TeMO：面向多物体网格的文本驱动 3D 风格化

张旭迎 [1]　尹博文 [1]　陈宇铭 [1]　林铮 [1]　侯淇彬 [1,2*]　程明明 [1,2]
[1] 南开大学，计算机学院　[2] 南开大学，深圳福田研究院

https://github.com/zhangxuying1004/TeMO/

Abstract

最近，在基于 *CLIP* 的方法的推动下，单个物体的文本驱动 *3D* 风格化取得了相当大的进展。然而，对于多物体 *3D* 场景的风格化仍然受到限制，因为用于预训练 *CLIP* 的图像-文本对主要由单个物体组成。与此同时，由于现有的监督方式主要依赖于图像-文本对的粗粒度对比，多个物体的局部细节可能容易被遗漏。为了克服这些挑战，我们提出了一种新颖的框架，称为 ***TeMO***，用于解析多物体 *3D* 场景，并在多个层次的对比监督下编辑其风格。首先，我们提出了一种解耦图注意力（***DGA***）模块，以明显增强 *3D* 表面点的特征。特别地，我们构建了一个跨模态图，以准确对齐从 *3D* 网格和文本描述中解耦出来的对象点和名词短语。然后，我们开发了一个交叉粒度对比（***CGC***）监督系统，其中构建了文本描述中的词语与随机渲染图像之间的细粒度损失，以补充粗粒度损失。大量实验证明，我们的方法可以合成高质量的风格化内容，并在广泛的多物体 *3D* 网格范围内优于现有方法。

1. 引言

通过风格化进行 3D 资产创建的目标是在基础网格上合成风格化的内容，以符合给定的文本描述 [15, 19, 25]、参考图像 [38, 50] 或 3D 形状 [34, 44]。这项研究在虚拟/增强现实 [5, 9]、游戏产业 [49] 和机器人技术 [13] 等广泛应用中扮演着重要角色。此外，它在计算机视觉和图形学中也展示出巨大的潜力，引起了越来越多的关注。考虑到文本提示的易用性和表达能力以及大规模对比语言-图像预训练（CLIP）[28] 模型的流行，我们选择使用文本驱动的 3D 风格化。

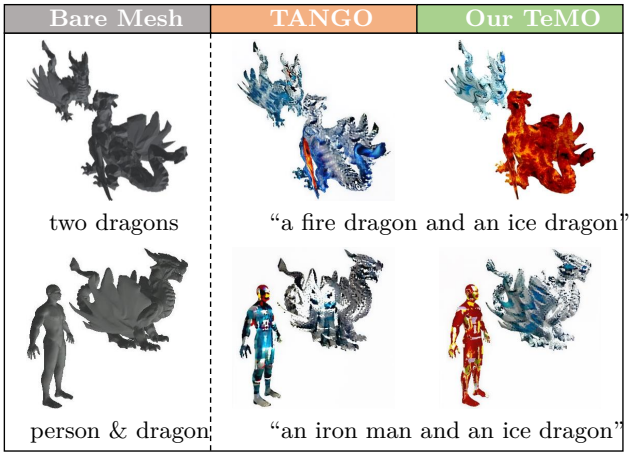


图 1. 现有的 3D 风格化方法（例如，TANGO [15]）和我们的 TeMO 在多物体场景中的视觉比较。对于具有相同/不同类别的多个物体的场景，现有方法容易在物体的不同属性之间产生干扰，而我们的 TeMO 能够准确地为每个物体合成所需的风格化内容。

近年来，一系列令人印象深刻的工作 [8, 15, 23, 25] 涌现出来，推动了文本驱动的 3D 风格化的发展。现有方法通常采用多层感知器（MLPs）来预测基础网格的位置属性偏移，并在 CLIP 的对比损失监督下进行训练。我们观察到这些工作主要关注单个 3D 物体的风格化，而在多物体场景中的表现较差，如图. 1 的第 2 列所示。我们认为 CLIP 的两个固有特性导致了这一问题：**i)** CLIP 主要通过由单个物体组成的图像-文本对进行预训练；**ii)** CLIP 损失采用来自图像和文本的全局表示向量来粗略匹配这两种模态，这不可避免地导致局部细节的丢失。此外，为多个 3D 物体合成期望风格的关键在于对这些 3D 场景的解析以及细节细化的多粒度监督。

为了同时生成多物体 3D 场景的风格化内容，首要

*侯淇彬是通讯作者

步骤是实现 3D 网格中的物体与目标文本之间的准确对齐。然而，现有方法采用文本的全局语义来风格化单个物体，这会不可避免地在多物体场景中进行风格化时产生噪声。为了解决这个问题，我们提出引入解耦图注意力（DGA）模块来解析 3D 场景。具体而言，从文本提示中解耦所有的名词短语，同时当前视图的网格表面点被划分为多个聚类。然后，构建一个跨模态图来建立名词短语与其对应的对象点之间的连接，同时将它们与不相关的对象点隔开。该图结构使得两种相关模态之间能够进行精确交互。最后，通过与图结构中相邻词节点的独立交叉注意力融合，增强 3D 物体的表面点特征。

