \sum_{i=1}^{n} \frac{e^{\mathcal{S}_I(i)}}{\sum_{k=1}^{n} e^{\mathcal{S}_I(k)}} \mathcal{S}_I(i), \tag{10}$$

$$\mathcal{L}_T = \sum_{j=1}^{m} \frac{e^{\mathcal{S}_T(j)}}{\sum_{k=1}^{m} e^{\mathcal{S}_T(k)}} \mathcal{S}_T(j), \tag{11}$$

where the weights are defined as the degree of correlation between the central and another modality. Finally, we adopt the average value of these two scores as the fine-grained contrast loss, which can be defined as:

$$\mathcal{L}_{fine} = -(\mathcal{L}_I + \mathcal{L}_T)/2. \tag{12}$$

The coarse-grained and fine-grained contrast supervision complement each other to build a cross-grained contrast supervision system. The former is utilized to align the global semantic information of the target text with the 3D objects, and the latter is used to achieve the local semantic alignment. This loss can be defined as:

$$\mathcal{L}_{cgcs} = \lambda_c \mathcal{L}_{coarse} + \lambda_f \mathcal{L}_{fine}, \tag{13}$$

where λ_c and λ_f are two hyper-parameters to balance the cross-grained and the fine-grained losses, set to 1.0 and 0.33, respectively.

4. Experiments

4.1. Experiment Setup

Datasets. To examine our method across a diverse set of 3D scenes, we first collect 3D object meshes from a variety of sources, *i.e.,* COSEG [33], Thingi10K [55], Shapenet [6], Turbo Squid [36], and ModelNet [43]. Then, we randomly place several objects from the collected 3D set into a mesh using Blender. Note that we down-sample the number of meshes' vertices and faces to ensure the robustness of our TeMO to low-quality meshes and reduce the burden of GPU during the stylization. The meshes used in this paper contain an average of 79,303 faces, 16% non-manifold edges, 0.2% non-manifold vertices, and 12% boundaries.

Implementation Details. Following the TANGO [15] network, we adopt 3 linear layers with 256 dimensions to build the normal estimation branch. In the reflectance branch, the point features are extracted by 2 shared layers with 256 dimensions, followed by 3 exclusive layers to predict diffuse, specular, and roughness. The dimension of our DGA module is also set as 256. The word features in our DGA module are extracted from the text encoder of CLIP, and so are the ones in our CGC loss. We choose ViT-B/32 as the backbone of the pre-trained CLIP model in this paper, which is consistent with previous works [15, 23, 25]. We also process the rendered images with 2D augmentation strategies [12, 15]

