E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2025)05-1238-19

论文引用格式:Lu L H, Zhang X H, Wei H, Li R Y, Du G G and Wang B Q. 2025. Text-guided 3D editing based on neural radiance fields and 3D Gaussian splatting: a review. Journal of Image and Graphics, 30(5):1238-1256(卢丽华, 张晓辉, 魏辉, 李茹杨, 杜国光, 王斌强. 2025. 以神经辐射场和三维高斯泼溅为基础的文本指导三维编辑综述. 中国图象图形学报, 30(5):1238-1256)[DOI:10.11834/jig.240589]

以神经辐射场和三维高斯泼溅为基础的文本指导三维编辑综述

卢丽华*,张晓辉,魏辉,李茹杨,杜国光,王斌强

山东海量信息技术研究院,济南　250101

摘　要:文本引导的三维编辑可以根据目标文本的引导,改变现有三维资产的几何形状和外观,从而创建多样化和高质量的三维资产。先进三维神经表示、文本引导图像生成与编辑等一系列关键技术的出现和发展,推动了文本引导三维编辑的进步。本文主要聚焦于基于神经辐射场和三维高斯泼溅的文本指导三维编辑的最新进展,并从方法本质与编辑能力两个维度对现有研究进行梳理与总结。具体地,本文将现有研究按照编辑约束,分为无约束、隐式约束和显式约束3个类别,以深入剖析各方法本质。此外,本文还从编辑类型(如几何、外观)、编辑范围(如物体、场景)、编辑鲁棒性(如全局或局部可控性)等多个方面,对现有研究的编辑能力进行了探讨。最后,本文分析了当前研究所面临的挑战,并展望了未来潜在的研究方向。

关键词:文本指导;三维编辑;神经辐射场(NeRF);三维高斯泼溅(3GS);编辑约束;三维编辑能力

Text-guided 3D editing based on neural radiance fields and 3D Gaussian splatting: a review

Lu Lihua*, Zhang Xiaohui, Wei Hui, Li Ruyang, Du Guoguang, Wang Binqiang

Shandong Academy of Massive Information Technology, Ji'nan　250101, China

Abstract: Artificial intelligence-generated content (AIGC), which refers to the use of artificial intelligence (AI) technology to generate digital content, such as text, images, videos, and three-dimensional (3D) assets, has developed rapidly in the past few years, triggering a technological revolution. In the field of 3D AIGC, text-guided 3D editing has research significance and application value. In accordance with the guidance of the target text, it can change the geometry and appearance of existing 3D assets, thereby creating diversified and high-quality 3D assets. Compared with other guiding conditions, such as reference images and sketches, the 3D content editing paradigm guided by natural language has the advantages of friendly interaction, high efficiency, and strong practicability. This paradigm also has wide application potential in virtual/augmented reality, automatic driving, robots, and other fields. In recent years, the emergence and development of a series of key technologies, such as advanced neural representation, generative models, and text-guided image generation and editing, have led to significant progress in text-guided 3D editing and achieved certain outcomes. However, editing 3D content with text guidance remains a challenging task. Unlike the text-guided 3D generation task of generating 3D assets

收稿日期:2024-10-11;**修回日期**:2025-01-14;**预印本日期**:2025-01-21

*通信作者:卢丽华　roryuna@126.com

基金项目:山东省自然科学基金创新发展联合基金项目(ZR2022LZH002)

Supported by:Shandong Provincial Natural Science Foundation (ZR2022LZH002)

from zero, text-guided 3D editing edits the existing 3D assets and changes their geometric structure and appearance, among others, to obtain a new asset that conforms to the description of the target text. In the process of 3D editing, the core problem is to ensure that the nonediting areas are not affected while completing the task that meets the requirements of the target text. Second, it is difficult to correctly understand the target text and edit 3D assets that are semantically consistent with the target text, especially when the target text describes complex scenes, including multiple objects and different attributes. Furthermore, selecting 3D representations that are suitable for editing is a complex task, and both explicit (e. g. , voxels and grids) and implicit (e. g. , neural radiation fields and distance functions) representations have advantages and disadvantages in terms of representation ability and efficiency. Finally, the lack of a large dataset of text-3D assets and the inconsistency of multiple perspectives make text-guided 3D editing more challenging. In recent years, neural radiance field and 3D Gaussian splatting have been proposed. Due to their advantages, such as continuity and high photorealistic rendering, significant progress has been made in the field of high-quality 3D reconstruction and rendering of scenes. With large pretrained text-image alignment models, neural radiance fields have also been extended to text-guided 3D generation. Therefore, a simple way to implement text-guided 3D editing is to finetune the pretrained text-guided 3D generation model and modify the geometry or appearance of the 3D asset, among many other processes, so that it meets the new target text description. Earlier methods supervised the adjustment of the neural radiance fields by contrast language-image pretraining loss to align it with the new target text. Recent methods mostly utilize score distillation sampling loss optimization to edit neural radiance fields. However, this approach based on fine-tuning generation models can only change 3D assets globally and does not support fine-grained 3D editing. At the same time, the emergence of large text-image datasets and pretrained text-image alignment models has promoted the flourishing development of text-guided image editing techniques. Representative image editing techniques are introduced into 3D editing, which is a promising direction to solve the task of text-guided 3D editing. This editing paradigm avoids the need for text-3D data pairs by elevating 2D image editing to neural radiance fields, thereby enabling key advances in text-guided 3D editing. Early methods have conducted image editing on images rendered by the existing 3D models to conform to the target text, using the edited image to reconstruct the target 3D models to complete 3D editing meeting the target text. Subsequent methods further improve the editing quality and efficiency through multiview consistent editing and generalized editing. However, such methods rely on the ability of text to guide image editing and can only use image editing to provide implicit constraints without explicit control of the 3D editing process, which is not ideal for high-quality 3D editing. To achieve more accurate editing, recent research work has focused on introducing explicit editing constraints in the editing process, thus limiting 3D editing to the editable area and avoiding unnecessary editing while meeting the requirements of the target text. These methods can automatically determine the editing region from the semantic correspondence between the target text and the image, thus enabling impressively high-quality 3D editing. In view of the significant advances mentioned above, the above literature must be systematically summarized and analyzed for researchers interested in the field of text-guided 3D editing. This paper focuses on the latest advancements in text-guided 3D editing based on neural radiance fields and 3D Gaussian splatting, summarizing existing research from the aspects of methodological essence and editing capabilities. Specifically, this paper categorizes current research into three types according to their editing constraints: unconstrained, implicit constraints, and explicit constraints, to deeply analyze the essence of each method. In addition, the paper discusses the editing capabilities of these methods from various perspectives, including types of editing (e. g. , geometry and appearance), scope of editing (e. g. , objects and scenes), and editing robustness (e. g. , global or local controllability). Finally, the paper analyzes the challenges faced by current research and offers insights and prospects for potential future research directions. In summary, the contributions of this paper are as follows: 1) it offers the first review of text-guided 3D editing based on neural radiance field and 3D Gaussian splatting, 2) it provides a set of effective classification criteria to summarize the existing research work from the essence of the methods, and 3) it discusses the 3D editing capabilities of existing studies based on the principle of effective classification.

Key words: text guidance; 3D editing; neural radiance fields(NeRF); 3D Gaussian splatting(3GS); editing constraints; 3D editing capacity

0 引 言

人工智能生成内容(artificial intelligence generated content, AIGC),是指利用人工智能(artificial intelligence,AI)技术来生成文本、图像、视频和三维资产等数字内容,在过去几年快速发展,引发技术革命。在三维AIGC领域,文本指导的三维编辑是一个具有研究意义和应用价值的方向。其根据目标文本指导,改变已有三维资产的几何、外观等,可以再创造多样化与高质量的三维资产。相比于其他指导条件(如参考图像、草图等),由自然语言指导的三维内容编辑范式,具有交互友好、效率高和实用性强等优点,在虚拟/增强现实、自动驾驶和机器人等多个领域具有广泛的应用潜力(Choudhary等,2024;Tseng等,2022)。近年来,先进神经表征、生成模型以及文本指导的图像生成与编辑等一系列关键技术的出现与发展,促进了文本指导的三维编辑的显著进步,取得了一定的研究成果(Haque等,2023;Kamata等,2023;Chen等,2024c;He等,2024)。但是,通过文本指导编辑三维内容仍是具有挑战性的工作。

不同于从零生成三维资产的文本指导三维生成任务,文本指导的三维编辑是对现有三维资产进行编辑,改变其几何结构、外观等,得到符合目标文本描述的新资产。在三维编辑过程中,核心问题是在完成符合目标文本要求的编辑的同时,保证非编辑区域不受影响。其次,正确理解目标文本,编辑得到与目标文本语义一致的三维资产是困难的,特别是当目标文本描述包含多个对象、不同属性等复杂场景时。此外,选择适合编辑的三维表征是复杂的,显式表征(如体素、网格等)与隐式表征(如神经辐射场、符号距离函数等)在表征能力、效率方面各有优缺点。最后,缺少大型文本—三维资产的数据集、多视角不一致等问题,都使得文本指导的三维编辑更具有挑战性。

近几年,神经辐射场(neural radiance field, NeRF)(Mildenhall等,2020)和三维高斯泼溅(3D Gaussian splatting,3GS)(Kerbl等,2023)相继提出,由于其连续性、高真实感渲染等优点,在高质量场景三维重建与渲染领域取得了重大进展(Mildenhall等,2020;Yu等,2021)。随着大型预训练的文本—图像对齐模型,如,对比语言—图像预训练(contrast language-image pre-training, CLIP)(Radford等,2021)、去噪扩散概率模型(denoising diffusion probabilistic model,DDPM)(Ho等,2020)等的出现,神经辐射场被扩展应用于文本指导的三维生成领域(Sanghi等,2022;Poole等,2022)。DreamFusion(Poole等,2022)提出得分蒸馏采样(score distillation sampling,SDS),将二维扩散模型提升到神经辐射场,解决了三维数据稀缺的问题,促使基于神经辐射场的三维生成取得长足进步。因此,实现文本指导三维编辑的一个朴素做法是,微调预训练好的文本指导三维生成模型,修改三维资产的几何或外观等,使其重新满足新的目标文本描述。早期方法(Wang等,2022;Wang等,2024)通过CLIP监督调整神经辐射场,使其对齐新的目标文本。受到DreamFusion(Poole等,2022)启发,新的方法(Kamata等,2023;Sella等,2023;Palandra等,2024)大多利用SDS优化编辑神经辐射场。然而,这种基于微调生成模型的方法,只能全局改变三维资产,不支持细粒度三维编辑。