此外，我们还设计了交叉粒度对比（CGC）损失，针对多物体风格化进行全面的跨模态监督。目标是引导网络为多个 3D 物体生成更多的风格化细节，以匹配目标文本。我们的损失由两个部分组成，即粗粒度对比和细粒度对比。在前者中，将文本提示视为句子级别的监督，利用 CLIP 模型的全局特征向量计算风格化 3D 网格渲染视图与文本提示之间的相似度。在后者中，我们从单词级别理解文本提示，并考虑句子中每个词与渲染图像集合之间的相似度。具体来说，我们通过提取 CLIP 文本编码器中的隐藏状态来生成文本提示的单词表示。受近期视频-文本检索进展的启发 [22]，我们基于每个单词或图像的重要性，对相似度向量中的元素进行加权求和，计算细粒度损失。

基于精心设计的 DGA 模块和 CGC 损失，我们提出了一种新颖的文本驱动的多物体网格 3D 风格化框架，称为 TeMO。为了验证 TeMO 的有效性，我们在各种多物体 3D 场景上进行了广泛的实验，如图. 1 的第 3 列所示。实验结果表明，与现有 3D 风格化方法相比，TeMO 更不容易受到多个物体的干扰，能够生成更高质量的风格化资产。

我们的贡献总结如下：

- 我们提出了一个新的 3D 风格化框架 TeMO。据我们所知，这是首次尝试解析文本和 3D 网格中的物体，并为多物体场景生成风格化内容。
- 我们提出了解耦图注意力 (DGA) 模块，通过构建一个图结构将多物体网格中的表面点与文本提示中的名词短语对齐。
- 我们设计了交叉粒度对比 (CGC) 损失，将文本与渲染图像在句子和单词级别进行对比。

2. 相关的工作

2.1. 文本驱动的 3D 操作

根据给定的提示生成或编辑 3D 内容是计算机视觉和图形学中的长期目标 [2, 39, 42]。在所有的提示形式中，文本因以下三大原因而备受关注：i) 文本描述在现有语料库中易于获取；ii) 文本描述对用户非常友好，因为它们易于修改，并且能够有效表达与风格化相关的复杂概念；iii) 大规模多模态模型 [16, 28] 的普及使得实现视觉-语言监督成为可能。

Text2Mesh [25] 提出了一个神经风格场网络，用于预测网格顶点的颜色和位移。TANGO [15] 通过将外观风格解耦为空间变化的双向反射分布、局部几何变化和光照条件来进行 3D 风格化。随后，X-Mesh [23] 在顶点特征提取过程中利用文本相关的空间和通道注意力来整合目标文本指导。受到文本驱动的 2D 生成显著进展的启发 [29, 31]，TEXTure [30] 和 Text2Tex [7] 结合了预训练好的深度感知图像扩散模型，以逐步从多个视角合成高分辨率的局部纹理。

为了充分利用预训练好的 2D 文本到图像扩散模型中的先验知识，DreamFusion [27] 引入了分数蒸馏采样（SDS）损失来进行文本到 3D 合成。在 SDS 损失的帮助下，Latent-NeRF [24] 和 Fantasia3D [8] 可以为 3D 物体生成 3D 形状和外观。尽管取得了令人印象深刻的结果，这些方法主要集中于单个 3D 物体的风格化，并很少探索多物体场景。CLIP-Mesh [26] 尝试为目标文本生成多个 3D 物体，但生成的内容不尽如人意。在本文中，我们通过两种精心设计的策略对渲染图像和文本提示中描述的物体进行解析和对齐。

2.2. 注意力机制

注意力机制的概念最早在神经机器翻译中被引入 [1]，其中根据候选向量的重要性得分计算加权和。这项技术已扩展到众多任务当中，例如自然语言处理 [10, 21, 37]、计算机视觉 [14, 17, 47, 54] 和多模态学习 [20, 41, 48, 53]。例如，Transformer [37] 采用自注意力操作来建立句子中单词之间的联系，并利用交叉注意力机制对齐源句子和目标句子。非局部神经网络 [40] 首次将自注意力引入计算机视觉，并在视频理解和目标检测中取得了巨大成功。ViT [11] 将图像视为一系列块，并采用基于自注意力的 Transformer 编码