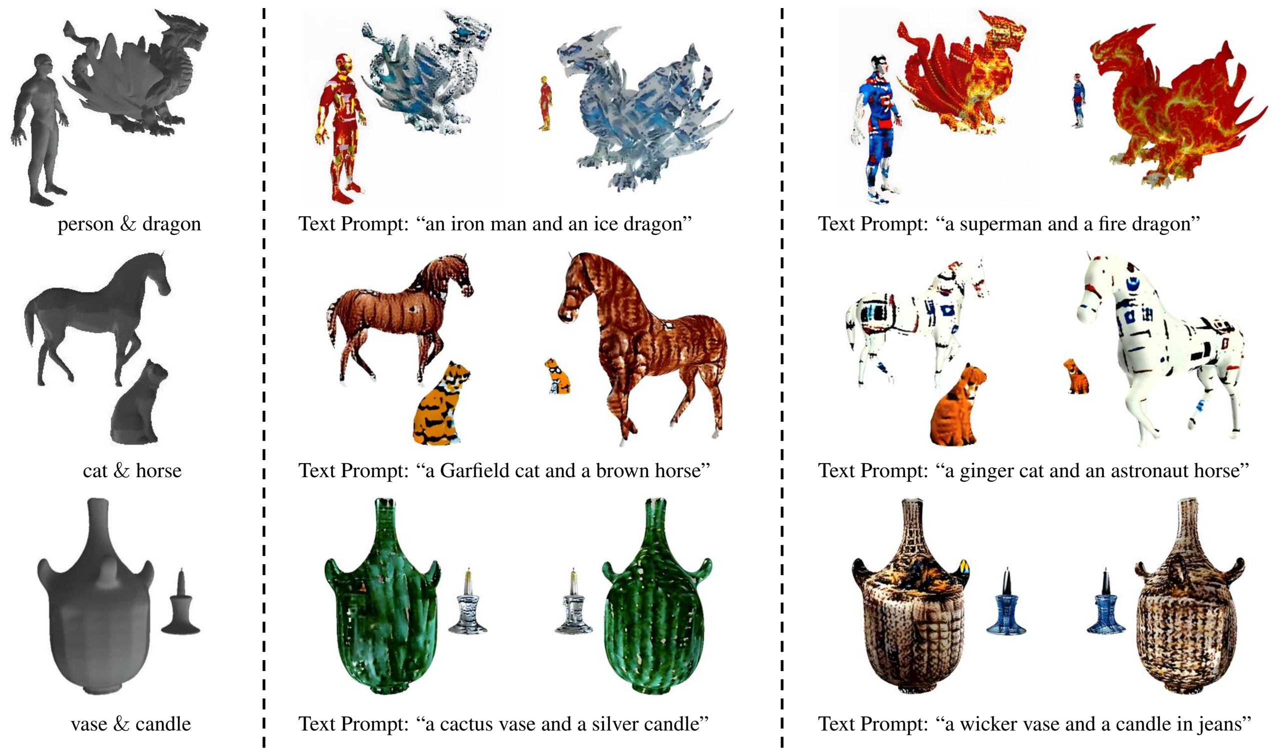


Figure 4. Given the same bare mesh, our TeMO produces various stylized contents for multi-object scenes to conform to the text prompts.