另一方面,大型文本—图像数据集和预训练文本—图像对齐模型的出现,促进了文本指导的图像编辑技术的繁荣发展(Brooks等,2023;Ruiz等,2023;Zhang等,2023;张浩宇等,2022)。将以InstructPix2Pix(Brooks等,2023)、DreamBooth(Ruiz等,2023)等为代表的图像编辑技术引入到三维编辑,是解决文本指导三维编辑这一任务极具前景的方向。这一编辑范式(Haque等,2023;Kamata等,2023;Song等,2023;Karim等,2024)通过将二维图像编辑提升到三维神经辐射场,避免了对文本—三维数据对的需求,实现了文本指导三维编辑的关键进步。典型地,InstructN2N(Haque等,2023)利用文本指导的图像编辑方法InstructPix2Pix(Brooks等,2023)对已有神经辐射场渲染得到的图像进行符合目标文本的图像编辑,利用编辑后的图像重建目标神经辐射场,完成满足目标文本的三维编辑。后续方法通过多视角一致性编辑(Song等,2023)、广义编辑(Karim等,2024)等进一步提高了编辑质量和效率。然而,这类方法依赖于文本指导图像编辑的能力,只能利用图像编辑提供隐式约束,无法显式控制三维编辑过程,这对于高质量的三维编辑来说并不理想。

为了达到更精确的编辑,最近的研究工作致力

于在编辑过程中引入明确的编辑约束，将三维编辑限定在可编辑区域，在满足目标文本要求的同时避免不必要的编辑。这些方法可以从目标文本和图像之间的语义对应关系中自动确定编辑区域。例如，Vox-E（Sella 等，2023）、DreamEditor（Zhuang 等，2023）等方法在三维编辑过程中引入二维交叉注意力图，实现局部精细编辑。这些方法实现了令人印象深刻的高质量三维编辑。

鉴于上述重大进展，有必要系统地总结与分析，供对文本指导的三维编辑领域感兴趣的研究者查阅。已有一些相关综述论文（Li 等，2024；Foo 等，2025；Chen 等，2024b）对文本指导的三维生成与编辑方法进行了调研与分析。Foo 等人（2025）总结了基于 AI 技术的各种数据模式生成，如图像、视频、文本、音频和三维物体等，其中文本指导的三维生成和编辑只做了简要总结。Li 等人（2024）全面讨论了三维生成的相关技术以及最新进展。Chen 等人（2024b）对文本和图像指导的三维神经风格化进行了调查，但其更多地关注图像指导的三维编辑方法，缺乏对文本指导的三维编辑的详细分类和讨论。上述综述侧重对三维生成方法的介绍与分析，且没有聚焦在基于神经辐射场和三维高斯的三维编辑方法最新进展。本文深入调研与讨论了文本指导的三维编辑分类和三维编辑能力。具体地，将该领域的最新研究进展依据编辑约束分为无约束、隐式约束和显式约束 3 类，并按照类别分析其方法本质。此外，本文从编辑类型（如几何、外观）、编辑粒度（如物体、场景）和编辑鲁棒性（如全局或可控性）等多个方面，对方法所能达到的三维编辑能力进行总结与讨论。综上所述，本文贡献如下：1）首次提出基于神经辐射场和三维高斯泼溅的文本指导三维编辑综述；2）提出有效的分类标准，用以从方法本质对比总结现有研究工作；3）在有效分类的基础上，探讨了现有研究的三维编辑能力。

1 背景知识

1.1 文本指导的图像生成与编辑

对于文本指导的图像生成和编辑，早期的研究（Reed 等，2016；Patashnik 等，2021）将生成对抗网络（generative adversarial network，GAN）和 CLIP 结合，在文本的指导下执行图像操作。最近，随着预训练扩散模型的提出与发展，基于扩散模型的方法（Rombach 等，2022；Zhang 等，2023）已经能够生成高分辨率和多样化的图像。开创性工作 GLIDE（Nichol 等，2022）和 Imagen（Saharia 等，2022）使用带有文本指导的扩散模型来控制图像生成过程，并在像素空间中生成图像。而以 Stable Diffusion（Rombach 等，2022）为代表的方法则将扩散模型引入到潜在空间，更高效地生成高视觉保真度的图像。为了在生成过程中实现更精确的控制，ControlNet（Zhang 等，2023）等代表性方法在文本之外加入了附加条件（如边界框、分割图）。例如，ControlNet（Zhang 等，2023）在基础扩散模型的每个编码器中加入新的控制结构，以支持各种特定输入条件，如深度图、法线和草图等。

在文本指导的图像编辑领域，一些方法致力于编辑现有图像，并生成可以保持某种主题或身份的新图像。给定同一主题的几幅图像，DreamBooth（Ruiz 等，2023）通过使用少量参考图像调整 Imagen（Saharia 等，2022）模型，将主题绑定到唯一标识符，并嵌入到扩散模型中。Instruct-Pix2Pix（Brooks 等，2023）侧重使用提供的蒙版进行局部编辑，其提出了一种基于文本指令编辑图像的方法，接受原始图像和目标文本作为条件，并在大型文本指令—图像数据集上进行训练，在完成符合文本指令描述编辑的同时，保持非编辑区域在编辑前后的一致性。更多文本指导的图像生成和编辑研究可参考刘安安等人（2024）的论文。

将文本指导的二维图像编辑提升到三维，可以充分利用大型预训练的视觉语言对齐模型，同时避免收集文本—三维资产数据集。Instruct N2N（Haque 等，2023）通过 InstructPix2Pix（Brooks 等，2023）编辑原始三维场景的渲染图像，然后使用编辑后的图像重建目标三维场景。随后，一些方法在编辑的三维一致性（Song 等，2023）、局部编辑（He 等，2024；Dong 和 Wang，2024）以及高效编辑（Karim 等，2024；Park 等，2024）等方面进行了进一步改进。还有一些方法倾向于将基于文本的二维编辑模型与分数蒸馏采样相结合来优化三维编辑。DreamEditor（Zhuang 等，2023）首先利用 DreamBooth（Ruiz 等，2023）在融合网格的神经辐射场上自动识别编辑区域，然后利用 SDS 提升 DreamBooth 的编辑结果，进行局部三维编辑。

1.2 三维表示

显式的三维表示，如体素、点云和网格，可以直观地表示物体的表面或轮廓，但消耗内存大，不适合表示大场景。为了解决上述问题，神经辐射场（NeRF）（Mildenhall等，2022）被提出，其在内存和拓扑上更加灵活和轻量级，可以用来表示大规模场景。具体地，神经辐射场利用全连接深度网络来表示三维空间，网络输入三维位置信息$\boldsymbol{X}=(x,y,z)$和二维视角方向$\boldsymbol{d}=(\theta,\phi)$，输出体密度和视角依赖的颜色信息，形式化表示为

$$F_{\Theta}:(\boldsymbol{X},\boldsymbol{d})\rightarrow(\boldsymbol{c},\sigma) \tag{1}$$

式中，F_{Θ}表示多层全连接网络。神经辐射场通过沿相机光线查询五维坐标来合成新视图，并使用体绘制技术将输出颜色和密度投影到图像中。具体地，给定相机位置$\boldsymbol{o}$，相机光线$\boldsymbol{r}(t)=\boldsymbol{o}+t\boldsymbol{d}$，和近平面与远平面的边界$t\in[t_n,t_f]$，在相机光线上采样$N$个点，则投影得到的颜色为

$$\hat{C}(\boldsymbol{r})=\sum\Omega_i\left(1-\exp\left(-\rho_i\delta_i\right)\right)\boldsymbol{c}_i \tag{2}$$

式中，ρ_i和$\boldsymbol{c}_i$分别表示第i个采样点的体密度和颜色，$\Omega_i=\exp\left(-\sum_{j=1}^{i-1}\rho_i\delta_i\right)$表示沿射线的累积透过率，$\delta_i$相邻采样点之间的距离。

但是，神经辐射场通常需要密集采样，训练和推理时间长。有方法结合显式三维表示对神经辐射场进行改进，Plenoxels（Fridovich-Keil等，2022）、Instant-NGP（Müller等，2022）、DVGO（Sun等，2022）和TensoRF（Chen等，2022）倾向于使用体素，DreamEditor（Zhuang等，2023）使用了网格。直接体素网格优化（direct voxel grid optimization，DVGO）（Sun等，2022）采用稠密体素网格来表示三维几何结构，并使用了一个浅层网络的特征体素网格来捕捉复杂的、视点依赖的外观。将训练时间减少到约15 min，同时保持了与NeRF相媲美的重建质量。英伟达提出即时神经图形基元（instant neural graphics primitives，Instant-NGP）（Müller等，2022），使用一个小规模的网络来实现全连接网络，该网络由多分辨率哈希编码技术进行优化。多分辨率结构有助于GPU并行计算，并且能够通过消除哈希冲突来减少计算量，在保证重建质量的同时，将训练时间缩短到几秒钟。TensoRF（Chen等，2022）通过将每个体素网格替换为平面和向量的张量分解来实现类似的模型压缩和加速。此外，三平面表示（Triplane）（Chan等，2022）则是采用另一种方式进行加速，其将三维空间分解为3个正交平面（如XY、XZ和YZ平面），并在这些平面上表示三维形状的特征。离散的显式表示有利于减少计算，有效地利用隐式表示的网络容量，加快神经辐射场的训练和渲染速度。三维高斯泼溅将三维空间表示为基于点云的各向异性高斯函数，实现实时渲染。

三维高斯泼溅（3D GS）（Sun等，2022）用基于点云的各向异性高斯分布表示三维场景，通过对三维高斯分布进行交替优化和密度控制，特别是优化各向异性协方差，以实现对场景的精确表示，并可以在1 080 p分辨率下实时渲染三维场景。

1.3 损失函数

文本指导的三维编辑优化主要是用3种损失函数：基于对比语言图像预训练（CLIP）的损失函数、基于分数蒸馏采样（SDS）的损失函数和基于重建的损失函数。

1）基于CLIP的损失函数。CLIP（Saharia等，2022）通过文本和图像编码器将文本和图像编码到隐式空间进行对齐，弥合了文本和图像之间的语义差距，早期的方法（Wang等，2022；Wang等，2024）通过基于CLIP的损失函数优化神经辐射场，使其对齐新的目标文本，实现编辑。基于CLIP的损失函数主要包括两种类型（Wang等，2022；Wang等，2024）：CLIP文本—图像相似度（CLIP text-image similarity）和CLIP方向文本—图像相似度（CLIP directional text-image similarity）。

CLIP文本—图像相似度$\mathcal{L}_{\text{sim}}(\boldsymbol{I},\boldsymbol{T})$通过在CLIP隐式空间中计算余弦距离来度量输出图像$\boldsymbol{I}$与目标文本提示$\boldsymbol{T}$之间的语义相似度，定义为

$$\mathcal{L}_{\text{sim}}(\boldsymbol{I},\boldsymbol{T})=1-\mathcal{D}_{\cos}\left(E_{\text{txt}}(\boldsymbol{T}),E_{\text{img}}(\boldsymbol{I})\right) \tag{3}$$