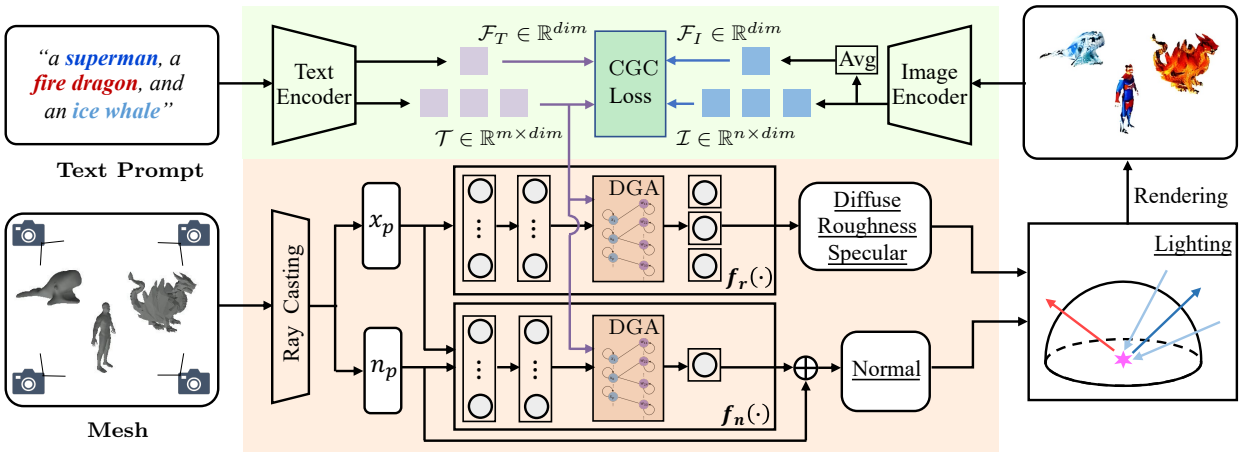


图 2. 所提出 TeMO 框架的整体架构。我们首先指定多个相机向 3D 网格场景中的物体投射光线。然后，可以从每条与物体相交的光线获得一个表面点 x_p 和法线 n_p。这些点和法线被输入到属性预测网络，解析 3D 物体的特征，并通过我们提出的 DGA 模块与解耦的文本特征进行交互。同时，我们使用一系列高斯来表示提示。最后，采用可微分的 SG 渲染器来渲染图像，并通过我们设计的 CGC 损失与文本提示进行对比。

器进行图像分类。Swin Transformer [17] 引入了移动窗口以增强自注意力的局部感知能力。最近，X-Mesh [23] 设计了一种用于 3D 物体顶点特征提取的文本引导动态注意力机制。然而，这种引导仅依赖于文本特征向量，而不考虑对文本和 3D 场景的解析。在本文中，我们通过跨模态图对从目标文本和 3D 网格中解耦出来的多个物体进行对齐，以实现精确的引导。

2.3. 多模态对比学习

对比学习由于能够对齐不同模态的表示，已成为多模态领域中越来越受欢迎的研究主题。基于这种策略，CLIP [28] 通过大量图像-文本对进行预训练，在跨模态监督中取得了巨大成功。TACo [45] 提出了一种基于词语句法类别的标记感知级联对比学习，以实现文本视频检索中的细粒度语义对齐。同时，FILIP [46] 提出了将图像块与句子中的词语进行对比的策略。对于文本驱动的 3D 风格化，CLIP 损失被广泛采用，用于计算 CLIP 嵌入空间中图像和文本向量之间的相似度。尽管这些方法在单物体风格化中取得了显著成果，但在多物体 3D 场景中表现不佳。我们认为导致这个问题的一个重要原因是这种粗粒度监督导致了局部细节的丢失。在本文中，我们提出了一种交叉粒度监督策略，考虑细粒度和粗粒度对比，以实现渲染图像与文本之间更精确的语义对齐。

3. 提出的方法

3.1. 整体架构

图. 2 显示了我们 TeMO 框架的端到端架构。给定一个基础网格和包含多个物体的文本提示，TeMO 的目标是在网格上合成风格化效果，使其与文本描述相匹配。我们使用一组顶点 $V \in \mathbb{R}^{e \times 3}$ 和面 $F \in \{1, ..., e\}^{u \times 1}$ 来显式定义输入的三角形网格，并在整个训练过程中保持固定。根据 TANGO [15]，我们将外观风格解耦为空间变化的双向反射分布函数 [4, 51, 52]（包括漫反射、粗糙度、镜面反射）、局部几何变化（法线图）和光照条件。

我们首先将顶点坐标归一化到一个单位球内。然后，使用高斯分布在网格周围随机采样点作为相机位置进行图像渲染。接下来，我们可以从采样的相机位置 c 和渲染图像中的像素 p 获取相机光线 $R_p = \{c + t\nu_p\}$，其中 ν_p 是光线的方向。进一步地，使用光线投射 [32] 找出光线与网格的第一个交点及其相交面。此外，相交面法线 $n_p \in \mathbb{R}^3$ 被用作点 $x_p \in \mathbb{R}^3$ 处的表面法线。