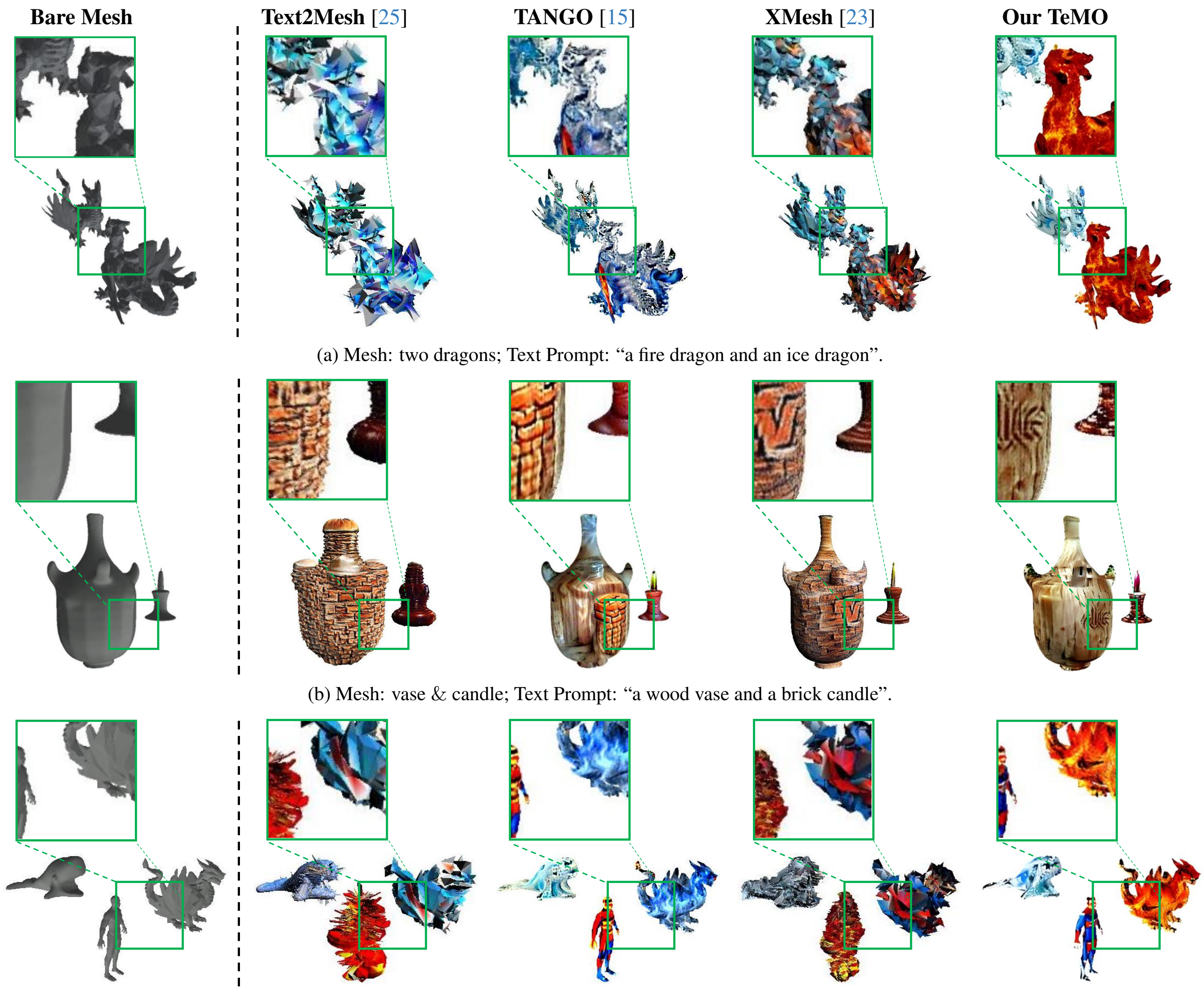
before feeding them into the pre-trained CLIP model. Our TeMO model is optimized with the AdamW [18] strategy for 1500 iterations, where the learning rate is initialized to 5×10^{-4} and decayed by 0.7 every 500 iterations. The entire training process takes about 10 minutes on a single NVIDIA RTX 3090 GPU.

4.2. Qualitative Evaluation

We conduct visualization experiments on a wide spectrum of multi-object scenes to verify the effectiveness of our TeMO. However, we observe that the 3D symmetry prior used widely in previous works [15, 25] can cause interference between different parts during the stylization process of multiple objects. We argue that the multiple objects of the meshes used in this paper are randomly placed to simulate a real 3D scene rather than along the z-axis. To avoid this issue, we remove this prior in our TeMO and previous methods involved in the comparison.

Neural Stylization and Controls. We present the stylization results of our TeMO driven by different text prompts for the same multi-object mesh in Fig. 4. As shown in the 1^{st} row where the 3D scene is composed of a person and a dragon, our TeMO can accurately distinguish between the person object and the dragon object and appropriately stylize different body parts for them according to the semantic roles described in each text prompt. Meanwhile, our TeMO also synthesizes desired stylizations for the 3D objects in the cat-horse mesh and vase-candle mesh, as shown in the 2^{nd} and 3^{rd} rows. These experimental results demonstrate that our TeMO method can generate photorealistic details with fine granularity and maintain global semantic understanding for the given multi-object 3D scene.