式中，$\mathcal{D}_{\cos}$表示计算余弦距离，E_{txt}和E_{img}分别表示CLIP的文本和图像编码器。

CLIP方向文本—图像相似度$\mathcal{L}_{\text{dir}}(\boldsymbol{I},\boldsymbol{T})$测量输入到输出图像变化方向与源文本和目标文本提示变化方向之间的余弦距离，定义为

$$\mathcal{L}_{\text{dir}}(\boldsymbol{T}_{\text{t}},\boldsymbol{T}_{\text{s}},\boldsymbol{I}_{\text{t}},\boldsymbol{I}_{\text{s}})=1-\mathcal{D}_{\cos}\left(\Delta(\boldsymbol{T}),\Delta(\boldsymbol{I})\right) \tag{4}$$

式中，$\Delta(\boldsymbol{T})=E_{\text{txt}}(\boldsymbol{T}_{\text{t}})-E_{\text{txt}}(\boldsymbol{T}_{\text{s}})$，$\Delta(\boldsymbol{I})=E_{\text{img}}(\boldsymbol{I}_{\text{t}})-E_{\text{img}}(\boldsymbol{I}_{\text{s}})$，$\boldsymbol{I}_{\text{t}}$和$\boldsymbol{T}_{\text{t}}$分别表示输出图像和其对应的文本提示，$\boldsymbol{I}_{\text{s}}$和$\boldsymbol{T}_{\text{s}}$分别表示输入图像和其对应的文本描述。

2）基于 SDS 的损失函数。DreamFusion（Poole 等，2022）提出得分蒸馏采样（SDS），将预先训练的文本到图像扩散模型中的先验知识提升到神经辐射场，奠定了基于SDS损失函数优化的基础。$\mathcal{L}_{\mathrm{SDS}}$ 具体定义为

$$\nabla_\theta \mathcal{L}_{\mathrm{SDS}}(\phi, \boldsymbol{x}=g(\theta)) = E_{t,\epsilon}\left[\left(\hat{\epsilon}_\phi(\boldsymbol{x}_t, y, t)-\epsilon\right)\frac{\partial \boldsymbol{x}}{\partial \theta}\right] \tag{5}$$

式中，图像 $\boldsymbol{x}=g(\theta)$ 由一个可微图像生成器 g 得到，θ 为对应图像生成器的参数，在 DreamFusion 中采用基于 NeRF 的可微渲染；$\hat{\epsilon}_\phi(\boldsymbol{x}_t, y, t)$ 为预训练的文本指导图像生成扩散模型预测得到的噪声，ϕ 为文本指导图像生成扩散模型的参数，ϵ 为扩散模型扩散过程中添加的噪声真值，t 为扩散模型去噪过程的时间步，y 为对应的文本提示，E 为期望值。SDS 损失通过计算预测噪声和噪声真值的差值进行梯度反向传播，优化基于NeRF的三维模型。

3）基于重建的损失函数。有些方法（Haque 等，2023；Song 等，2023）使用文本指导的图像编辑模型直接对源三维模型渲染得到的图像进行编辑，基于这些编辑后的图像，利用重建损失优化得到编辑后的三维模型。具体重建损失计算为

$$\mathcal{L}_{\mathrm{rec}} = \left\| \boldsymbol{I}_{\mathrm{r}} - \hat{\boldsymbol{I}}_{\mathrm{r}} \right\|_2^2 \tag{6}$$

式中，$\hat{\boldsymbol{I}}_{\mathrm{r}}$ 为编辑后三维模型渲染得到的图像，$\boldsymbol{I}_{\mathrm{r}}$ 为经过文本指导图像编辑模型编辑后的图像。重建损失利用编辑后三维模型渲染得到图像与文本指导图像编辑模型编辑后的图像之间的差值进行梯度反向传播，优化编辑后的三维模型。

2 基于神经辐射场和三维高斯泼溅的文本指导三维编辑

文本指导的三维编辑旨在对现有三维资产进行修改，以满足目标文本的要求。其任务的起点是已存在的三维资产，在编辑过程中需要根据文本指导完成指定的三维编辑，同时保证非编辑区域在编辑前后保持不变，引入编辑约束是解决这一难题的有效途径。

本文将现有研究按照编辑约束，分为无约束、隐式约束和显式约束 3 种，并对其方法本质进行具体分析。图 1 总结了基于神经辐射场和三维高斯泼溅的文本指导三维编辑方法近几年的进展，并按照不同类别进行展示。此外，本节从三维表示（如基于神经辐射场、基于融合体素的神经辐射场、基于融合网格的神经辐射场、基于三平面、基于三维高斯泼溅）和损失函数等方面，对现有研究进行了进一步总结与分析，具体细节参考表 1。

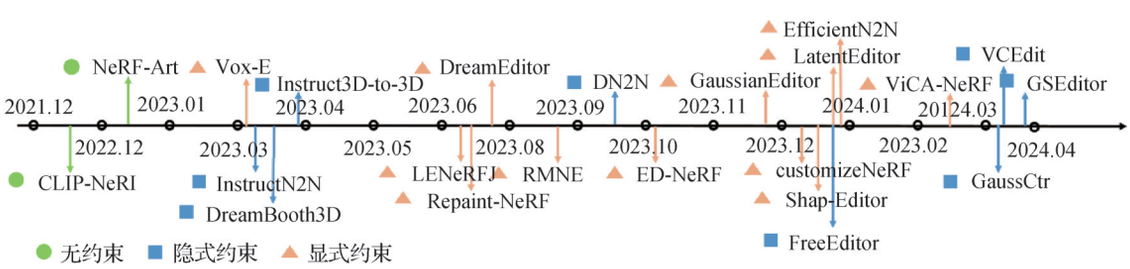

图 1　基于神经辐射场和三维高斯泼溅的文本指导三维编辑方法进展总结

Fig. 1　A summary of the progress of text-guided 3D editing methods based on neural radiance fields and 3D Gaussian splatting

2.1 无约束的文本指导三维编辑

如图 2 所示，无约束的文本指导三维编辑利用基于 CLIP 的损失函数，优化编辑模块，从而达到编辑三维模型的目的。但是，由于 NeRF 几何和外观表示的纠缠，以及 CLIP 损失优化能力受限，直接利用 CLIP 损失优化 NeRF 是有难度的。NeRF-Art（Wang 等，2024）和 CLIP-NeRF（Wang 等，2022）从解缠外观和几何表示以及优化 CLIP 损失的角度解决上述问题。

NeRF-Art（Wang 等，2024）提出了一种基于 CLIP 的全局—局部对比损失，该方法鼓励编辑结果更接近目标文本，远离 CLIP 嵌入空间中预定义的负样本，并结合定向 CLIP 损失，使全局结构和局部细节都遵循目标文本的语义。此外，在改变 NeRF 体密度场时，采用权值正则化来抑制混浊伪影和几何噪声。CLIP-NeRF（Wang 等，2022）设计了一个解纠缠的条件 NeRF 框架，将三维编辑分解为几何编辑和外观编辑。其中，几何和外观代码用于控制体积场和颜色场，然后提出几何和外观映射器，在几何和外观变形中引入文本指导，与解纠缠的条件 NeRF 配

表1　基于神经辐射场和三维高斯泼溅的文本指导三维编辑方法汇总表

Table 1　A summary of text-guided 3D editing methods based on neural radiance fields and 3D Gaussian splatting

编辑约束		方法	三维表示	损失函数
无约束		CLIP-NeRF(Wang等,2022)	神经辐射场	CLIP
		NeRF-Art(Wang等,2024)	神经辐射场	CLIP
隐式约束		InstructN2N(Haque等,2023)	神经辐射场	重建损失
		DreamBooth3D(Raj等,2023)	神经辐射场	重建损失
		EfficientN2N(Song等,2023)	神经辐射场	重建损失
		DN2N(Fang等,2023)	神经辐射场	重建损失
		FreeEditor(Karim等,2024)	神经辐射场	重建损失
		Instruct3D-to-3D(Kamata等,2023)	直接体素网格优化	SDS
		GSEditor(Palandra等,2024)	三维高斯泼溅	SDS
		GaussCtrl(Wu等,2024)	三维高斯泼溅	重建损失
		VCEditor(Wang等,2025)	三维高斯泼溅	重建损失
显示约束	二维交叉注意力图	Shap-Editor(Sella等,2023)	神经辐射场	重建损失、SDS
	二维相关图	RMNE(Mirzaei等,2023)	神经辐射场	重建损失
	二维交叉注意力图	ViCA-NeRF(Dong和Wang,2024)	神经辐射场	重建损失
	二维分割图	customizeNeRF(He等,2024)	神经辐射场	重建损失、SDS
	二维相关图	LatentEditor(Khalid等,2024)	神经辐射场	重建损失
	二维交叉注意力图	Vox-E(Sella等,2023)	直接体素网格优化	SDS
	二维分割图	ED-NeRF(Park等,2024)	即时神经图形基元	重建损失、SDS
	二维分割图	RepaintNeRF(Zhou等,2023)	即时神经图形基元	重建损失、CLIP、SDS
	二维交叉注意力图	LENeRF(Hyung等,2023)	三平面	CLIP
	二维分割图	DreamEditor(Zhuang等,2023)	融合网格的神经辐射场	SDS
	二维分割图	GaussianEditor(Chen等,2024c)	三维高斯泼溅	重建损失

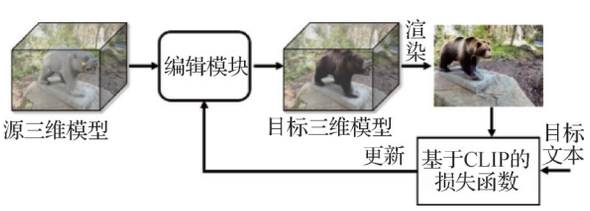

图2　无约束的文本指导三维编辑框架图

Fig. 2　The frame diagram of unconstrained text-guide 3D editing

合,在优化CLIP损失的情况下实现单独的几何和外观编辑。

2.2　隐式约束的文本指导三维编辑

如图3所示,隐式约束的文本指导三维编辑方法的核心为文本指导的图像编辑模型,主要通过交替进行目标三维模型的渲染图像编辑和目标三维模型更新,并以迭代优化的方式实现三维编辑。InstructN2N(Haque等,2023)和Instruct3D-to-3D(Kamata等,2023)奠定了此类方法的基础。

InstructN2N(Haque等,2023)提出了一种基于文本指令的三维编辑方法(如图4所示)。给定场景的源NeRF模型以及目标文本指令,该方法使用基于文本指令的图像编辑模型InstructPix2Pix(Brooks等,2023)来编辑目标NeRF渲染得到的图像,基于这些编辑后的图像,利用重建损失优化目标NeRF。此外,InstructN2N设计了一种迭代数据集更新策略,在不同视角上重复上述图像编辑和目标NeRF优化操作,以不同视角中传播图像编辑结果完成三维编辑。

