题目

摘要

文中提出了 Text2Tex，一种从给定文本提示生成高质量 3D 网格纹理的新方法。方法结合了一个预训练的深度感知图像扩散模型，从多个视点逐步合成高分辨率的部分纹理。为了避免在视图中积累的不一致和拉伸的工件，我们动态地将渲染视图分割成一个生成掩码，它表示每个可见文本的生成状态。这种划分的视图表示指导深度感知的绘图模型生成和更新相应区域的部分纹理。此外，我们提出了一种自动视图序列生成方案，以确定更新部分纹理的下一个最佳视图。大量的实验表明，我们的方法明显优于现有的文本驱动方法和基于 gan 的方法。

关键词： Text2tex；深度感知图像扩散模型；高质量 3D 网格纹理；自动视图序列生成

1 引言

最近，利用扩散模型架构，文本到图像生成器在 2D 领域取得了显著进展，实现了基于文本描述的高分辨率 2D 内容生成。然而，通过这种 2D 视觉语言先验知识来产生 3D 纹理存在着显著的挑战。具体来说，合成的纹理不仅要对语言线索具有高保真度，而且要对目标网格具有高质量和一致的质量。因此，之前从文本输入绘制 3D 几何图形的尝试通常无法提供纹理良好的 3D 内容。

在本文中，介绍了一种新的纹理合成方法 Text2Tex，该方法使用预训练的深度感知文本到图像扩散模型无缝地对 3D 物体进行纹理合成，能够根据已知的文字以及模型生成高质量的 3D 纹理。

其技术贡献有如下三点：1. 设计了一种高质量纹理合成方法，通过 depth-aware 扩散模型逐步地绘制以及更新 3D 纹理。2. 提出了一种动态确定纹理空间生成和更新顺序的自动视图序列生成方案。3. 对大量的 3D 对象进行了广泛的研究，证明了所提出的方法对于大规模的 3D 内容生成是有效的。

2 相关工作

2.1 从 3D 和 2D 数据对 3D 内容的生成

与 2D 图像相比，有几种 3D 的表示方式，所以也就有几种基于不同表现方式的生成模型，比如基于体素 [20] [28] [31] [33]，基于点云 [2] [21]，基于网格 [35]，或者是基于有符号距离函数 [9] [10] [11] [12] [14]。然而 3D 的数据是稀缺的，因此，从 3D 生成模型合成的样本，

在 3D 数据上训练，在结构和纹理方面的质量和多样性有限。最近的工作利用可微分渲染来学习并仅使用 2D 图像生成纹理 [29] [34]。然而，他们通常接受特定形状类别的训练，并在纹理质量方面挣扎。

2.2 依据文本的内容生成

近年来，视觉语言 [18] [30] [32] [25] [6] [5] [7] [8] 领域取得了巨大的进步。具体来说，对比语言图像预训练 (contrast Language-Image Pre-Training, CLIP) [25] 的出现，通过对文本对进行语义丰富的训练，促进了文本引导图像生成的发展。然而扩散模型 [13] [16] [17] 因为其视觉质量以及训练的稳定性，获得越来越多的关注。在这项工作中，利用了 Stable Diffusion [27] 的深度调节特性来提供更一致的纹理。

2.3 从 2D 数据的文字到 3D 内容生成

因为 NeRF [23] 的成功，提出了基于 NeRF 的生成器 [1] [3] [4] [15]，使用基于 gan 的框架从 2D 图像中学习 3D 结构。一个新的研究方向是将 NeRF 技术与基于扩散的文本-图像模型相结合，使文本到 3d 的学习仅在 2D 监督下实现。为了解决优化 NeRF 场的挑战，提出了分数蒸馏损失 [24]，该方法利用预训练的 2D 扩散模型作为批评家来提供基本梯度。随后的研究重点是在潜在空间 [10] [22] 中采用这种损失，并采用从粗到细的细化方法 [19]。然而，基于优化的方法被较长的收敛时间所困扰。最近的一项并行研究 [26] 提出了一种从多个预设视点逐步更新的非优化方法。

相比之下，文中的方法从自动选择的视点序列迭代更新和细化合成纹理，最大限度地减少了人为设计不同几何形状的不同视点顺序的工作量。

3 本文方法

3.1 本文方法概述

这是它大致的 pipeline，如图 1所示:

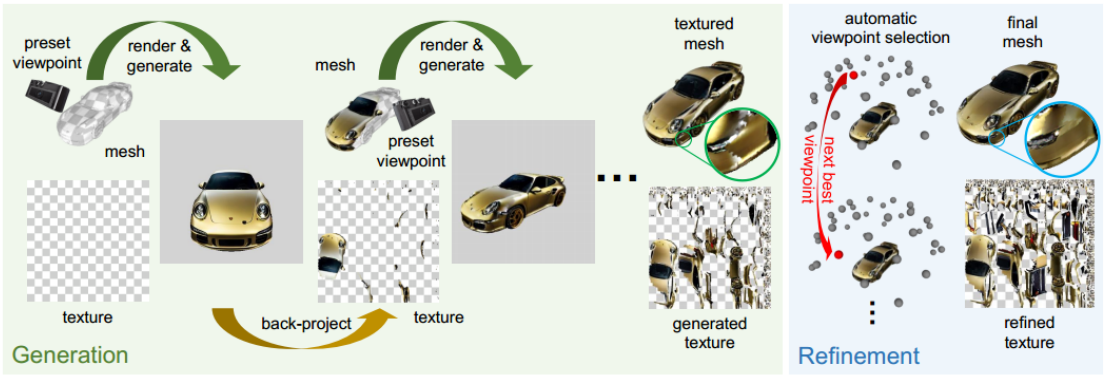

图 1. pipeline

给定一个汽车的 mesh，首先从初始的视点进行渲染，通过 depth-to-image diffusion model 根据输入提示生成新的外观生成，并将生成的图像投影回局部纹理，然后重复这个过程，直

到最后一个预设视点输出初始的纹理网格；再通过纹理细化，从一系列自动选择的视点更新初始纹理，以处理拉伸以及模糊的部位。

3.2 diffusion model

在实现中，使用了 latent diffusion model 作为生成模型，在扩散过程之前，首先训练好一个自编码模型，这样，我们就可以利用这个编码器对图片进行压缩，将输入的图像编码为 latent code，在潜在表示空间中做 diffusion 操作，最后再用解码器恢复到原始像素空间。相比于普通的 diffusion，能大大减少计算复杂度，同时有很好的生成效果。

同时，为了防止在扩散过程中的完全随机性，引入了一个值在 0-1 之间的比例因子去控制扩散的步数，通过这个因子，我们就可以中间某一步的时间开始对隐码进行去噪，从而用原始图像的部分信息引导扩散的过程。

3.3 Depth aware Image inpainting

纹理生成的核心是在网格表面去绘制缺失的区域，生成的纹理应该要高度与网格的几何以及文字吻合，为了实现这一目标，使用了 depth-to-image model，可以根据输入的文字以及深度图生成高质量的图片，如图 2所示：

图 2. Depth2img

3.4 Texture generation

在 depth2img model 的基础上，我们只能通过文字得到一张图片，而我们想要的是 3D 纹理图，所以我们还需要额外的信息去指导整个生成的过程。

文中引入了 similarity mask 以及 generation mask，如图 3所示。similarity mask 用来表示从当前视角看去的 mesh 的可见度，根据这个 mask 就可以判断某区域相对于视角是正还是斜。在此基础上，去生成 generation mask，指示哪些区域需要更新，哪些区域需要保留，如果没有纹理就 new；有纹理但是本次的视角更正，就覆盖旧纹理 update；有纹理且本次的视角更斜，就保持 keep。然后对每个视点都进行这样的视点，就可以得到一个完整的纹理。