Qualitative Comparisons. We provide the visual comparisons of prediction results between our TeMO and previous pioneering works in text-driven 3D object stylization, including Text2Mesh [25], TANGO [15], and X-Mesh [23]. To ensure a fair comparison, we adopt the official implementations of these methods and also train them with the default settings without the symmetry prior. The experimental results show it is a real struggle for Text2Mesh [25] and TANGO [15] to understand the detailed semantics of the text prompt with multiple objects. As shown in the 1^{st} row of Fig. 5 where the 3D scene contains two objects of the same category, given a text prompt "a fire dragon and an ice dragon", they tend to capture the "ice" property, missing the "fire" property. For a 3D scene containing two objects of different categories, they are prone to mixing the properties of these objects, as shown in the 2^{nd} row where the text prompt is "a wood vase and a brick candle". Therefore, the stylized assets they generate for these multi-object scenes are unsatisfactory. X-Mesh generates more accurate results that align with the text prompts, as shown in the 1^{st} and 2^{nd} rows, which can be attributed to incorporating the text vector while extracting vertex features. However, it can pro-



(a) Mesh: two dragons; Text Prompt: "a fire dragon and an ice dragon".

(b) Mesh: vase & candle; Text Prompt: "a wood vase and a brick candle".

(c) Mesh: person & dragon & whale; Text Prompt: "a superman, a fire dragon, and an ice whale".

Figure 5. Visual comparisons of our TeMO with previous text-driven 3D stylization methods on several multi-object scenes, including two objects of the same or different categories, and three different objects.

duce semantic noises due to its utilization of the text vector containing attributes of multiple objects to process all vertex features. With an increasing number of objects, it will also encounter challenges related to comprehending text details and the alignment between the text and 3D objects. As shown in the 3^{rd} row, this method still fails to generate stylized assets without mixed properties. In contrast, our TeMO equipped with 3D scene parsing and multi-grained supervision, is able to generate photorealistic stylized content for each object in these 3D scenes to conform to the descriptions in the text prompts.

4.3. Quantitative Evaluation

Objective Metric. We adopt the CLIP score to objectively evaluate the semantic alignment achieved by our TeMO and recent 3D stylization methods. Specifically, 8 views spaced $45°$ around the stylized meshes are chosen to obtain the rendered 2D images. Then, the visual objects are compared with the textual objects in CLIP's embedding space via the cosine function. As shown in the 2^{nd} column of Tab. 1, our TeMO surpasses previous methods by a large margin. These results demonstrate the superiority of our TeMO over existing methods on multi-object stylization.

User Study. We further conduct a user study to evaluate these 3D stylization methods subjectively. We randomly select 10 mesh-text pairs and recruit 60 users to evaluate the quality of the stylization assets generated by our TeMO and previous methods. Particularly, the participants include experts in the field and individuals without specific background knowledge. Moreover, each of them will be asked three questions [25]: (Q1) "How natural is the output results?"(Q2) "How well does the output match the original

Figure 6. Ablation experiments on the proposed designs of our TeMO. Mesh: two dragons; Text Prompt: "a fire dragon and an ice dragon".

Table 1. Quantitative comparisons of our TeMO and previous text-driven 3D stylization methods in multi-object scenes, including an objective alignment score (0-1) and three subjective opinion scores (1-5). Note that the higher these metrics, the better the method.

	Alignment	User-Q1	User-Q2	User-Q3
Text2Mesh [25]	0.262	1.750	1.506	1.472
TANGO [15]	0.274	2.406	2.450	2.539
X-Mesh [23]	0.265	1.839	1.722	1.761
Our TeMO	**0.285**	**3.344**	**3.311**	**3.261**

content?" (Q3) "How well does the output match the target style?", and then assign a score (1-5) to them. We report the mean opinion scores in parentheses for each factor averaged across all style outputs. As shown in Tab. 1, our TeMO still outperforms other methods across all questions. Therefore, the 3D assets generated by our method are more in line with people's understanding of the text prompts.

4.4. Ablation Studies

To verify the effectiveness of the proposed designs in our TeMO, we conduct ablation studies by gradually adding them to our baseline model, *i.e.,* TANGO [15]. We chose the two-dragon mesh with the text prompt "a fire dragon and an ice dragon", and the experimental results are shown in Fig. 6. Compared to the baseline model, introducing our DGA module enables the model to distinguish two dragons, yet it falls short in endowing them with precise texture details. Meanwhile, incorporating our CGC loss facilitates the model to capture more semantic details, *e.g.,* "fire" and "ice". Nevertheless, it fails to distinguish the two objects. It is noteworthy that the model equipped with these two designs together is not only capable of accurately distinguishing between two objects but can also synthesize high-quality texture details for them. These experiments indicate that our DGA module and CGC loss can effectively assist the model in generating desired stylized content for multiple 3D objects to conform to the target text.