不同于InstructN2N,Instruct3D-to-3D(Kamata

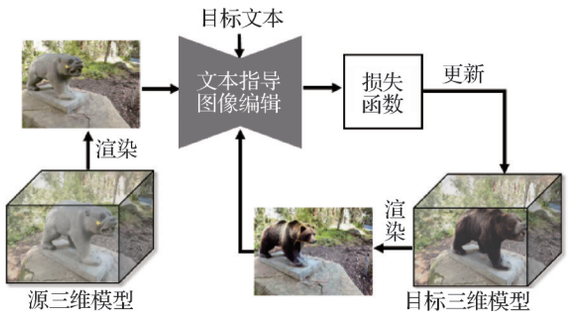

图 3　隐式约束下的文本指导三维编辑框架图

Fig. 3　The frame diagram of text-guide 3D editing under implicit constraints

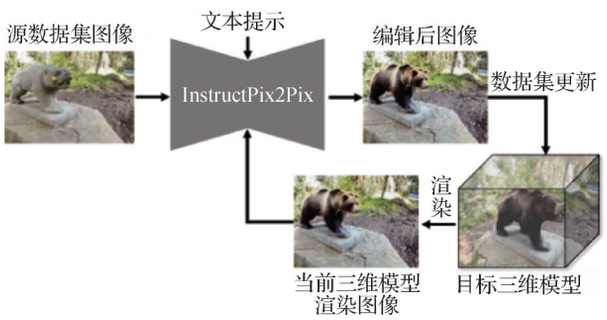

图 4　InstructN2N（Haque 等，2023）流程框架图

Fig. 4　The frame diagram of InstructN2N（Haque et al. ，2023）

等，2023）采用直接体素网格优化（DVGO）表示三维模型，根据目标文本指令，利用 InstructPix2Pix（Brooks 等，2023），对目标三维模型渲染得到的带噪声图像进行去噪，并利用 SDS 损失将去噪结果提升到三维，完成基于文本指令的三维编辑。通过从不同视角迭代执行此过程，获取符合目标文本指令的三维模型。此外，它还提出了一种动态缩放方案来动态调整几何转换的强度，使三维编辑过程更加可控和流畅。上述方法在编辑过程中需要迭代优化不同视角以实现三维编辑，训练效率低且可能导致多视角不一致。

为了解决上述问题，EfficientN2N（Song 等，2023）将迭代数据集更新策略替换为多视角图像联合编辑，在 InstructPix2Pix 基础上引入视角对应正则化，避免了迭代优化，提高编辑效率（如图 5 所示）。具体地，在任意两个视角间，将对应正则化应用于扩散模型，在去噪过程中对齐来自多个视角的样本，一次图像编辑操作可获取目标 NeRF 的多个渲染图像跨视角一致编辑结果，用以重建目标 NeRF。同时，视角一致性编辑使得 EfficientN2N 的速度是

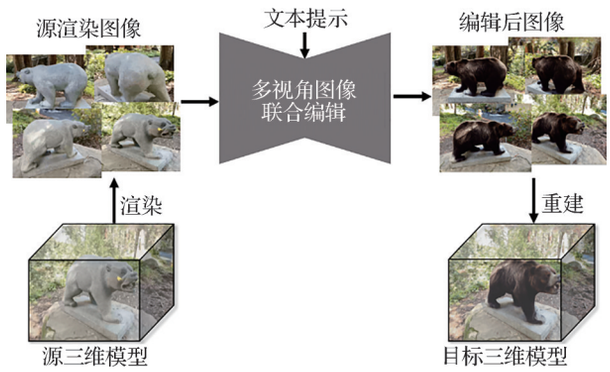

图 5　EfficientN2N（Song 等，2023）流程框架图

Fig. 5　The frame diagram of EfficientN2N（Song et al. ，2023）

InstructN2N（Haque 等，2023）的 10 倍。

DN2N（Fang 等，2023）和 FreeEditor（Karim 等，2024）通过训练广义 NeRF 扩展了文本指导三维编辑的泛化能力，针对新场景，允许用户在推理阶段无需再训练即可直接编辑三维资产，减少了编辑时内存消耗。DN2N（Fang 等，2023）首先过滤图像编辑中多视角一致性较差的结果，然后将剩余的不一致性视做去除噪声扰动的问题。具体来讲，DN2N 利用大语言模型和文本指导的图像编辑模型，构建具有相似扰动特性的数据对训练广义 NeRF，并在训练过程中引入跨视角正则化项，协助去除噪声扰动。在推理过程中，针对新场景以及新的目标文本，利用文本指导图像编辑模型获取一组编辑后的图像，并使用广义 NeRF 去除多视角不一致性，获取编辑后的任意视角图像。广义 NeRF 无需针对新场景进行再训练，提高了编辑的泛化性和效率。与上述方法不同的是，FreeEditor（Karim 等，2024）提出了一种“单视图编辑”方案，利用文本指导图像编辑模型编辑初始视角，并引入编辑转换器，利用自注意力和交叉注意力来强制视角内的一致性和从起始视角到目标视角的跨视角编辑迁移。在推理过程中，只需要编辑起始视角图像满足目标文本要求，即可构建目标 NeRF，避免了多视角不一致性，节省编辑时间。

DreamBooth3D（Raj 等，2023）从多阶段优化的角度解决多视角不一致性问题。其提出了一个可生成具有特定主题 3D 资产的生成框架，该框架可以根据给定的目标文本编辑 NeRF，同时保持编辑后的 NeRF 具有源模型的几何和外观身份。具体来说，设计了一种三阶段优化方案，在第 1 阶段和第 2 阶段对图像编辑扩散模型 DreamBooth（Ruiz 等，2023）进行部分和完全微调，生成多视角伪主体图像。在最后

阶段，用这些多视角图像对DreamBooth进一步微调，获取多视角DreamBooth，作为最终的文本指导图像编辑模型，配合SDS损失和重建损失优化目标NeRF。

与前述方法不同，GSEditor（Palandra等，2024）、GaussCtrl（Wu等，2024）和VCEdit（Wang等，2025）利用3D GS表示3D空间。GSEditor（Palandra等，2024）利用InstructPix2Pix对目标3D GS模型渲染得到的带噪声图像进行去噪，并利用SDS损失进行优化。在基本编辑完成后，使用DreamGaussian（Tang等，2024）对目标3D GS模型进行网格提取和纹理优化。GaussCtrl（Wu等，2024）使用深度引导和基于注意力的潜在编码对齐模块来实现多视图一致编辑。具体地，采用深度引导的图像编辑模型ControlNet（Zhang等，2023），利用深度图来强制多视角的几何一致性；同时，利用自注意力和交叉注意力机制对齐多视角图像的潜在编码，统一编辑后图像的外观。VcEdit（Wang等，2025）采用了迭代优化模式：交替进行编辑多视角渲染图像与更新3D GS模型；在编辑多个视角图像的过程中，通过将交叉注意力一致性模块引入到图像编辑扩散模型中，并利用编辑一致性模块直接校准图像编辑输出，减少多视角编辑不一致性。

总之，基于隐式约束的三维编辑，需要解决多视角图像编辑不一致、迭代优化等问题。EfficientN2N（Song等，2023）、DN2N（Fang等，2023）和FreeEditor（Karim等，2024）等方法从多视角图像联合编辑、训练广义NeRF等角度提出了优化方案。GSEditor（Palandra等，2024）、GaussCtrl（Wu等，2024）和VCEdit（Wang等，2025）利用3D GS表示三维空间，在改进算法的同时，借助于3D GS的高效表达与快速渲染能力，提高编辑质量与效率。但是，此类方法隐式约束来源于文本指导的图像编辑，编辑区域定位与三维编辑的准确性和质量受到图像编辑能力的限制，容易造成过度编辑或欠编辑等情况。

2.3 显式约束的文本指导三维编辑

为了得到更精确和细粒度的三维编辑，需要在编辑过程中引入显式约束，将三维编辑限制在可编辑区域中进行，在实现目标文本指定编辑的同时避免不必要的修改。如图6所示，基于显式约束的文本指导三维编辑方法，利用预训练的文本指导图像编辑模型，根据目标文本与渲染图像的语义对齐关系，自动获取二维交叉注意力图（2D cross-attention map）、二维分割图（2D segmentation map）、二维相关图（2D relevance map）等，这些编辑约束进一步与优化损失协作，在编辑约束指定的掩码区域施加梯度的反向传播优化三维模型。具体细节如表1所示。

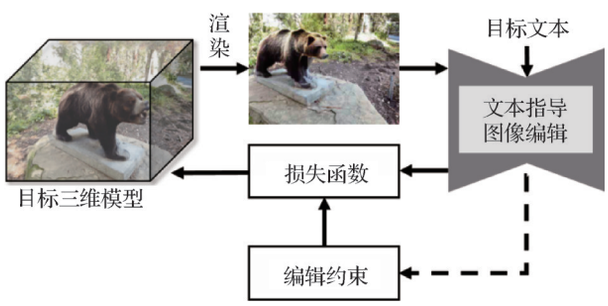

图6 显式约束下的文本指导三维编辑框架图

Fig. 6 The frame diagram of text-guided 3D editing under explicit constraints

RMNE（relevance map-guided NeRF editing）（Mirzaei等，2023）和LatentEditor（Khalid等，2024）采用InstructN2N（Haque等，2023）的迭代数据集更新策略，同时引入了二维相关图来避免InstructN2N中存在的过度编辑问题。其中二维相关图由带文本条件和无条件时图像编辑模型InstructPix2Pix（Brooks等，2023）预测的差异计算得来。具体地，RMNE（Mirzaei等，2023）（如图7所示）提出基于二维相关图的图像编辑，即在InstructPix2Pix基础上训练了一个相关图引导的文本指导图像编辑模型，用相关图作为二维遮罩定位可编辑像素，用以精确编辑目标NeRF渲染得到的图像和更新相关域渲染得到的相关图，并以此基础更新目标NeRF和相关域。LatentEditor（Khalid等，2024）将真实场景编码到潜在空间中，并在该潜在空间中迭代进行数据集更新和NeRF编辑。将二维相关图作为潜在空间中的二维掩码，指导潜在空间中图像的局部修改，使用编辑后的图像迭代更新潜在空间的NeRF，实现细粒度三维编辑。一旦训练好潜在的NeRF，就可以通过解码器对其进行解码。与InstructN2N相比，在潜在空间中完成三维编辑，使得编辑时间缩短为InstructN2N的五分之一。

GaussianEditor（Chen等，2024c）、customizeNeRF（He等，2024）和ED-NeRF（Park等，2024）等方法利用二维分割图定位可编辑区域。customizeNeRF（He等，2024）提出了一种局部—全局迭代编辑训练策