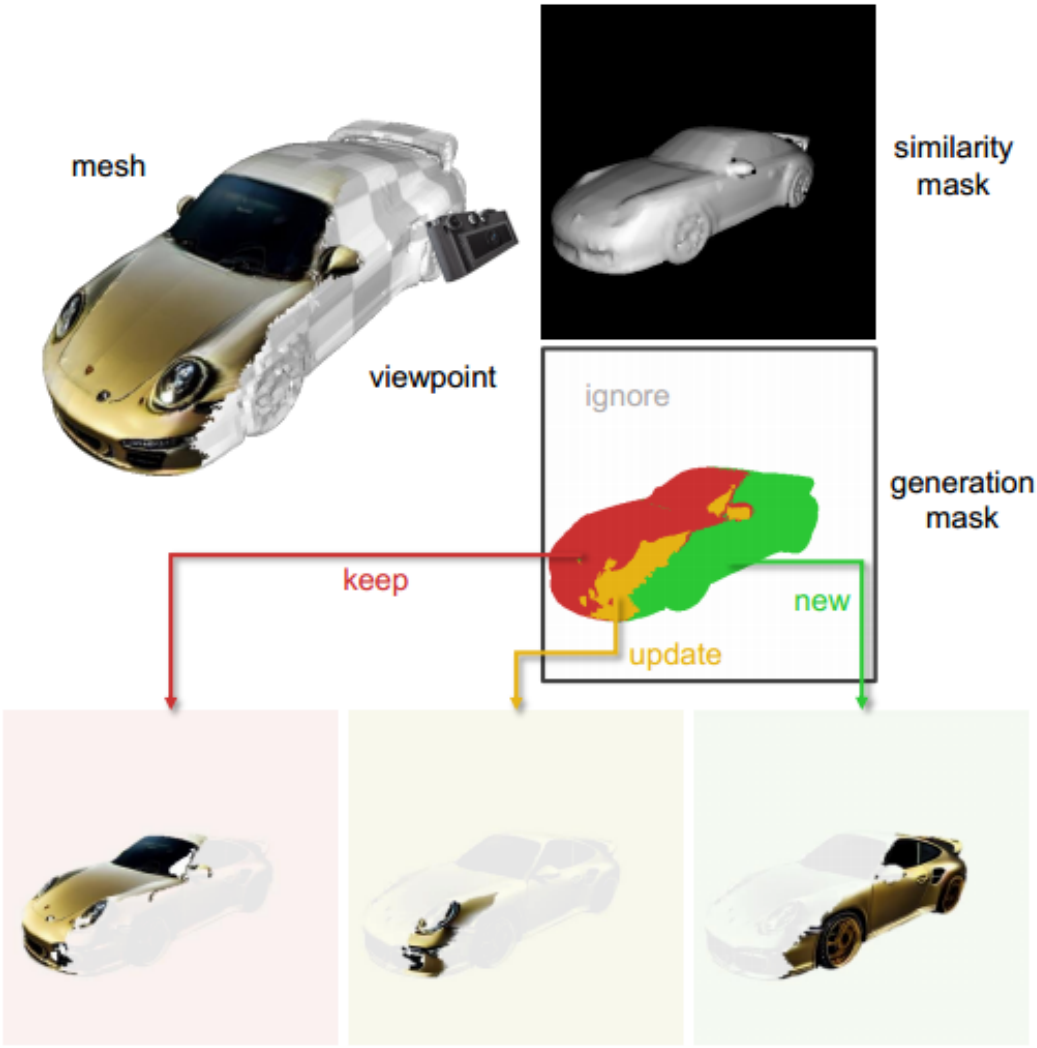

图 3. generation mask

3.5 Automatic viewpoint selection

然而，因为这些视点是预先选定的，所以对于不同的模型可以效果并不一样，纹理之间会存在不一致的情况。因此在 generation 后，又加入了一个 Refinement 的步骤，即通过增加更多的视点，然后对这些视点进行选择，再对不一致的纹理进行更新。

首先，会定义一组以物体为圆心的半球上均匀分布的密集视点，对于所有视点，得到对应的 genration mask，并计算一个 view heat 的值，这个值就是表示 update 区域的面积与当前可视区域面积的占比，归一化而得到的值。然后每次选择 view heat 值最大的视角进行更新，为了防止与之前生成的纹理有过大的差异，这里会使用一个比较温和的 denoising strength，去保留原有纹理的外观。过程如图 4所示：

