Pure Mathematics 理论数学, 2025, 15(4), 31-45
Published Online April 2025 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2025.154106

Hans汉斯

基于三维场景的语义感知风格迁移

焦　傲

中国科学技术大学人工智能与数据科学学院，安徽　合肥

收稿日期：2025年2月10日；录用日期：2025年3月12日；发布日期：2025年4月9日

摘　要

随着电影和游戏行业的快速发展，三维场景的创建与编辑方法不断优化，逐渐向更高效、更便捷的方向发展。相比传统的网格和点云表示方法，三维高斯提供了一种更灵活且高效的三维场景表示方式，能够在保证高质量渲染效果的同时，生成逼真的新视角图像。然而，现有的三维高斯模型在风格化方面仍存在局限性，难以满足创意设计和艺术表达的需求。因此，如何在保持三维结构信息的同时，实现高质量的风格迁移，成为一个值得深入研究的问题。针对这一问题，本文提出了一种基于三维高斯的语义风格迁移方法。首先，通过多视角图像训练三维高斯模型，并在这些图像上进行风格迁移，以确保三维模型的风格一致性和结构完整性。具体而言，我们利用LSeg模型对内容图像和风格图像进行语义分割，提取对应区域后，基于图像复杂度自适应确定聚类类别数量，在颜色空间采用K均值聚类进行分割，并以聚类区域面积筛选有效的结构信息。随后，通过语义匹配进行风格迁移，并结合WCT进行风格融合，最终使用VGG解码器生成风格化图像。实验结果表明，本文方法在风格质量、结构保持性和多视角一致性方面均优于现有方法，为三维艺术创作提供了更高质量的风格迁移效果和更强的可控性。

关键词

风格迁移，三维高斯，语义感知

Semantic-Aware Style Transfer Based on 3D Scenes

Ao Jiao

School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei Anhui

Received: Feb. 10th, 2025; accepted: Mar. 12th, 2025; published: Apr. 9th, 2025

Abstract

With the rapid development of the film and gaming industries, methods for creating and editing 3D scenes have been continuously optimized, evolving toward greater efficiency and convenience.

文章引用: 焦傲. 基于三维场景的语义感知风格迁移[J]. 理论数学, 2025, 15(4): 31-45.
DOI: 10.12677/pm.2025.154106

Compared to traditional representations based on meshes and point clouds, 3D Gaussian Splatting provides a more flexible and efficient way to represent 3D scenes, enabling high-quality novel view synthesis while maintaining superior rendering performance. However, existing 3D Gaussian models still have limitations in stylization, making it difficult to meet the demands of creative design and artistic expression. Therefore, achieving high-quality style transfer while preserving 3D structural information remains a challenging research problem. To address this issue, we propose a semantic style transfer method based on 3D Gaussians. First, a 3D Gaussian model is trained using multi-view images, and style transfer is performed on these images to ensure consistency and structural integrity in the final 3D model. Specifically, we utilize the LSeg model for semantic segmentation of content and style images. After extracting corresponding regions, we adaptively determine the number of clusters based on image complexity and apply K-means clustering in the color space to segment the images. The clustered regions are then filtered based on their area to retain essential structural information. Subsequently, style transfer is performed using semantic matching, and style fusion is achieved with the Whitening and Coloring Transform (WCT). Finally, a VGG-based decoder generates the stylized images. Experimental results demonstrate that our method outperforms existing approaches in terms of style quality, structural preservation, and multi-view consistency, providing better controllability and higher-quality style transfer for 3D artistic content creation.

Keywords

Style Transfer, 3D Gaussian Splatting, Semantic Perception

1. 引言

随着电影和游戏行业的不断发展，越来越多的研究致力于探索更加高效、便捷的方式来创建和编辑三维场景。相比基于网格(Mesh)或点云(Point Cloud)的传统三维表示方法，三维高斯(3D Gaussian Splatting) [1]提供了一种更加灵活且高效的方式来表示三维场景，能够在保持高质量渲染效果的同时，生成高质量的新视角图片，适应不同的优化需求。然而，尽管现有三维高斯模型能够精确地重建场景，其风格化能力仍然有限，难以满足创意设计和艺术表达的需求。因此，如何在保持三维结构信息的同时，实现高质量的风格迁移，成为一个值得探索的问题。