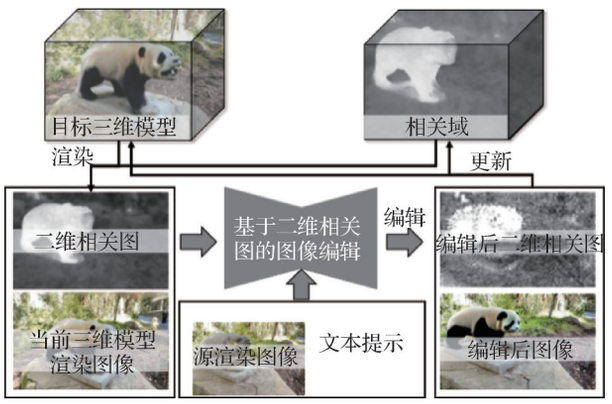

图 7　RMNE（Mirzaei 等，2023）流程框架图

Fig. 7　The frame diagram of RMNE（Mirzaei 等，2023）

略，可以在前景区域编辑和全场景编辑之间交替进行，在利用全局和局部 SDS 损失进行编辑优化的同时，利用 Grounded-SAM（Ren 等，2024）得到分割图作为伪前景掩码，约束编辑前后背景区域重建，从而做到编辑前景内容的同时保持背景区域不变。此外，设计了一个类引导的正则化，利用生成模型中的类先验来缓解图像编辑中不同视角之间的不一致问题。Repaint-NeRF（Zhou 等，2023）在第 1 阶段利用二维分割图作为伪标签，训练分离出要更改的部分。随后，在编辑三维场景与目标文本对齐的同时，重建了编辑后三维场景中的非编辑区域，使其与源模型相似。此外，还采用了 CLIP 损失，使编辑后的三维模型渲染图像中不可见的区域被背景信息填充。ED-NeRF（Park 等，2024）将三维场景的即时神经图形基元（Instant-NGP）表示编码到潜在空间，利用基于潜在空间的扩散模型进行 3D 编辑，以获得更快的编辑速度和高质量的编辑结果。它利用现有分割模型 SAM（segment anything model）（Kirillov 等，2023）通过文本指导分割目标区域而确定编辑区域掩码，并配合改进的损失函数一起监督潜在空间的三维编辑。具体地，它提出 Delta 去噪分数（delta denoising score，DDS）代替 SDS，其中 DDS 由目标文本和源文本对应的 SDS 分数相减得到，将其与二进制掩码相结合，得到掩码 DDS 损失函数，保证目标 NeRF 向着文本指定的方向优化。此外，其引入了一个额外的重建损失，以减少超出掩膜区域的编辑。Gaussian-Editor（Chen 等，2024c）使用 3D GS 表示三维场景，并利用 InstructPix2Pix（Brooks 等，2023）编辑 3D GS 的渲染图像，然后使用编辑后的图像更新目标 3D GS 模型。此外，此方法结合大型语言模型和分割模型，自动获取感兴趣的图像分割区域，并反投影到 3D GS，保证优化梯度回传只在特定 3D GS 上进行，实现精细化编辑。

与上述方法不同，Vox-E（Sella 等，2023）、DreamEditor（Zhuang 等，2023）、ViCA-NeRF（Dong 和 Wang，2024）和 Shap-Editor（Chen 等，2024a）利用二维交叉注意力图实现局部编辑。Vox-E（Sella 等，2023）利用直接体素网格优化（DVGO）来表示三维场景，在编辑过程中，以源三维场景模型初始化目标三维模型，并以 SDS 损失进行优化，以满足目标文本要求的三维编辑。对于全局编辑，Vox-E 引入体积正则化损失以强制编辑前后三维模型的体密度具有高度相关性，从而保证编辑前后的几何一致性。对于局部编辑，Vox-E 利用 Stable Diffusion（Rombach 等，2022）获取文本与渲染图像之间的交叉注意力图，进一步得到编辑与非编辑体素的掩码，用源体素模型替换编辑后非编辑区域体素结果，从而保证非编辑区域不受影响。然而，由于几何和纹理的耦合，Vox-E 受到噪声编辑结果的影响。DreamEditor（Zhuang 等，2023）首先将源三维模型提取为融合网格的神经场，解耦纹理和几何表示；然后，利用预训练的图像编辑模型 DreamBooth（Ruiz 等，2023）确定编辑区域，并利用 SDS 损失优化的编辑区域内三维模型的纹理和几何形状获得符合目标文本的三维编辑结果。Shap-Editor（Chen 等，2024a）利用文本指导三维生成方法 Shap-E（Jun 和 Nichol，2023）在潜在空间中完成三维编辑。为了实现局部精确编辑，利用 InstructPix2Pix 计算文本与图像之间的交叉注意力图，从而在编辑过程中利用重建损失约束非编辑区域的颜色和深度尽量不变。借助 Shap-E，在推理过程中无需重新训练，即可在大约 1 s 内完成三维编辑。但是，其编辑后三维模型的质量受限于 Shap-E 的生成能力。

ViCA-NeRF（Dong 和 Wang，2024）提出几何正则化和可学习正则化模块，将图像编辑从编辑视图传播到未编辑视图，从而增强视图一致性，避免迭代数据集更新。几何正则化利用深度图来建立图像对应关系，而可学习的正则化进一步对齐文本指导的图像编辑模型中编辑和未编辑图像之间的潜在代码，从而能够编辑关键视图并在整个空间中传播更新。这样做可以用多视图一致编辑的图像更新数据集，从而促进目标 NeRF 的重建。为了实现局部编辑，

利用SAM(Kirillov等,2023)和少量用户交互得到二维掩码,约束编辑过程。LENeRF(Hyung等,2023)提出直接在三平面(Chan等,2022)上进行文本指导的三维局部编辑。利用CLIP损失确定文本与图像的二维交叉注意力图作为伪标签,训练注意力场模块得到三平面特征空间的三维掩码,利用此掩码融合源和目标三平面特征,最终在CLIP损失优化下,完成局部精确编辑。

基于显式约束的文本指导三维编辑方法利用显式约束明确指定编辑区域,增加编辑过程的可控性,并与优化损失协作,在指定区域编辑和优化三维表示,可以实现局部可控编辑。但是,此类方法受限于文本与渲染图像语义对齐准确程度,不准确的语义对齐会造成过度编辑或者欠编辑。此外,利用文本与图像语义对齐约束三维编辑,会面临编辑约束多视角不一致等问题,ViCA-NeRF(Dong和Wang,2024)、LENeRF(Hyung等,2023)等方法在一定程度上缓解了此类问题,但无法彻底解决上述问题。

3 基于神经辐射场和三维高斯泼溅的文本指导三维编辑能力

如表2所示,本文从编辑类型(几何、外观、风格化)、编辑范围(物体、场景)和编辑的鲁棒性(全局或局部可控)3个角度讨论现有文本指导的三维编辑方法的编辑能力。其中,几何编辑是指对三维模型的几何形状进行修改,使其符合目标文本提示的描述。外观编辑是指在文本的指导下,对三维模型的

表2 基于神经辐射场和三维高斯泼溅的文本指导三维编辑能力一览表

Table 2 A summary of text-guided 3D editing capabilities based on neural radiance fields and 3D Gaussian splatting

方法	编辑类型	编辑范围	编辑可控性
CLIP-NeRF(Wang等,2022)	几何/外观	物体	全局可控
NeRF-Art(Wang等,2024)	几何/外观/风格化	场景	全局可控
InstructN2N(Haque等,2023)	几何/外观/风格化	场景	全局可控
DreamBooth3D(Raj等,2023)	几何/外观	场景	全局可控
EfficientN2N(Song等,2023)	几何/外观/风格化	场景	全局可控
DN2N(Fang等,2023)	几何/外观/风格化	场景	全局可控
FreeEditor(Karim等,2024)	几何/外观/风格化	场景	全局可控
Instruct3D-to-3D(Kamata等,2023)	几何/外观	场景	全局可控
GSEditor(Palandra等,2024)	几何/外观	场景	全局可控
GaussCtrl(Wu等,2024)	几何/外观/风格化	场景	全局可控
VCEditor(Wang等,2025)	几何/外观/风格化	场景	全局可控
Shap-Editor(Chen等,2024a)	几何/外观	物体	全/局部可控
RMNE(Mirzaei等,2023)	几何/外观/风格化	场景	局部可控
ViCA-NeRF(Dong和Wang,2024)	几何/外观/风格化	场景	全/局部可控
customizeNeRF(He等,2024)	几何/外观	场景	局部可控
LatentEditor(Khalid等,2024)	几何/外观/风格化	场景	局部可控
Vox-E(Sella等,2023)	几何/外观	物体	全/局部可控
ED-NeRF(Park等,2024)	几何/外观	场景	局部可控
RepaintNeRF(Zhou等,2023)	几何/外观	场景	局部可控
LENeRF(Hyung等,2023)	几何/外观	场景	局部可控
DreamEditor(Zhuang等,2023)	几何/外观/风格化	场景	局部可控
GaussianEditor(Chen等,2024c)	几何/外观	场景	局部可控

纹理、颜色、材质等进行编辑。风格化是指对三维模型的风格进行变换，如将已有三维模型转换成梵高风格等。此外，为了进一步说明现有方法的编辑能力，给出了物体编辑、场景编辑和风格化编辑 3 个编辑任务上的对比实验，并针对实验结果进行了详细讨论分析。由于文本指导三维编辑任务的开放性，大多数研究方法通过可视化不同编辑实例上的编辑结果来展示其编辑能力。针对定量对比，使用用户研究来评估整体的编辑质量，同时使用基于 CLIP 的文本—图像相似度来评估编辑结果与编辑文本的对齐程度，并讨论方法的编辑效率。因此，在不同的三维编辑任务中，本文分别给出对比实验所使用的编辑用例和评价指标，并展示在编辑用例上的可视化或定量对比实验结果。对比实验可以在一块 NVIDIA Tesla A100 GPU 上完成。

3.1 文本指导的三维物体编辑

为了评估文本引导的物体编辑的性能，本文在 Shap-Editor（Chen 等，2024a）的编辑实例上展示了 3 种代表方法：InstructN2N（Haque 等，2023）、Vox-E（Sella 等，2023）和 Shap-Editor（Chen 等，2024a）的编辑结果。编辑示例包含两个：全局编辑的“让它看起来像金子做的”或者局部编辑“加上一顶圣诞老人的帽子”，分别展示了对比方法在全局外观编辑和局部几何和外观联合编辑任务上的编辑能力。如图 8 所示。