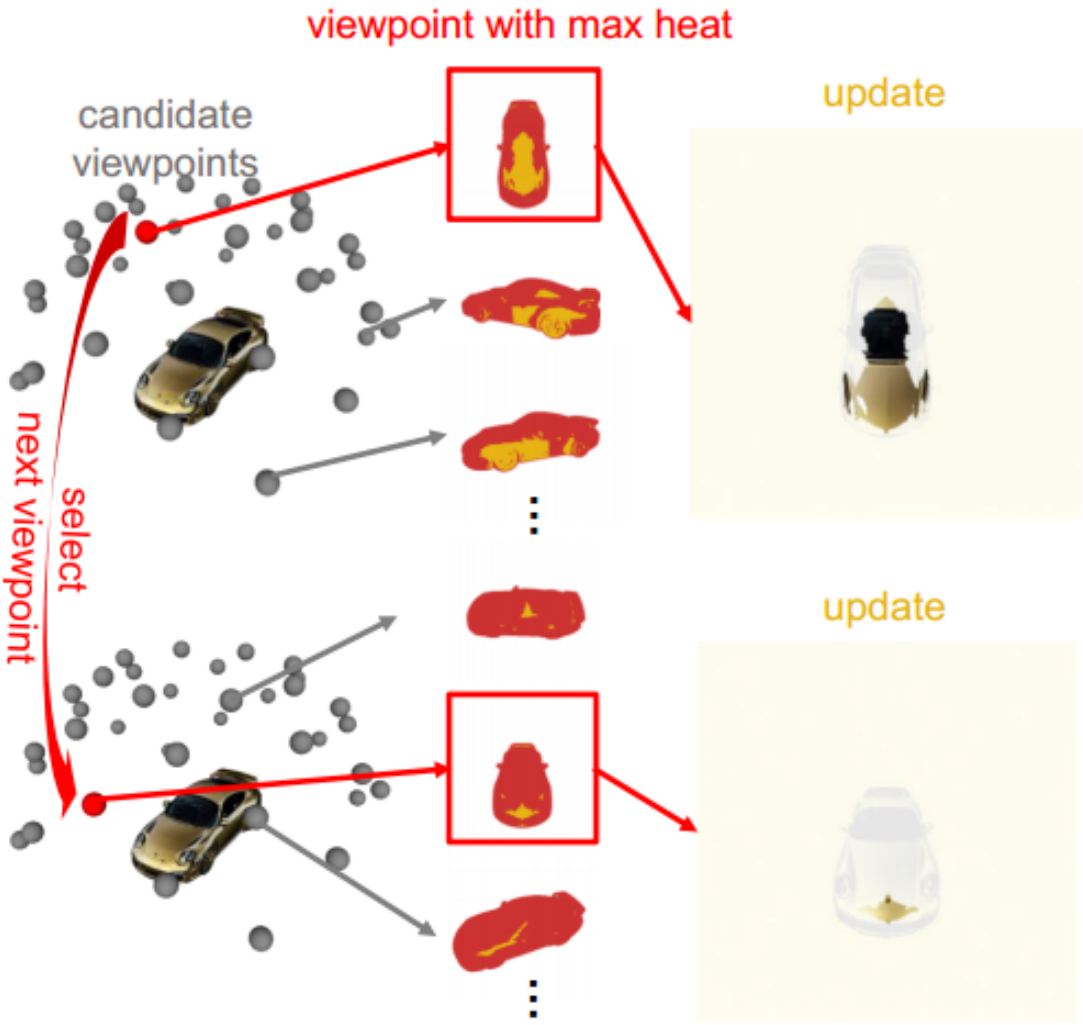

图 4. viewpoint selection

4 复现细节

4.1 实验环境搭建

- 环境：Ubuntu+Pytorch1.21.1+CUDA 11.3
- 显卡：Quadro RTX 6000 24GB

配置过程如图 5所示：

```
# create and activate the conda environment
conda create -n text2tex python=3.9.15
conda activate text2tex

# install PyTorch 1.12.1
conda install pytorch==1.12.1 torchvision==0.13.1 torchaudio==0.12.1 cudatoolkit=11.3 -c pytorch
# install runtime dependencies for PyTorch3D
conda install -c fvcore -c iopath -c conda-forge fvcore iopath
conda install -c bottler nvidiacub

# install PyTorch3D
conda install pytorch3d -c pytorch3d
# please don't use pip to install it, as it only supports PyTorch>=2.0
conda install xformers -c xformers
pip install -r requirements.txt
```

图 5. 配置

4.2 使用说明

1. 将输入的 obj 文件放入 data 文件夹中（附：如果使用自己的 Obj 文件，需要先对其进行预处理，需要满足一下条件：1. Y 轴向上 2. 输入网格需要面向 Z 的正轴方向 3. 其包围盒与原点对齐 4. 包围盒最大边长应该在 1 左右）代码中提供了预处理的程序，同样，还提供了参数化的程序去生成 UV map。