针对图像风格迁移，已有大量研究在 2D 领域取得了显著进展。例如，VGG 特征提取网络常被用于计算内容与风格的特征差异，自适应实例归一化(AdaIN) [2]能够高效地在内容图像中融合风格信息，而白化彩色变换(WCT) [3]提供了一种基于特征变换的方法来进行风格转换。此外，多模态风格迁移(MST) [4]等技术也在提升风格精细度和多模式风格控制方面发挥了重要作用。然而，将这些方法直接应用于三维场景仍然面临诸多挑战，特别是在如何保证多视角一致性以及如何提取和匹配语义信息方面，仍然存在较大改进空间。

为了解决上述问题，我们提出了一种基于三维高斯的语义风格迁移方法。我们的方法首先通过多视角图像训练出三维高斯模型，然后在这些多视角图像上进行风格迁移，以保证最终三维模型的风格一致性和结构保持性。具体而言，我们利用 LSeg 模型[5]对内容图像和风格图像进行语义分割，提取对应区域后，将图像从 RGB 空间转换到 LAB 颜色空间，并基于图像复杂度(如颜色数量、亮度等)自适应地确定聚

类类别数目，采用 K 均值聚类进行分割，并以聚类区域的面积作为标准筛选有效的结构信息。随后，我们根据语义匹配结果进行风格迁移，并结合 WCT 进行风格融合，最后通过 VGG 解码器生成最终的风格化图像。由于我们的方法基于颜色聚类进行风格迁移，可以有效保持图像结构，并确保多视角图像之间的风格一致性。

我们的方法主要贡献如下：

1) 提出了一种基于多视角图像的三维语义风格迁移框架，能够在保证结构保持性的同时，实现高质量风格迁移。

2) 结合 LSeg 语义分割和自适应 K 均值聚类，提取并匹配内容与风格图像的语义对应区域，提升风格迁移的准确性。

3) 采用颜色聚类结合 WCT 方法进行风格融合，确保多视角风格迁移的一致性，并提升最终渲染质量。

通过实验验证，我们的方法在风格化质量、结构保持性和多视角一致性方面均优于现有方法，为三维艺术创作提供了更强的可控性和更高质量的风格迁移效果。

2. 相关工作

2.1. 二维图像风格迁移

二维风格迁移技术已经取得了显著进展，早期的风格迁移方法包括非参数化算法[6]和纹理合成[7]，但自基于卷积神经网络(CNN)的风格迁移方法以来，深度学习技术极大提升了风格迁移的效果。通过使用 VGG-19 网络[8]提取的特征，并计算 Gram 矩阵来实现风格约束，优化噪声图像并生成最终的风格化图像。随后的一些研究工作[9]-[12]探索了替代的风格损失公式，以增强语义一致性并捕捉笔触等高频风格细节。前馈式风格迁移方法[13]-[15]通过训练神经网络，捕捉风格图像的风格信息，并通过单次前向传播将其转移到输入图像中，从而确保了更快的风格化处理。

近期在风格损失方面的改进[16] [17]包括将全局 Gram 矩阵替换为最近邻特征矩阵，以提高纹理的保留。一些方法采用了图像生成的补丁匹配技术，如 Fast PatchMatch [18]和 PatchMatch [19]，但它们通常仅限于特定的视角。

这些方法为二维风格迁移提供了更为灵活和精细的调控手段，使得风格迁移不仅仅限于简单的视觉效果转换，而是能够精准地控制风格的传递与内容的保持。在我们的研究中，我们结合了这些先进的技术，通过引入语义感知方法、特征匹配和自适应算法，进一步提高了风格迁移的质量与可控性，尤其是在多视角图像的风格迁移中，确保了风格一致性和结构的精确保持。

2.2. 三维场景表示

近年来，三维场景表示在计算机视觉领域取得了显著进展。隐式神经表示，尤其是 NeRF (Neural Radiance Fields) [20]，通过使用位置编码神经网络成功地对三维场景进行了表示，特别在新视角合成中表现出色。然而，NeRF 的局限性在于渲染速度慢和训练过程中较高的内存消耗。这些问题源于其需要在渲染过程中反复查询神经网络来计算场景的每个细节，导致大规模场景渲染的效率较低。