5. Limitation and Future Work

Despite achieving excellent results on text-driven multi-object stylization, our TeMO framework still has a few limitations, which can also facilitate future research:

1) 3D Symmetry Prior. As stated in Sec. 4.2, our TeMO fails to incorporate 3D symmetry prior, whose important role has been demonstrated by Text2Mesh [25] in promoting style consistency of a single object. To generate more photorealistic stylization assets for multi-object scenes, it will be valuable to calculate symmetry planes for each object and apply symmetry priors to them.

2) Diffusion Model. We observe that current diffusion technologies struggle to generate multi-object images according to the text prompt, which hinders the application of diffusion-based stylization methods in multi-object 3D scenes. We argue it would be interesting to extend the concept of scene parsing to the diffusion models for the release of their potential in multi-object editing or generation.

6. Conclusion

In this paper, we present TeMO, an innovative framework proposing scene parsing and multi-grained cross-modal supervision to achieve text-driven multi-object 3D stylization for the first time. Specifically, we first develop a DGA module to precisely align the objects in the 3D mesh and the text prompt and enhance the 3D point features with the word features belonging to the same object as them. Then, we design a CGC loss, in which the fine-grained loss at the local level and coarse-grained contrast loss at the global level are both constructed and complement each other. Further, extensive experiments are conducted to demonstrate the effectiveness and superiority of our methods over the existing methods among a wide range of multi-object 3D scenes. We believe it is promising to achieve content editing of multiple objects in 3D scenes simultaneously, and we hope the scene-parsing perspective provided by the proposed TeMO framework will inspire future works.

Acknowledgments. This research was supported by NSFC (NO. 62225604, NO. 62276145), the Fundamental Research Funds for the Central Universities (Nankai University, 070-63223049), CAST through Young Elite Scientist Sponsorship Program (No. YESS20210377). Computations were supported by the Supercomputing Center of Nankai University (NKSC).

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2014. 2

[2] Guirong Bai, Shizhu He, Kang Liu, and Jun Zhao. Example-guided stylized response generation in zero-shot setting. *Science China Information Sciences*, 2022. 2

[3] Edward Loper Bird, Steven and Ewan Klein. Natural language processing with python. *O'Reilly Media Inc*, 2009. 4

[4] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *IEEE ICCV*, 2021. 3

[5] Arthur Caetano and Misha Sra. Arfy: A pipeline for adapting 3d scenes to augmented reality. In *ACM UIST*, 2022. 1

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[7] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. In *IEEE ICCV*, 2023. 2

[8] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *IEEE ICCV*, 2023. 1, 2

[9] Shaoyu Chen, Budmonde Duinkharjav, Xin Sun, Li-Yi Wei, Stefano Petrangeli, Jose Echevarria, Claudio Silva, and Qi Sun. Instant reality: Gaze-contingent perceptual optimization for 3d virtual reality streaming. *IEEE TVCG*, 2022. 1

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[12] Kevin Frans, Lisa Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems*, 2022. 5

[13] Mengdi Han, Xiaogang Guo, Xuexian Chen, Cunman Liang, Hangbo Zhao, Qihui Zhang, Wubin Bai, Fan Zhang, Heming Wei, Changsheng Wu, et al. Submillimeter-scale multimaterial terrestrial robots. *Science Robotics*, 2022. 1

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE CVPR*, 2018. 3

[15] Jiabao Lei, Yabin Zhang, Kui Jia, et al. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[16] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. *arXiv preprint arXiv:2403.02781*, 2024. 2