为了实现多视图一致特征，我们的 TeMO 限制法线位移仅作为位置的函数进行预测，同时允许颜色材料作为位置和视角的函数进行预测。因此，我们用 MLP 表示 TeMO，包括了两个分支，即：法线分支 $f_n(\cdot)$ 和反射分支 $f_r(\cdot)$。具体来说，前者用于预测点 x_p 处的法线偏移，后者被设计用于预测位置 x_p 处材料的表面反射系数，即漫反射、粗糙度和镜面反射。为了合成高频

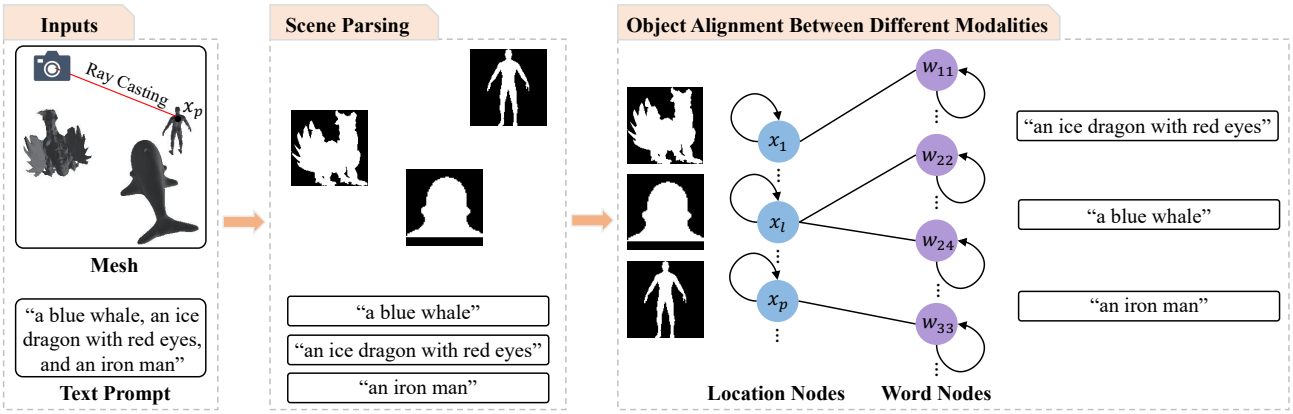


图 3. 我们 DGA 模块中跨模态图架构的构建流程。注意，x_p 是 3D 物体的表面点，w_{ij} 是第 i 个名词短语中的第 j 个词，仅当它们对应于同一个物体时才相互连接。

细节，我们还对每个输入应用了傅里叶位置编码 [35]。此外，由于球面高斯函数具有闭式形式和解析解，我们用它来表示每个光强度 $L_i(\cdot)$ 。基于所获得的几何和外观分量，可以通过半球渲染器 [15] 计算渲染图像中每个像素的颜色：

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

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

其中 $\Omega = \{w_i : w_i \cdot \hat{n}_p \geq 0\}$ 表示半球，w_i 是入射光方向，$\hat{n}_p$ 是表面点 x_p 上的估计法线。

3.2. 解耦图注意力

为了实现多个 3D 物体的文本驱动风格化，关键问题是解决文本中描述的物体与网格中的物体之间的准确对齐。X-Mesh [23] 引入了文本引导的动态线性层，其中利用文本中目标物体的全局表示向量作为引导，获取文本感知的顶点特征。然而，这种包含多个物体信息的全局向量容易产生相互干扰，并在多物体场景的引导过程中产生语义噪声。

为了解决这个问题，我们提出了对文本和网格中的物体进行解析的方法。首先，我们使用 NLTK 工具 [3] 从文本中提取由形容词或介词短语修饰的名词短语。接着，我们采用高斯混合模型（GMM）[56] 对当前光线与网格的交点集 $\{x_1, ..., x_p, ...\}$ 进行聚类。同时，我们可以根据光线是否与网格相交来获取当前视图中物体的二值图。进一步地，我们根据聚类的点将二值图中的物体进行解耦，并获取多个单独物体的二值图。对于解耦的名词短语和多物体的二值图，我们可以根据它们的语义相似性匹配得到正确的对。最后的结果是，文本中描述的物体与网格中对应的物体对齐，并用于构建跨模态图 $G = (\mathcal{V}, \mathcal{E})$ ，如图. 3 所示。具体而言，所有表面点特征和词语特征被视为独立的节点，形成节点集 $\mathcal{V}$ 。对于边集 $\mathcal{V}$ ，如果表面点节点和词语节点属于相同的语义对象，则在它们之间建立连接。

在这个跨模态图的设置下，我们可以在表面点节点与其相邻的词语节点之间单独执行交叉注意力，其中解析后的表面点特征用作 Query，解析后的文本特征作为 Key 和 Value。表面点节点 $v_i \in \mathbb{R}^{dim}$ 的增强可以用公式表示为：