2. 在 run.sh 中修改输入路径、输出路径、文件名以及输入的提示词，如图 6所示：

```
python scripts/generate_texture.py \
    --input_dir data/Hitec \
    --output_dir outputs/Hitec \
    --obj_name mesh \
    --obj_file mesh.obj \
    --prompt "an office tower with aligned windows" \
```

图 6. 输入

5 实验结果分析

运行实验结果如图 7所示：

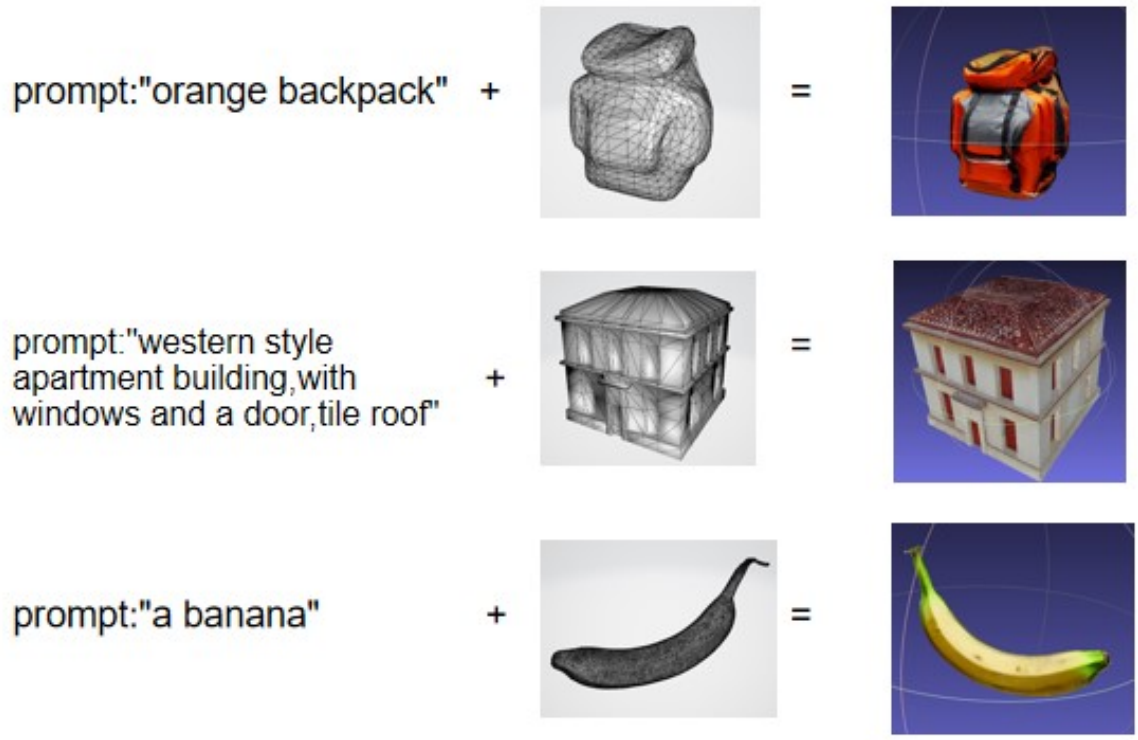

图 7. 运行结果

5.1 改进思路

在实验过程中发现，因为视点选择的问题，该方法对于一些自遮挡较多的模型效果并不是特别好，如图 8所示。所以对于这类型的模型可以通过进一步改进视点的选择机制来获得更好的效果。