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE ICCV*, 2021. 3

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[19] Haoyu Lu, Yuqi Huo, Mingyu Ding, Nanyi Fei, and Zhiwu Lu. Cross-modal contrastive learning for generalizable and efficient image-text retrieval. *Machine Intelligence Research*, 2023. 1

[20] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. In *AAAI*, 2021. 3

[21] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *ACL EMNLP*, 2015. 3

[22] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, 2022. 2, 5

[23] Yiwei Ma, Xiaioqing Zhang, Xiaoshuai Sun, Jiayi Ji, Haowei Wang, Guannan Jiang, Weilin Zhuang, and Rongrong Ji. X-mesh: Towards fast and accurate text-driven 3d stylization via dynamic textual guidance. In *IEEE ICCV*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[24] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *IEEE CVPR*, 2023. 2

[25] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *IEEE CVPR*, 2022. 1, 2, 5, 6, 7, 8

[26] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *ACM SIGGRAPH*, 2022. 2

[27] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 5

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2

[30] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH*, 2023. 2

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE CVPR*, 2022. 2

[32] Scott D Roth. Ray casting for modeling solids. *Computer graphics and image processing*, 1982. 3

[33] Oana Sidi, Oliver Van Kaick, Yanir Kleiman, Hao Zhang, and Daniel Cohen-Or. Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering. In

ACM SIGGRAPH, 2011. 5

[34] Jia-Mu Sun, Tong Wu, and Lin Gao. Recent advances in implicit representation-based 3d shape generation. *Visual Intelligence*, 2024. 1

[35] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 2020. 4

[36] TurboSquid. Turbosquid 3d model repository. In *https://www.turbosquid.com/*, 2021. 5

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[38] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *IEEE CVPR*, 2022. 1

[39] Wei Wang, Qiulei Dong, and Zhanyi Hu. Asppr: active single-image piecewise planar 3d reconstruction based on geometric priors. *Science China Information Sciences*, 2023. 2

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE CVPR*, 2018. 3

[41] Mingrui Wu, Xuying Zhang, Xiaoshuai Sun, Yiyi Zhou, Chao Chen, Jiaxin Gu, Xing Sun, and Rongrong Ji. Difnet: Boosting visual information flow for image captioning. In *IEEE CVPR*, 2022. 3

[42] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769*, 2023. 2

[43] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE CVPR*, 2015. 5

[44] Qun-Ce Xu, Tai-Jiang Mu, and Yong-Liang Yang. A survey of deep learning-based 3d shape generation. *Computational Visual Media*, 2023. 1

[45] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *IEEE ICCV*, 2021. 3

[46] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 3

[47] Bowen Yin, Xuying Zhang, Qibin Hou, Bo-Yuan Sun, Deng-Ping Fan, and Luc Van Gool. Camoformer: Masked separable attention for camouflaged object detection. *arXiv preprint arXiv:2212.06570*, 2022. 3

[48] Bowen Yin, Xuying Zhang, Zhongyu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. In *ICLR*, 2024. 3

[49] Bo Zhang, Lizbeth Goodman, and Xiaoqing Gu. Novel 3d contextual interactive games on a gamified virtual environment support cultural learning through collaboration among intercultural students. *SAGE Open*, 2022. 1

[50] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *ECCV*, 2022. 1

[51] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *IEEE CVPR*, 2021. 3

[52] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM TOG*, 2021. 3

[53] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *IEEE CVPR*, 2021. 3

[54] Xuying Zhang, Bowen Yin, Zheng Lin, Qibin Hou, Deng-Ping Fan, and Ming-Ming Cheng. Referring camouflaged object detection. *arXiv preprint arXiv:2306.07532*, 2023. 3

[55] Qingnan Zhou and Alec Jacobson. Thingi10k: A dataset of 10,000 3d-printing models. *arXiv preprint arXiv:1605.04797*, 2016. 5

[56] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *IEEE ICPR*, 2004. 4