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

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

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

其中，$\mathrm{Adj}(v_i)$ 是 v_i 的相邻节点，$\mathrm{Linear}(\cdot)$ 表示线性变换。通过这种注意力机制，网格中不同物体的表面点特征在解析文本中词语特征的引导下可以得到显著增强。

3.3. 交叉粒度对比监督

为了指导神经网络进行 3D 风格化的优化，第一步是从多个 2D 视角渲染风格化的 3D 网格。大多数现有方法通常使用 CLIP [28] 的视觉编码器和文本编码器分别提取渲染图像和目标文本的全局特征向量，这些向

量通过余弦相似度进行对比，实现跨模态监督：

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

其中 $\mathcal{F}_I \in \mathbb{R}^{512}$ 是从不同视角渲染的图像的平均特征向量，$\mathcal{F}_T \in \mathbb{R}^{512}$ 表示目标文本的全局特征向量，$\|\cdot\|_2$ 表示欧几里得范数。

尽管这些方法在单个 3D 物体的风格化中取得了显著成果，但在多物体场景中仍存在局限性。由于描述多个物体的句子仍然被单个特征向量表示，因此物体的细节可能会大量丢失。因此，这种粗粒度的对比监督不足以指导神经网络合成多个 3D 物体的真实感风格化内容。

为了解决这个问题，我们构建了一个细粒度对比监督来补充粗粒度对比。具体来说，我们首先计算文本中的词语特征与渲染图像的视觉特征之间的相关性图，即 $\mathcal{S} \in \mathbb{R}^{n \times m}$ ，这些特征同样从 CLIP 的文本编码器和视觉编码器中提取：

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

其中 $\mathcal{I} \in \mathbb{R}^{n \times 512}$ 表示从 n 个视角渲染的图像特征，$\mathcal{T} \in \mathbb{R}^{m \times 512}$ 表示文本中 m 个词语的特征。然后，我们分别沿图像轴和文本轴对相关矩阵进行归一化，以提取感兴趣的文本和视觉分量。这个过程可以表述为：

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

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

受 [22] 的启发，我们进一步通过相似度向量的加权求和计算以图像为中心的细粒度对比分数和以文本为中心的细粒度对比分数，其公式如下：

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

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

其中权重被定义为中心模态和另一模态之间的相关度。最后，我们采用这两个分数的平均值作为细粒度对比损失：

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

粗粒度和细粒度对比监督相互补充，构建了一个交叉粒度对比监督系统。前者用于对齐目标文本的全局语义信息与 3D 物体，后者用于实现局部语义对齐。该损失定义为：

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

其中 λ_c 和 λ_f 是平衡粗粒度和细粒度损失的两个超参数，分别设置为 1.0 和 0.33。

4. 实验

4.1. 实验设置

数据集 为了在各种 3D 场景中检验我们的方法，我们首先从多个来源收集了 3D 物体网格，包括 COSEG [33]、Thingi10K [55]、Shapenet [6]、Turbo Squid [36] 和 ModelNet [43]。然后，我们使用 Blender 将从收集的 3D 集合中选择的几个物体随机放置到一个网格中。需要注意的是，我们对网格的顶点和面数进行了下采样，以确保 TeMO 对低质量网格的鲁棒性，并在风格化过程中减少对 GPU 的负担。本文使用的网格平均包含 79,303 个面，16%为非流形边，0.2%为非流形顶点，12%为边界。

实现细节 根据 TANGO [15] 网络，我们采用了 3 层维度为 256 的线性层来构建法线估计分支。在反射分支中，点特征首先通过 2 层维度为 256 的共享层提取，随后通过 3 个专用层来预测漫反射、镜面反射和粗糙度。我们的 DGA 模块的维度也设置为 256。DGA 模块中的词语特征和 CGC 损失中的词语特征都是从 CLIP 的文本编码器中提取的。我们选择 ViT-B/32 作为预训练的 CLIP 模型的主干网络，这与之前的工作 [15, 23, 25] 保持一致。在将渲染图像输入到预训练的 CLIP 模型之前，我们使用了 2D 增强策略 [12, 15] 对其进行处理。我们的 TeMO 模型使用 AdamW [18] 优化策略进行优化，一共进行了 1500 次迭代，其中学习率初始化为 5×10^{-4}，并且每 500 次迭代的衰减为 0.7。整个训练过程在单个 NVIDIA RTX 3090 GPU 上大约需要 10