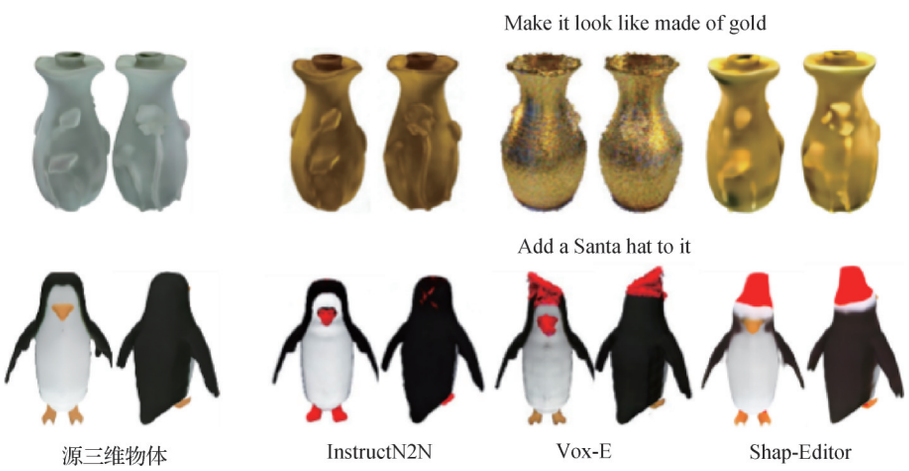

图 8　文本指导三维物体编辑的可视化结果（图片来自于 Shap-Editor（Chen 等，2024a））

Fig. 8　Visual results of text-guided 3D object editing（Image from Shap-Editor（Chen et al. ，2024a））

图 8 展示了 InstructN2N（Haque 等，2023）、Vox-E（Sella 等，2023）和 Shap-Editor（Chen 等，2024a）的可视化编辑结果。Vox-E（Sella 等，2023）和 Shap-Editor（Chen 等，2024a）都侧重编辑单个物体，可以实现对单个物体的纹理替换等全局的外观编辑，或者插入新物体的局部几何与外观联合编辑。针对全局编辑的目标文本“让它看起来像金子做的”（第 1 行）或者局部编辑的目标文本“加上一顶圣诞老人的帽子”（第 2 行），在保持非编辑区域不变以及完成指定编辑时，Shap-Editor 的编辑能力要优于 Vox-E。将花瓶变成金子做的同时，Shap-Editor 可以保留花瓶的几何细节；给企鹅带上圣诞帽的同时可以保持企鹅本身不变（企鹅嘴巴等细节保持不变）。这是因为 Vox-E（Sella 等，2023）利用直接体素网格优化（DVGO）来表示三维场景，几何和纹理的耦合使得 Vox-E 的编辑结果受到噪声影响。Shap-Editor（Chen 等，2024a）利用文本指导三维生成方法 Shap-E（Jun 和 Nichol，2023）在潜在空间中完成三维编辑。而且，为了实现局部精确编辑，会在编辑过程中利用重建损失约束非编辑区域的颜色和深度尽量不变。上述做法都进一步提高了 Shap-Editor 的编辑能力。但是 Shap-Editor 编辑后模型外观质量受限于 Shap-E 的生成能力，模型外观不够真实。比如添加的圣诞帽还只是看起来比较像，不真实。此外，可以认为单个物体是比较简单的场景，因为对比了典型方法 InstructN2N（Haque 等，2023），InstructN2N 利用文本指导的图像编辑模型隐式约束三维编辑，通过交替进行目标三维模型的渲染图像编辑和目标三维模型更新完成三维编辑。可以看到，针对全局编辑“让它看起来像金子做的”（图 8 第 1 行），InstructN2N 可以完成符合目标文本三维编辑，只是编辑结果质量不高；而针对局部编辑 InstructN2N 会存在编辑失败的情况，如针

对局部编辑“加上一顶圣诞老人的帽子”(图8第2行),InstructN2N所依赖的文本指导图像编辑模型无法完成多视角一致的图像编辑,从而导致三维编辑失败。

为了评估定量实验结果,设置了20个编辑实例,其中60%用于全局编辑,40%用于局部编辑,编辑的源三维模型来自于Shap-E(Jun和Nichol,2023)生成的对象或OmniObject3D数据集。表3给出了InstructN2N(Haque等,2023)、Vox-E(Sella等,2023)和Shap-Editor(Chen等,2024a)的定量对比结果,其中$CLIP_{sim}$和$CLIP_{dir}$分别表示CLIP文本—图像相似度和CLIP方向文本—图像相似度。实验结果来自于Shap-Editor(Chen等,2024a)。从表3可以看到,在$CLIP_{sim}$和$CLIP_{dir}$两个指标上,无论是局部还是全局的物体编辑,Shap-Editor都取得了更好的结果。此外,InstructN2N和Vox-E需要针对每个场景进行单独优化,且需要迭代优化,完成一次编辑的时间分别需要约45 min和53 min,不适合现实应用;而Shap-Editor借助Shap-E,在推理过程中,无需重新训练,即可在大约1 s内完成三维编辑。

表3 文本指导三维物体编辑的定量对比结果

Table 3 Quantitative comparison results of text-guided 3D object editing

方法	局部编辑		全局编辑		编辑时间
	$CLIP_{sim}$	$CLIP_{dir}$	$CLIP_{sim}$	$CLIP_{dir}$	
InstructN2N(Haque等,2023)	0.253	0.051	0.239	0.057	约45 min
Vox-E(Sella等,2023)	0.277	0.075	0.271	0.066	约53 min
Shap-Editor(Chen等,2024a)	0.292	0.097	0.272	0.072	约1 s

3.2 文本指导的三维场景编辑

为了评估文本引导的三维场景编辑能力,本文在InstructN2N EfficientN2N(Song等,2023)提供的编辑实例上,对比了InstructN2N(Haque等,2023)、RMNE(Mirzaei等,2023)、GaussianEditor(Chen等,2024c)和EfficientN2N(Song等,2023)4种典型方法。为了充分讨论不同方法的场景编辑能力以及泛化性,本文编辑实例包括室内和室外两种场景,并兼顾动物和人物两类常见编辑对象。图9展示了可视化编辑结果。给定目标编辑文本“把熊变成北极熊”或者“把他变成小丑”,InstructN2N(Haque等,2023)、RMNE(Mirzaei等,2023)和GaussianEditor(Chen等,2024c)3种方法都能取得不错的结果,但仍然存在编辑不完整和模糊编辑的问题,如在“把熊变成北极熊”的编辑实例中,熊的面部和脚部等区域存在不完整编辑;在“把他变成小丑”的编辑实例中,除了面部变为小丑以外,衣服等非编辑区域存在过度编辑。这是由于上述方法依赖于提升图像编辑完成三维编辑,编辑准确性受图像编辑准确性的限制。另外,RMNE(Mirzaei等,2023)和GaussianEditor(Chen等,2024c)使用二维相关图或分割图来实现局部可控的精确编辑,但是自动从目标文本提示与图像之间的语义关系中获得的二维相关图或分割图,依然存在编辑边界不准确问题。

EfficientN2N(Song等,2023)利用多视角联合编辑提高多视角一致性以及编辑效率,编辑后模型外观质量有所下降(如在图9“把熊变成北极熊”的编辑实例中,编辑后北极熊的外观皮毛质感缺失),但是,相比InstructN2N和RMNE需要30 min以上的编辑时间,EfficientN2N训练速度加快,约2 min可以完成编辑。GaussianEditor(Chen等,2024c)采用了训练和渲染速度更快的3D GS作为三维表示,可以在5~10 min完成编辑。此外,现有的场景编辑规模局限于以单一物体为中心的场景,难以直接拓展到多物体的复杂场景。

3.3 文本指导的三维风格化编辑

除了几何和外观等内容编辑,InstructN2N(Haque等,2023)和DreamEditor(Zhuang等,2023)等方法可以做到风格化编辑。为了评估文本引导的三维风格编辑方法的能力,本文对比了CLIP-NeRF(Wang等,2022)、NeRF-Art(Wang等,2024)、InstructN2N(Haque等,2023)、DreamEditor(Zhuang等,2023)、ViCA-NeRF(Dong和Wang,2024)和Free-Editor(Karim等,2024)6种典型方法,图10中展示了6种方法的在“梵高风格”和“野兽派风格”下的三维风格化编辑可视结果。

图 9　文本指导三维场景编辑的可视化结果

Fig. 9　Visual results of text-guided 3D scene editing

给定目标文本“梵高”或“野兽派”风格，CLIP-NeRF（Wang 等，2022）和 NeRF-Art（Wang 等，2024）利用基于 CLIP 的损失，使全局结构和局部编辑细节都遵循目标文本的语义，但是两个方法都缺乏编辑约束，导致风格化后外观质量下降。InstructN2N（Haque 等，2023）、ViCA-NeRF（Dong 和 Wang，2024）和 FreeEditor（Karim 等，2024）均采用提升图像编辑到三维的方式完成风格编辑，因此，风格编辑结果相似。此外，InstructN2N（Haque 等，2023）和 ViCA-NeRF（Dong 和 Wang，2024）等方法还可实现天气或季节转换等其他风格化编辑。不同于上述方法，DreamEditor（Zhuang 等，2023）遵循文本提示直接修改三维模型，也可实现风格化编辑。但是，上述方法仍然存在改变非编辑区域的过度编辑问题，如背景、衣服等的几何发生了不应有的改变。

为了评估不同方法的风格化编辑能力，本文进行了一项用户研究，要求参与者对 InstructN2N（Haque 等，2023）、DreamEditor（Zhuang 等，2023）、ViCA-NeRF（Dong 和 Wang，2024）和 FreeEditor（Karim 等，2024）4 种典型方法的编辑结果进行投票。投票依据为：编辑结果总体质量（V_{qua}）和与给定文本描述的对齐程度（V_{ali}）。用户研究涉及 51 名参与者，收到 100 个回复，表 4 展示了用户研究的对比结果。由于 ViCA-NeRF（Dong 和 Wang，2024）可以产生更生动的色彩和细节，更符合梵高风格，因此 ViCA-NeRF 在编辑结果总体质量和与给定文本描述