为了解决这些问题，许多方法尝试结合显式表示和隐式表示，以在效率和效果之间取得平衡。例如，Triplane [21]、TensoRF [22]和 K-Plane [23]等方法通过张量分解技术，将显式表示引入到隐式辐射场中，显著提升了表示效率。与这些方法不同，三维高斯模型(3D Gaussian Splatting) [1]采用了点云和三位高斯分布的显式表示方法，通过结合 α 混合和高效的光栅化技术，能够在渲染过程中显著减少计算开销，并且能更好地捕捉细节。与传统的体积渲染方法相比，三维高斯不仅在渲染大规模场景时保持了高效率，

还能够在保留细节的同时提供更好的视觉效果。

本研究使用三维高斯作为骨干模型，采用显式的基于点的三维表示，利用各向异性三维高斯分布来表示三维场景。通过这种方式，我们不仅解决了传统隐式表示的计算瓶颈，还能有效提升渲染效率，特别是在复杂场景下，能够显著提高计算性能和视觉效果。

2.3. 三维风格迁移

三维场景风格迁移的目标是将风格应用到场景中，同时保持风格的保真度和多视角的一致性。随着对 3D 内容需求的增加，神经风格迁移技术已扩展到多种三维表现形式。在网格风格化中，通常采用差异渲染技术，将风格迁移目标从渲染图像传播到三维网格，从而实现几何或纹理的转移[24]-[26]。另一些研究则使用点云作为三维场景的代理，在风格化新视角时确保三维一致性。例如，[27]方法使用特征化的三维点云与风格图像结合，经过 CNN 渲染器生成风格化渲染图像。然而，显式方法的性能常受到几何重建质量的限制，尤其在复杂的现实世界场景中，往往会出现明显的伪影。

包括 NeRF [20]和三维高斯在内的新模型能够较好地表达复杂场景，逐渐成为研究重点。许多基于 NeRF 的风格化网络在训练时引入了图像风格迁移损失[28] [29]，或者通过相互学习的图像风格化网络来优化基于参考风格的颜色相关参数[30]。一些方法[31] [32]支持同时进行外观和几何的风格化，能够在新视角下保持风格的一致性，取得良好的效果。然而，这些方法通常需要时间较长的优化过程，且由于体积渲染中的随机采样代价较高，渲染速度较慢，无法实现实时渲染。本研究立足于三维高斯模型的显示表达之上，利用 LSeg 模型[5]和自适应匹配算法，能实现更加迅速且高质量的三维风格迁移任务。

2.4. 语义识别模型

自然语言驱动的图像识别模型在 CLIP [33]发布后受到了广泛关注。CLIP (Contrastive Language-Image Pretraining)通过对比学习将图像和文本映射到同一特征空间，从而实现了零样本分类预测，这一创新为图像与文本的联合理解提供了强大的基础。通过这种方式，CLIP 能够利用大量的图像和文本数据进行预训练，使得模型能够在没有特定任务标签的情况下，直接根据文本描述进行图像分类。这种方法极大地推动了多模态学习的发展，并为后续的相关研究奠定了基础。

LSeg [5]在 CLIP 的基础上进一步发展，采用了 CLIP 中的预训练文本编码器，并结合基于 Vision Transformer (ViT) [34]的图像编码器，成功实现了像素级的语义特征预测。这使得 LSeg 能够更精确地处理图像中的细粒度信息，尤其是在图像分割和语义理解任务中，表现出了优异的效果。此外，其他研究[19] [25] [35]也证实了 CLIP 特征的损失函数在语义信息控制上的潜力，展示了对语义内容的良好可控性，进一步验证了 CLIP 在多模态任务中的有效性。

在我们的工作中，我们结合了 LSeg 中提取的语义特征与 VGG 网络提取的视觉特征，综合在一起进行自适应的特征匹配，从而更好地保留图像内容的结构，实现风格的精细变化。这种跨网络的特征融合，不仅提高了风格迁移的精度，还增强了模型对复杂图像和多样风格的适应能力。