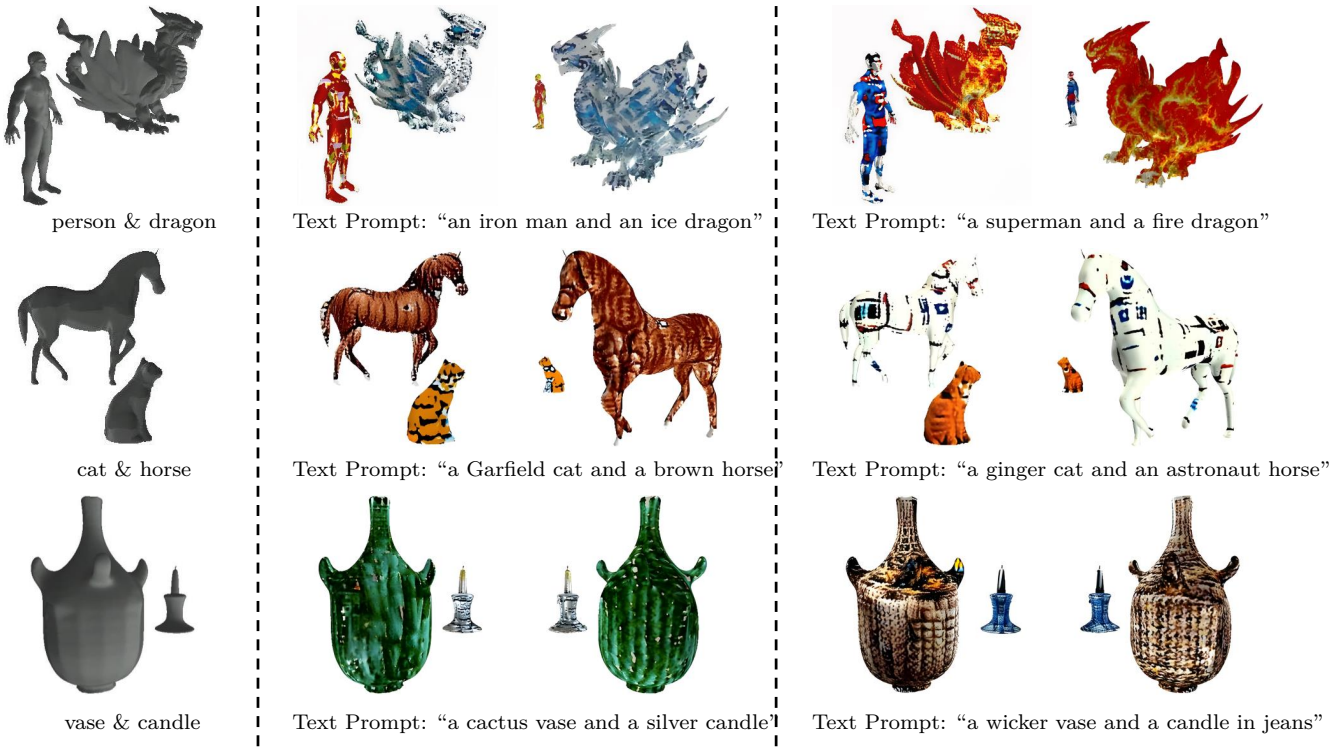


图 4. 给定相同的基础网格，我们的 TeMO 能够为多物体场景生成各种风格化内容，以符合文本提示。

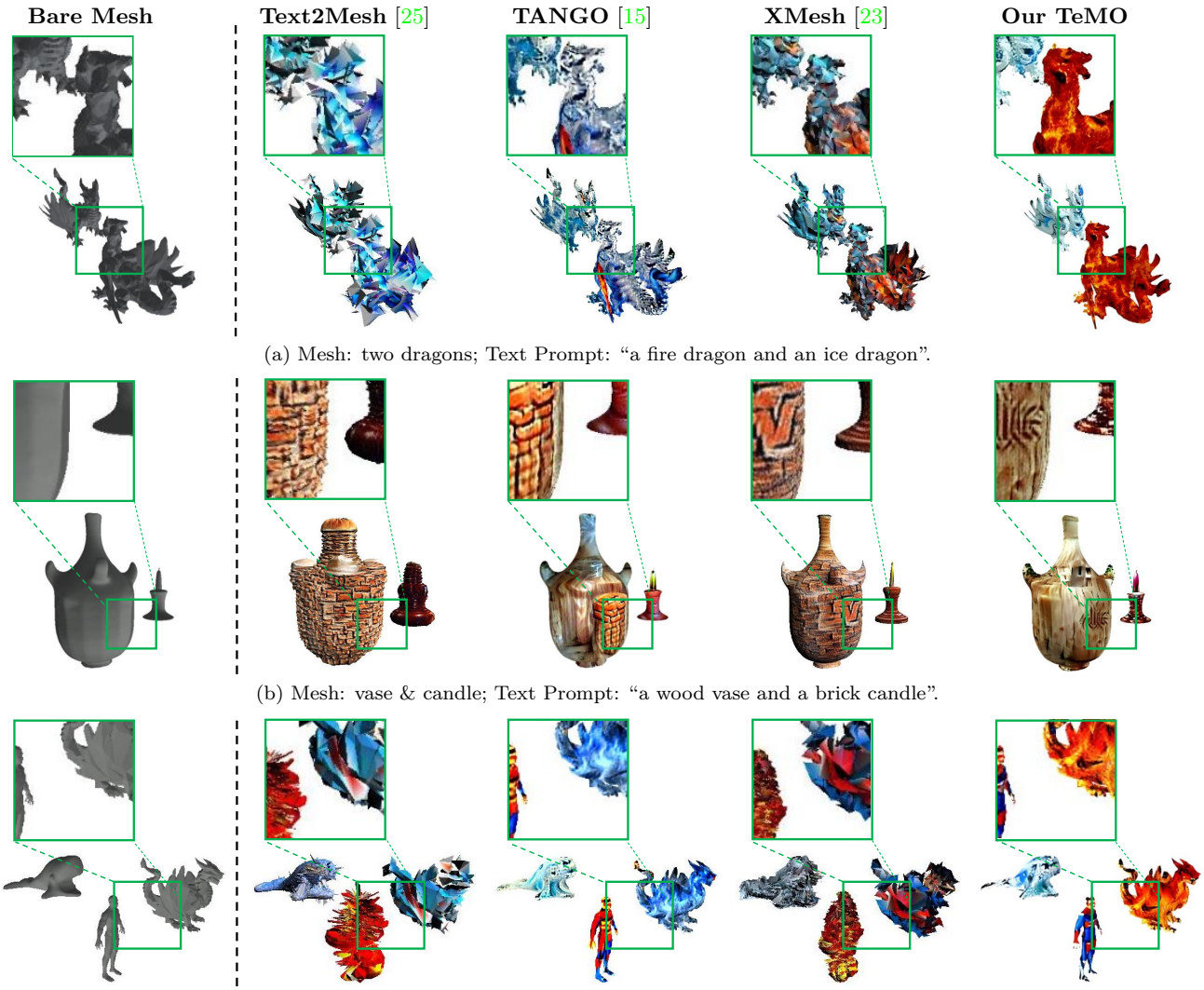
分钟。

4.2. 定性评估

我们在广泛的多物体场景中进行了可视化实验，以验证 TeMO 的有效性。然而，我们观察到，以往工作中广泛使用的 3D 对称性先验 [15, 25] 在多物体风格化过程中会导致不同部分之间的干扰。我们认为，本文使用的网格中的多个物体是随机放置的，旨在模拟真实的 3D 场景，而不是沿着 z 轴排列。为了避免这个问题，我们在 TeMO 以及参与对比的其他方法中去除了这个先验。

神经风格化与控制 如图. 4 所示，我们展示了由不同文本提示驱动的 TeMO 对同一多物体网格的风格化结果。正如第 1 行所示，3D 场景由一个人和一条龙组成，TeMO 能够准确区分人物对象和龙对象，并根据每个文本提示中描述的语义角色，适当地对它们的不同身体部分进行风格化。同时，如第 2 行和第 3 行所示，TeMO 还为猫-马网格和花瓶-蜡烛网格生成了理想的风格化效果。这些实验结果表明，TeMO 方法能够生成具有精细粒度的真实感细节，并且能够在给定的多物体 3D 场景中保持全局语义理解。