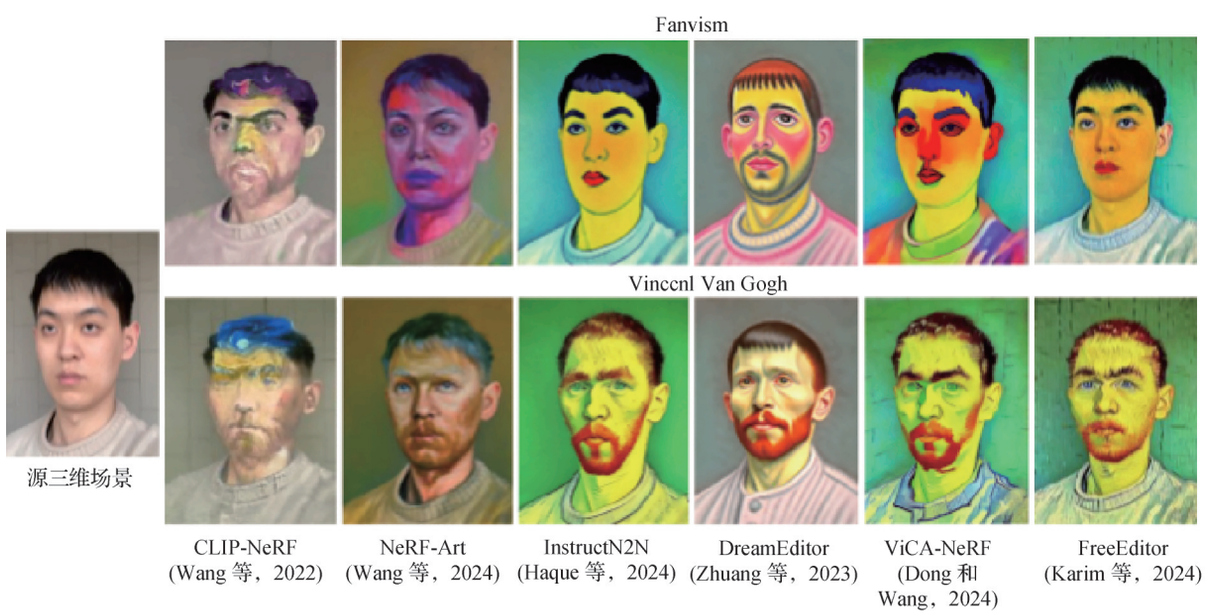

图 10　文本指导三维场景风格化编辑的可视化结果

Fig. 10　Visual results of text-guided 3D scene stylized editing

的对齐程度上获得了最高的投票数。

为了进一步讨论方法的可用性，本文对比了上述方法的编辑效率，对比结果如表5所示。相比于NeRF-Art长达数小时的编辑时间，InstructN2N、DreamEditor、ViCA-NeRF和FreeEditor可以将编辑时间缩短到数十分钟，特别是FreeEditor允许用户在推理时可直接编辑三维场景，而无需进一步重新训练模型，不仅减少了总编辑时间，还获得了更好的空间效率，而其他方法需要为每个不同的场景重新训练模型，从而增加了空间复杂度O(n)，n表示不同场景的数量。

表4　文本指导风格化编辑的用户研究对比结果

Table 4　Comparative results of user study on text-guided stylized editing

结果	方法			
	InstructN2N (Haque等，2023)	DreamEditor (Zhuang等，2023)	ViCA-NeRF (Dong和Wang，2024)	FreeEditor (Karim等，2024)
V_{ali}/%	26.7	10.8	51.7	10.8
V_{qua}/%	19.4	9.8	51.2	19.6

表5　文本指导风格化编辑效率对比结果

Table 5　Comparative results of editing efficiency of text-guided stylized editing

方法	编辑时间/min	时间复杂度	空间复杂度
NeRF-Art(Wang等，2024)	约780	O(n)	O(n)
InstructN2N(Haque等，2023)	约45	O(n)	O(n)
DreamEditor(Zhuang等，2023)	约70	O(n)	O(n)
ViCA-NeRF(Dong和Wang，2024)	约15	O(n)	O(n)
FreeEditor(Karim等，2024)	约3	O(n)	O(1)

4　挑战与未来研究方向

尽管现有文本指导的三维编辑技术取得了一定的发展与进步，但该领域仍面临着一些值得深入研究的挑战。

1)三维编辑的可控性。三维编辑的可控性主要表现在两个方面：编辑位置准确判定、外观和几何独立编辑。(1)现有的文本指导三维编辑方法根据文本与图像的对应关系自动确定编辑位置，高度依赖于文本指导的图像生成与编辑模型的语义对齐能力，会导致编辑区域边界定位模糊、过度编辑和多视角不一致等问题。引入用户交互是一种可行的策略。例如，鼓励用户用三维边界框(Cheng等，2024)、二维草图(Mikaeili等，2023)等指定编辑区域，增加编辑可控性。(2)外观和几何独立准确编辑是三维可控编辑的另一难题。尽管神经辐射场和三维高斯泼溅在渲染质量方面已经达到逼真的效果，但这两种表示形式无法准确解耦几何和外观，且在几何编辑方面受到隐式表示或离散几何表示的影响，几何编辑可控性受限。利用更先进的渲染技术提取几何、纹理等以促进独立编辑，并结合网格等显式表示，增加神经辐射场和三维高斯泼溅的几何表示能力是解决此问题的一种有前景的方向。此外，针对外观编辑可控性问题，可对现有三维编辑模型进行调整，使其能够接受更多的输入条件，以实现更可控、更确定性的三维编辑。例如，TIP-Editor(Zhuang等，2024)可以接收文本以及参考图像作为输入条件，利用参考图像消除文本提示的模糊性，实现符合文本描述通用特征和参考图像独有特征的三维编辑。

2)可信的评价标准。对编辑质量的评价，除了比较可视化结果、进行用户研究等主观评价方式外，现有方法常用文本指导的三维生成中的定量评价标准。例如，基于CLIP相似度只能评估三维资产与目标文本的对齐程度，缺乏对编辑模型的多视图一致性等其他方面的评估，且不能反映非编辑区域的保持程度，然而，这在三维编辑中是至关重要的。未来研究需要有更可靠的定性标准，从模型质量、对非编

辑区域的保持度以及与目标文本的匹配度等方面客观判断编辑质量。利用大语言模型来评估编辑结果，是一个有前景的研究方向。

3）大规模场景编辑。现有文本指导三维编辑可支持编辑物体和场景，但场景规模局限于以物体为中心的室内或室外场景，不适用于多街区等大规模场景。而且，编辑这些大规模场景，需要复杂的文本描述，这对复杂文本理解以及文本—三维场景语义对齐也提出了挑战。最后，神经辐射场和三维高斯泼溅等三维表示仍然存在针对大规模场景占用内存多、几何表示不一致等问题。针对上述问题，本文认为现有的大语言模型可以帮助理解复杂的文本提示，然后将复杂文本提示转换为简单的文本指令，输入到已有文本指导的三维编辑方法中。其次，探索更高效、轻量化的三维表示形式（Lin 等，2024）也是解决上述问题的可行办法。

4）大规模三维编辑数据集。收集大型配对的文本—三维编辑数据集具有挑战性，这是多方面原因导致的。首先，不同于从网络上可以轻松获取的图像数据，构建高质量的三维资产需要专业技能。其次，人工标注文本描述需要花费大量时间，且会存在主观认知偏差，导致三维资产对应的文本标注存在噪声。此外，三维编辑存在插入新物体、替换已有物体等多种编辑需求，构建符合不同任务需求的、大型成对的源三维资产—文本—编辑后三维资产的数据集是困难的。由于上述问题，现有文本指导三维编辑方法大多依赖于文本指导的图像编辑和每个物体或场景单独优化，无法构建适用于不同场景或编辑任务的大模型。将文本指导的三维编辑任务转换为三维生成任务，利用文本指导的三维生成大模型（Zhang 等，2024）是解决此类问题的一个可行方法。例如，针对插入新物体的三维编辑任务，可以使用文本指导的三维生成大模型先生成指定物体，然后插入到场景中指定位置。针对替换已有物体的三维编辑任务，可以先将已有物体删除，然后利用三维生成大模型生成要替换的物体，并将替换后物体插入指定位置。

5 结 语

随着文本引导图像编辑、三维表示等领域的进展，基于神经辐射场和三维高斯泼溅的文本引导三维编辑研究已取得丰硕成果。本文从方法本质和编辑能力两个维度，对最新研究进展进行了全面的调查与总结分析。本文依据编辑约束构建了一个基本的分类框架，并对各类方法本质进行了深入的分析。进一步地，本文从编辑类型、范围和可控性等多个角度，探讨了这些方法对编辑能力的贡献，并进行了详细对比讨论。此外，本文还探讨了该领域面临的一些挑战，并提出了若干具有前景的研究方向。期望本综述能够为感兴趣的读者提供一个系统的概览，并激发研究者们在后续工作中的创新思维。

参考文献（References）

Brooks T, Holynski A and Efros A A. 2023. InstructPix2Pix: learning to follow image editing instructions//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 18392-18402 [DOI: 10.1109/CVPR52729.2023.01764]

Chan E R, Lin C Z, Chan M A, Nagano K, Pan B X, De Mello S, Gallo O, Guibas L, Tremblay J, Khamis S, Karras T and Wetzstein G. 2022. Efficient geometry-aware 3D generative adversarial networks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 16102-16112 [DOI: 10.1109/CVPR52688.2022.01565]

Chen A P, Xu Z X, Geiger A, Yu J Y and Su H. 2022. TensoRF: tensorial radiance fields//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 333-350 [DOI: 10.1007/978-3-031-19824-3_20]

Chen M H, Xie J Y, Laina I and Vedaldi A. 2024a. Shap-editor: instruction-guided latent 3D editing in seconds//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 26446-26456 [DOI: 10.1109/CVPR52733.2024.02498]

Chen Y S, Shao G C, Shum K C, Hua B S and Yeung S K. 2024b. Advances in 3D neural stylization: a survey [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2311.18328.pdf

Chen Y W, Chen Z L, Zhang C, Wang F, Yang X F, Wang Y K, Cai Z, Yang L, Liu H P and Lin G S. 2024c. GaussianEditor: swift and controllable 3D editing with Gaussian splatting//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 21476-21485 [DOI: 10.1109/CVPR52733.2024.02029]

Cheng X H, Yang T Y, Wang J, Li Y, Zhang L, Zhang J and Yuan L. 2024. Progressive3D: progressively local editing for Text-to-3D content creation with complex semantic prompts [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2310.11784.pdf

Choudhary T, Dewangan V, Chandhok S, Priyadarshan S, Jain A,

Singh A K, Srivastava S, Jatavallabhula K M and Krishna K M. 2024. Talk2BEV: language-enhanced bird's-eye view maps for autonomous driving//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE: 16345-16352 [DOI: 10.1109/ICRA57147.2024.10611485]

Dong J H and Wang Y X. 2024. ViCA-NeRF: view-consistency-aware 3D editing of neural radiance fields [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2402.00864.pdf

Fang S K, Wang Y F, Yang Y, Tsai Y H, Ding W R, Zhou S C and Yang M H. 2023. Editing 3D scenes via text prompts without retraining [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2309.04917.2023.12.05.pdf

Foo L G, Rahmani H and Liu J. 2025. AI-Generated Content (AIGC) for various data modalities: a survey [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2308.14177.pdf

Fridovich-Keil S, Yu A, Tancik M, Chen Q H, Recht B and Kanazawa A. 2022. Plenoxels: radiance fields without neural networks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 5491-5500 [DOI: 10.1109/CVPR52688.2022.00542]