3. 背景知识

3.1. 三维高斯模型

三维高斯模型[1]将三维场景表示为一组三维高斯基元 $\mathbb{G}=\left\{\boldsymbol{g}_p=\left(\boldsymbol{\mu}_p,\boldsymbol{\Sigma}_p,\sigma_p,\boldsymbol{c}_p\right)\right\}$，其中每个三维高斯基元 $\boldsymbol{g}_p$ 由以下参数描述：表示位置的中心点 $\boldsymbol{\mu}_p\in\mathbb{R}^3$，表示形状和大小的协方差矩阵 $\boldsymbol{\Sigma}_p\in\mathbb{R}^{3\times3}$，不透明度 $\sigma_p\in\mathbb{R}_+$ 和颜色 $\boldsymbol{c}_p\in\mathbb{R}^3$。给定相机内参 $\boldsymbol{K}$ 和外参(旋转矩阵 $\boldsymbol{R}$ 和平移向量 $\boldsymbol{t}$)，三维高斯基元的中心点 $\boldsymbol{\mu}_i$ 投影到图像平面上的像素坐标 $\boldsymbol{p}_i$ 由以下公式计算：

$$\boldsymbol{p}_i = \boldsymbol{K}\left(\boldsymbol{R}\boldsymbol{\mu}_i + \boldsymbol{t}\right) \tag{1}$$

同时，高斯的协方差矩阵 $\boldsymbol{\Sigma}_i$ 需要变换到相机坐标系，并投影到图像空间。令 $\boldsymbol{J}$ 为投影变换的雅可比矩阵，则 $\Sigma_i' = \boldsymbol{J}\Sigma_i\boldsymbol{J}^{\mathrm{T}}$ 将三维高斯在相机视角下变换为二维椭圆。

在图像平面上，每个二维椭圆通过光栅化确定其覆盖的像素区域，并计算其颜色贡献。使用 α 混合计算最终颜色，按照从前到后的顺序进行累积。令 α_i 等于高斯基元的协方差矩阵 $\boldsymbol{\Sigma}_i$ 乘以不透明度 σ_p，则计算像素颜色的公式为：

$$\boldsymbol{C} = \sum_{i\in N} \boldsymbol{c}_i\alpha_i \prod_{j=1}^{i-1}\left(1-\alpha_j\right) \tag{2}$$

三维高斯模型是一种高效的可微分渲染技术，能够实现实时渲染。相比传统的基于采样的方法(如 Nerf [20])，它在计算速度和内存占用方面大幅提升，更加高效。此外，与 NeRF 类似，三维高斯模型可以针对特定场景进行重建，并利用多张带有位姿信息的图像进行监督学习，从而生成更精确的渲染效果。

3.2. 多模态风格迁移模型

VGG 网络[8]是一种深度卷积神经网络，因其层次清晰、特征提取能力强，被广泛用于计算机视觉任务。其主要特点是采用连续的小尺寸 3×3 卷积核，并通过增加网络深度来提升特征表达能力。在风格迁移任务中，VGG 作为特征提取网络，其不同层的特征映射可表示为：

$$\boldsymbol{F}_l = \phi_l\left(I\right) \tag{3}$$

其中 $\phi_l(\cdot)$ 表示 VGG 在第 l 层的特征提取操作，$\boldsymbol{F}_l$ 为对应的特征映射。风格信息通常使用 Gram 矩阵，该矩阵捕捉了通道间的相关性，可用于衡量图像风格，具体表示为：

$$\boldsymbol{G}_l = \boldsymbol{F}_l\boldsymbol{F}_l^{\mathrm{T}} \tag{4}$$

然而，该方法无法区分风格图像的不同语义模式。在此基础上，多模态风格迁移模型(MST) [4]对风格图像的特征进行聚类，形成子风格集合 $\mathcal{S} = \{S_1, S_2, \cdots, S_K\}$，每个子风格代表一种风格模式。然后计算内容特征 f_i^c 与每个子风格 S_k 之间的相似性，通常使用马氏距离衡量：

$$D\left(f_i^c, S_k\right) = \left(f_i^c - \mu_k\right)^{\mathrm{T}} \Sigma_k^{-1}\left(f_i^c - \mu_k\right) \tag{5}$$

基于此，MST 通过图割在内容图像中为每个区域找到最优的子风格匹配，定义能量函数为：

$$E(X) = \sum_i D\left(f_i^c, S_{x_i}\right) + \lambda \sum_{(i,j)\in\mathcal{N}} \boldsymbol{I}\left(x_i \neq x_j\right) \tag{6}$$

完成匹配后，在风格合成阶段，MST 通过一个风格化网络 G_θ 将内容图像转换为目标风格：

$$\hat{F}^c = G_\theta\left(F^c, S_X\right) \tag{7}$$

MST 通过显式建模风格图像中的多种模式，使风格迁移更加灵活精准，然而它是在 VGG 特征空间里进行聚类和分割，无法保证内容图像的结构完整，K 均值聚类的个数也需要人为指定，不会依据图像自适应变化，此外，图割会造成一些风格特征的丢失，从而导致不自然的风格迁移结果。