定性对比 我们提供了 TeMO 与以前在文本驱动 3D 物体风格化领域的开创性工作的视觉对比结果，这些工作包括 Text2Mesh [25]、TANGO [15] 和 X-Mesh [23]。为确保公平对比，我们采用了这些方法的官方实现，并在去掉对称性先验的情况下使用默认设置进行训练。实验结果表明，对 Text2Mesh [25] 和 TANGO [15] 来说，理解多物体场景中的详细语义是一件非常困难的事情。如图. 5 的第 1 行所示，3D 场景包含两个同类物体，给定文本提示为 “a fire dragon and an ice dragon” 时，它们倾向于捕捉 “ice” 的特性，遗漏了 “fire” 的特性。而在包含两种不同类别物体的 3D 场景中，它们容易混淆这些物体的属性，如第 2 行所示，文本提示为 “a wood vaster and a brick candle”，它们生成的风格化资产对于这些多物体场景来说并不令人满意。如第 1 行和第 2 行所示，X-Mesh 生成的结果与文本提示更为一致，这得益于在提取顶点特征时结合了文本向量。然而，由于该方法利用包含多个物体属性的文本向量来处理所有顶点特征，可能会产生语义噪声。随着物体数量的增加，它也会在理解文本细节和文本与 3D 物体对齐方面



(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”.