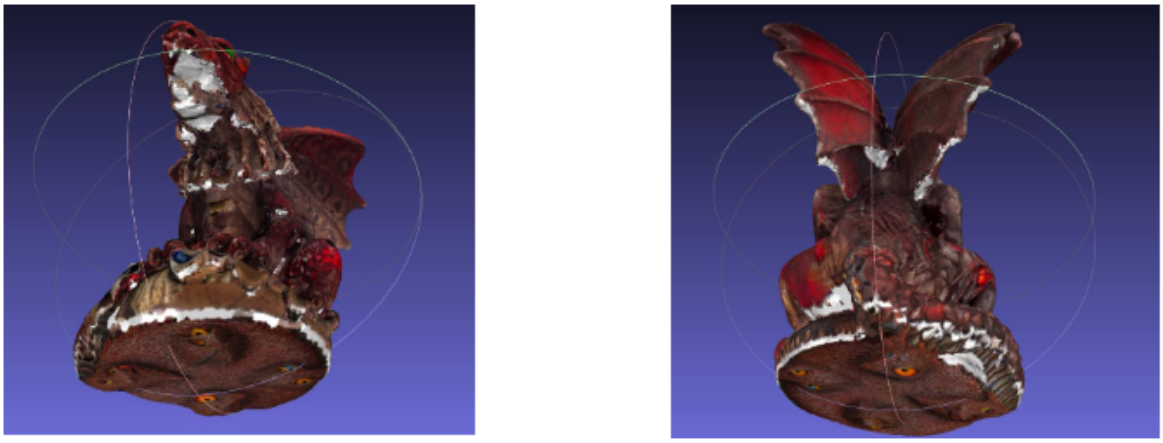

图 8. 自遮挡模型的运行结果

6 总结与展望

在本文中，我们提出了一种新的方法，Text2Tex，用于从给定的文本提示合成高质量的 3D 网格纹理。我们的方法利用深度感知图像的扩散模型，从多个视点逐步生成高分辨率的部

分纹理。

为了避免在视点间积累不一致和拉伸的伪影，我们动态地将渲染视图分割成一个生成mask，有效地引导扩散模型生成和更新相应的部分纹理。此外，我们提出了一种自动视点序列生成方案，该方案利用生成掩码来自动确定下一个最佳视图来优化生成的纹理。大量的实验表明，我们的方法可以有效地合成各种物体几何形状的一致性和高度详细的 3D 纹理，而无需额外的人工操作。

尽管这种方法能够产生高质量的 3D 纹理，但同时也发现它会倾向于生成带有阴影效果的纹理。虽然这个问题可以通过仔细调整输入提示来解决，但这样做需要额外的人力工程工作，并且可能无法很好地扩展到大规模生成目标。一个潜在的解决方案是微调扩散模型，以消除纹理中的阴影。

参考文献

[1] Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 3davatargan: Bridging domains for personalized editable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4552–4562, 2023.

[2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018.

[3] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.

[4] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021.

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

[6] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D 3 net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In *European Conference on Computer Vision*, pages 487–505. Springer, 2022.

[7] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021.

[8] Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18109–18119, 2023.

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019.

[10] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023.

[11] Zezhou Cheng, Menglei Chai, Jian Ren, Hsin-Ying Lee, Kyle Olszewski, Zeng Huang, Subhransu Maji, and Sergey Tulyakov. Cross-modal 3d shape generation and manipulation. In *European Conference on Computer Vision*, pages 303–321. Springer, 2022.

[12] Angela Dai, Yawar Siddiqui, Justus Thies, Julien Valentin, and Matthias Nießner. Spsg: Self-supervised photometric scene generation from rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1747–1756, 2021.

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[14] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022.

[15] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[17] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022.

[18] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021.

[19] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.

[20] Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey Tulyakov. Infinicity: Infinite-scale city synthesis. *arXiv preprint arXiv:2301.09637*, 2023.

[21] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021.

[22] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023.

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[24] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[25] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[26] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[28] Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Jian Ren, Hsin-Ying Lee, Menglei Chai, and Sergey Tulyakov. Unsupervised volumetric animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4658–4669, 2023.

[29] Yawar Siddiqui, Justus Thies, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Texturify: Generating textures on 3d shape surfaces. In *European Conference on Computer Vision*, pages 72–88. Springer, 2022.

[30] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[31] Edward J Smith and David Meger. Improved adversarial systems for 3d object generation and reconstruction. In *Conference on Robot Learning*, pages 87–96. PMLR, 2017.

[32] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[33] Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Learning descriptor networks for 3d shape synthesis and analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8629–8638, 2018.

[34] Rui Yu, Yue Dong, Pieter Peers, and Xin Tong. Learning texture generators for 3d shape collections from internet photo sets. In *British Machine Vision Conference*, 2021.

[35] Song-Hai Zhang, Yuan-Chen Guo, and Qing-Wen Gu. Sketch2model: View-aware 3d modeling from single free-hand sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6012–6021, 2021.