Haque A, Tancik M, Efros A A, Holynski A and Kanazawa A. 2023. Instruct-NeRF2NeRF: editing 3D scenes with instructions//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 19683-19693 [DOI: 10.1109/ICCV51070.2023.01808]

He R Z, Huang S F, Nie X C, Hui T R, Liu L Q, Dai J, Han J Z, Li G B and Liu S. 2024. Customize your NeRF: adaptive source driven 3D scene editing via local-global iterative training//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 6966-6975 [DOI: 10.1109/CVPR52733.2024.00665]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Advances in neural information processing systems. Vancouver, Canada, 33, 6840-6851 [DOI: 10.5555/3495724.3496298]

Hyung J, Hwang S, Kim D, Lee H and Choo J. 2023. Local 3D editing via 3D distillation of CLIP knowledge//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 12674-12684 [DOI: 10.1109/CVPR52729.2023.01219]

Jun H and Nichol A. 2023. Shap-E: generating conditional 3D implicit functions [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2305.02463.pdf

Kamata H, Sakuma Y, Hayakawa A, Ishii M and Narihira T. 2023. Instruct 3D-to-3D: text instruction guided 3D-to-3D conversion [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2303.15780.pdf

Karim N, Iqbal H, Khalid U, Hua J and Chen C. 2024. Free-editor: zero-shot text-driven 3D scene editing [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2312.13663.pdf

Kerbl, Bernhard and Kopanas, Georgios and Leimkuehler, Thomas and Drettakis, George. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4): 139: 1-139:14 [DOI: 10.1145/3592433]

Khalid U, Iqbal H, Karim N, Hua J and Chen C. 2024. LatentEditor: text driven local editing of 3D scenes [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2312.09313.pdf

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, Xiao T T, Whitehead S, Berg A C, Lo W Y, Dollár P and Girshick R. 2023. Segment anything//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 3992-4003 [DOI: 10.1109/ICCV51070.2023.00371]

Li X Y, Zhang Q, Kang D, Cheng W H, Gao Y M, Zhang J B, Liang Z H, Liao J, Cao Y P and Shan Y. 2024. Advances in 3D generation: a survey [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2401.17807.pdf

Lin J Q, Li Z H, Tang X, Liu J Z, Liu S Y, Liu J Y, Lu Y D, Wu X F, Xu S C, Yan Y L and Yang W M. 2024. VastGaussian: vast 3D Gaussians for large scene reconstruction [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2402.17427.pdf

Liu A A, Su Y T, Wang L J, Li B, Qian Z X, Zhang W M, Zhou L N, Zhang X P, Zhang Y D, Huang J W and Yu N H. 2024. Review on the progress of the AIGC visual content generation and traceability. Journal of Image and Graphics, 29(6): 1535-1554 (刘安安, 苏育挺, 王岚君, 李斌, 钱振兴, 张卫明, 周琳娜, 张新鹏, 张勇东, 黄继武, 俞能海. 2024. AIGC视觉内容生成与溯源研究进展. 中国图象图形学报, 29(6): 1535-1554) [DOI: 10.11834/jig.240003]

Mikaeili A, Perel O, Safaee M, Cohen-Or D and Mahdavi-Amiri A. 2023. SKED: sketch-guided text-based 3D editing//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 14561-14573 [DOI: 10.1109/ICCV51070.2023.01343]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2022. NeRF: representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99-106 [DOI: 10.1145/3503250]

Mirzaei A, Aumentado-Armstrong T, Brubaker M A, Kelly J, Levinshtein A, Derpanis K G and Gilitschenski I. 2023. RMNE-watch your steps: local image and scene editing by text instructions [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2308.08947.pdf

Müller T, Evans A, Schied C and Keller A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics, 41(4): 102 [DOI: 10.1145/3528223.3530127]

Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I and Chen M. 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models [EB/OL]. [2024-10-11]. http://arxiv.org/odf/2112.10741.pdf

Palandra F, Sanchietti A, Baieri D and Rodolà E. 2024. GSEdit: efficient text-guided editing of 3D objects via Gaussian splatting [EB/

OL]. [2024-10-11]. http://arxiv.org/pdf/2403.05154.pdf

Park J, Kwon G and Ye J C. 2024. ED-NeRF: efficient text-guided editing of 3D scene with latent space NeRF [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2310.02712. pdf

Patashnik O, Wu Z Z, Shechtman E, Cohen-Or D and Lischinski D. 2021. StyleCLIP: text-driven manipulation of StyleGAN imagery//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 2065-2074 [DOI: 10.1109/ICCV48922.2021.00209]

Poole B, Jain A, Barron J T and Mildenhall B. 2022. DreamFusion: text-to-3D using 2D diffusion [EB/OL].[2024-10-11]. http://arxiv.org/pdf/2209.14988.pdf

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2103.00020.pdf

Raj A, Kaza S, Poole B, Niemeyer M, Ruiz N, Mildenhall B, Zada S, Aberman K, Rubinstein M, Barron J, Li Y Z and Jampani V. 2023. DreamBooth3D: subject-driven text-to-3D generation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 2349-2359 [DOI: 10.1109/ICCV51070.2023.00223]

Reed S, Akata Z, Yan X C, Logeswaran L, Schiele B and Lee H. 2016. Generative adversarial text to image synthesis [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/1605.05396.pdf

Ren T H, Liu S L, Zeng A L, Lin J, Li K C, Cao H, Chen J Y, Huang X Y, Chen Y K, Yan F, Zeng Z Y, Zhang H, Li F, Yang J, Li H Y, Jiang Q and Zhang L. 2024. Grounded SAM: assembling open-world models for diverse visual tasks [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2401.14159.pdf

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Ruiz N, Li Y Z, Jampani V, Pritch Y, Rubinstein M and Aberman K. 2023. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 22500-22510 [DOI: 10.1109/CVPR52729.2023.02155]

Saharia C, Chan W, Saxena S, Li L L, Whang J, Denton E, Ghasemipour S K S, Ayan B K, Mahdavi S S, Lopes R G, Salimans T, Ho J, Fleet D J and Norouzi M. 2022. Photorealistic text-to-image diffusion models with deep language understanding [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2205.11487.pdf

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//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 18582-18592 [DOI: 10.1109/CVPR52688.2022.01805]

Sella E, Fiebelman G, Hedman P and Averbuch-Elor H. 2023. Vox-E: text-guided voxel editing of 3D objects//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 430-440 [DOI: 10.1109/ICCV51070.2023.00046]

Song L C, Cao L L, Gu J T, Jiang Y F, Yuan J S and Tang H. 2023. Efficient-NeRF2NeRF: streamlining text-driven 3D editing with multiview correspondence-enhanced diffusion models [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2312.08563.pdf

Sun C, Sun M and Chen H T. 2022. Direct voxel grid optimization: super-fast convergence for radiance fields reconstruction//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 5449-5459 [DOI: 10.1109/CVPR52688.2022.00538]

Tang J X, Ren J W, Zhou H, Liu Z W and Zeng G. 2024. DreamGaussian: generative Gaussian splatting for efficient 3D content creation [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2309.16653.pdf

Tseng K W, Huang J Y, Chen Y S, Chen C S and Hung Y P. 2022. Pseudo-3D scene modeling for virtual reality using stylized novel view synthesis//ACM SIGGRAPH 2022 Posters. Vancouver, Canada: ACM: #66 [DOI: 10.1145/3532719.3543232]

Wang C, Chai M L, He M M, Chen D D and Liao J. 2022. CLIP-NeRF: text-and-image driven manipulation of neural radiance fields//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 3825-3834 [DOI: 10.1109/CVPR52688.2022.00381]

Wang C, Jiang R X, Chai M L, He M M, Chen D D and Liao J. 2024. NeRF-Art: text-driven neural radiance fields stylization. IEEE Transactions on Visualization and Computer Graphics, 30(8): 4983-4996 [DOI: 10.1109/TVCG.2023.3283400]

Wang Y X, Yi X Y, Wu Z K, Zhao N, Chen L and Zhang H W. 2025. View-consistent 3D editing with Gaussian splatting [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2403.11868.pdf

Wu J, Bian J W, Li X H, Wang G R, Reid I, Torr P and Prisacariu V A. 2024. GaussCtrl: multi-view consistent text-driven 3D Gaussian splatting editing [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2403.08733z.pdf

Yu A, Ye V, Tancik M and Kanazawa A. 2021. pixelNeRF: neural radiance fields from one or few images [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2012.02190.pdf

Zhang B W, Cheng Y J, Yang J L, Wang C Y, Zhao F, Tang Y S, Chen D and Guo B N. 2024. GaussianCube: a structured and explicit radiance representation for 3D generative modeling [EB/OL]. [2024-10-11]. http://arxiv.org/pdf/2403.19655.pdf

Zhang H Y, Wang T B, Li M Z, Zhao Z, Pu S L and Wu F. 2022. Comprehensive review of visual-language-oriented multimodal pre-

training methods. Journal of Image and Graphics, 27(9): 2652-2682 (张浩宇, 王天保, 李孟择, 赵洲, 浦世亮, 吴飞. 2022. 视觉语言多模态预训练综述. 中国图象图形学报, 27(9): 2652-2682) [DOI: 10.11834/jig.220173]

Zhang L, Rao A Y and Agrawala M. 2023. Adding conditional control to text-to-image diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 3813-3824 [DOI: 10.1109/ICCV51070.2023.00355]

Zhou X C, He Y, Yu F R, Li J Q and Li Y. 2023. RePaint-NeRF: NeRF editting via semantic masks and diffusion models//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macau, China: International Joint Conferences on Artificial Intelligence Organization: 1813-1821 [DOI: 10.24963/ijcai.2023/201]

Zhuang J Y, Kang D, Cao Y P, Li G B, Lin L and Shan Y. 2024. TIP-Editor: an accurate 3D editor following both text-prompts and image-prompts. ACM Transactions on Graphics, 43(4): #121 [DOI: 10.1145/3658205]

Zhuang J Y, Wang C, Lin L, Liu L J and Li G B. 2023. DreamEditor: text-driven 3D scene editing with neural fields//Proceedings of SIGGRAPH Asia 2023 Conference Papers. Sydney, Australia: ACM: #26 [DOI: 10.1145/3610548.3618190]

作者简介

卢丽华,女,研究员,主要研究方向为三维计算机视觉。
E-mail:roryuna@126.com
张晓辉,男,研究员,主要研究方向为三维计算机视觉。
E-mail: xiaohuizhang06@163.com
魏辉,男,研究员,主要研究方向为三维计算机视觉。
E-mail:weihui@126.com
李茹杨,女,研究员,主要研究方向为三维计算机视觉、自动驾驶。E-mail: lryheadquarters@163.com
杜国光,男,研究员,主要研究方向为三维计算机视觉。
E-mail: guoguangdu@126.com
王斌强,男,研究员,主要研究方向为情感计算、三维计算机视觉。E-mail: binqiang2wang@qq.com