图 5. 我们的 TeMO 与以前的文本驱动 3D 风格化方法在多个多物体场景中的视觉比较，包括相同或不同类别的两个物体，以及三个不同的物体。

遇到挑战。正如第 3 行所示，该方法仍然未能生成没有混合属性的风格化资产。相比之下，我们的 TeMO 通过 3D 场景解析和交叉粒度监督，能够为这些 3D 场景中的每个物体生成真实感的风格化内容，以符合文本提示中的描述。

4.3. 定量评估

客观指标 我们采用 CLIP 评分来客观评估 TeMO 和最新的 3D 风格化方法在语义对齐方面的表现。具体来说，我们选择了围绕风格化网格每隔 45° 采样的 8 个视角，生成渲染的 2D 图像。然后，使用余弦函数在 CLIP 的嵌入空间中比较视觉对象和文本对象的相似性。正如表. 1 的第 2 列所示，TeMO 相比于之前的方法取得了很大的优势。这些结果表明，TeMO 在多物体风格化方面优于现有方法。

表 1. 我们的 TeMO 与以前的文本驱动 3D 风格化方法在多物体场景中的定量比较，包括一个客观对齐评分（0-1）和三个主观评分（1-5）。请注意，分数越高，方法效果越好。

	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**

图 6. 对我们的 TeMO 提出的设计进行消融实验。网格：two dragons；文本提示：“a fire dragon and an ice dragon”。

用户研究 我们进一步进行了用户研究，以主观评估这些 3D 风格化方法。我们随机选择了 10 对网格-文本对，并邀请了 60 位用户来评价 TeMO 和之前方法生成的风格化资产的质量。特别地，参与者包括该领域的专家和没有特定背景知识的普通用户。此外，我们要求每个参与者回答三个问题 [25]：（Q1）“输出结果的自然程度如何?”（Q2）“输出结果与原始内容的匹配程度如何?”（Q3）“输出结果与目标风格的匹配程度如何?”，然后给出评分（1-5）。如表. 1 所示，我们展示了所有风格化方法输出的平均得分的均值意见分数（MOS）。在所有问题上，TeMO 依然优于其他方法。因此，我们的方法生成的 3D 资产更符合人们对文本提示的理解。

4.4. 消融实验

为了验证 TeMO 中所提出设计的有效性，我们通过逐步将它们添加到我们的基准模型（即 TANGO [15]）中，进行消融实验。我们选择了 “two dragons” 的网格，并使用文本提示 “a fire dragon and an ice dragon”，实验结果如图. 6 所示。与基准模型相比，引入我们的 DGA 模块使模型能够区分两条龙，但在赋予它们精确的纹理细节方面仍然不足。同时，加入 CGC 损失后，模型能够捕捉到更多的语义细节，例如 “fire” 和 “ice”，但仍未能完全区分两个物体。值得注意的是，结合这两个设计的模型不仅能够准确区分两个物体，还能为它们合成高质量的纹理细节。这些实验表明，DGA 模块和 CGC 损失能够有效地帮助模型为多个 3D 物体生成符合目标文本的理想风格化内容。

5. 局限与未来工作

尽管我们的 TeMO 框架在文本驱动的多物体风格化方面取得了优异的结果，但仍然存在一些局限性，这也可以促进未来的研究：

1) 3D 对称先验。如第 4.2 节所述，我们的 TeMO 未能结合 3D 对称先验，而 Text2Mesh [25] 已证明其在促进单个物体的风格一致性方面的重要作用。为了生成更具真实感的多物体场景风格化资产，计算每个物体的对称平面并将对称先验应用于这些物体将非常有价值。

2) 扩散模型。我们观察到，当前的扩散技术在根据文本提示生成多物体图像方面存在困难，这阻碍了基于扩散的风格化方法在多物体 3D 场景中的应用。我们认为，将场景解析的概念扩展到扩散模型中，以释放其在多物体编辑或生成方面的潜力，将是一个有趣的方向。

6. 结论

在本文中，我们提出了一个创新框架，TeMO，首次通过场景解析和交叉粒度跨模态监督实现文本驱动的多物体 3D 风格化。具体来说，我们首先开发了一个 DGA 模块，以精确对齐 3D 网格中的物体和文本提示，并增强与它们属于同一物体的词语特征相对应的 3D 点特征。然后，我们设计了 CGC 损失，其中局部的细粒度损失和全局的粗粒度对比损失相互补充。进一步地，我们进行了广泛的实验，以证明我们的方法在多种多物体 3D 场景中的有效性和优越性。我们相信，同时实现 3D 场景中多个物体的内容编辑是有前景的，并希望所提出的 TeMO 框架提供的场景解析视角能够启发未来的研究工作。

致谢: 本研究得到了国家自然科学基金（编号：62225604、62276145）、中央高校基础研究基金（南开大学，070-63223049）、中国科学院青年科学家资助计划（编号：YESS 20210377）的资助。计算得到了南开大学超级计算中心（NKSC）的支持。

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. 2

[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. 2

[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. 2

[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. 2, 3

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[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. 2

[21] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *ACL EMNLP*, 2015. 2

[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

[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, 4

[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. 2

[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. 2

[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. 2

[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. 2

[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. 2

[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. 2

[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. 2

[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