4. 方法

我们提出了一种新的可控且可推广的三维模型语义感知风格迁移模型，该模型不仅可以实现对象选择和风格融合，还可以为语义风格图像提供更好的风格转换质量，整体的流程图见图 1。我们首先在 4.1 节中介绍了核心算法，然后在 4.2 节中展示如何实现不同的可控任务，在 4.3 节里介绍了创新点并与现有

方法进行了对比分析，最后，我们在 4.4 节中提供了本文方法的实现细节。

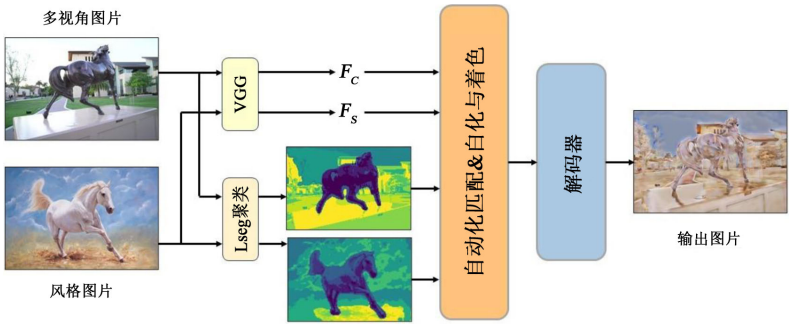

Figure 1. Pipeline of the semantic style transfer model for 3D scenes
图 1. 三维场景语义风格迁移模型流程图

4.1. 核心算法

给定一个相机视图的渲染图像 I_r，可以使用任何二维分割算法，如 LSeg [5]，得到具有 M 个类的分割掩码 M_r，其中每个像素 $M_r(x,y)\in\{0,1,\cdots,M\}$。对于每个类别内部的像素，在色彩空间里通过自适应的聚类和匹配算法确定对应关系，然后在对应的类别之间进行白化和着色，最后通过解码器输出最终结果。

4.1.1. 自适应聚类算法

MST 在 VGG 特征空间中使用 K 均值进行聚类和特征匹配，但由于特征维度过高且分辨率较低，所以难以保持图像的结构。三维风格迁移对结构完整性要求较高，所以我们先一步对原始图像进行分割来提取结构信息，具体来说，在颜色空间对内容图像和风格图像进行 K 均值聚类并生成聚类图。

在本研究中，我们选择 K 均值聚类算法进行颜色空间分割，主要基于以下几个原因：

1、简单有效，计算效率高

K 均值聚类算法是一种经典的无监督学习方法，其通过迭代优化聚类中心来对数据进行分组。相比其他更为复杂的聚类方法(如谱聚类、层次聚类等)，K 均值算法具有较高的计算效率，尤其适合在大规模数据上进行处理。在颜色空间分割中，K 均值能够迅速且有效地将图像中的像素进行分组，符合风格迁移过程中对处理速度的需求。

2、能够有效处理颜色空间的复杂性

图像的颜色信息包含了丰富的语义信息，尤其在进行风格迁移时，不同颜色区域的分割对保持风格一致性至关重要。K 均值聚类能够将颜色空间中相似的颜色像素聚集到一起，从而为后续的风格迁移提供更为清晰的区域划分。此外，K 均值聚类能够处理颜色空间中的高维数据，在颜色空间中能有效降低颜色变换对视觉效果的影响。

3、适应性强，能够根据图像内容自适应分割

K 均值聚类的另一大优势在于它能够根据图像的实际内容自适应地进行颜色分割。与基于固定阈值的传统方法相比，K 均值聚类不需要人工设定颜色类别，而是根据图像像素的分布情况自动决定类别。这样一来，K 均值聚类能够更好地适应不同风格和不同场景的图像内容，提升风格迁移的效果。

尽管 K 均值聚类算法在颜色空间分割中表现出色，但它的一个固有问题是聚类数量的选择。在传统的 K 均值算法中，K 值通常是事先设定的，这可能导致聚类效果不理想，尤其是在颜色变化丰富的图像中。为了克服这一问题，我们采用了**自适应聚类数量的策略**，基于图像复杂度(如颜色数量、亮度和面积等)自适应地确定聚类类别数目。具体算法流程见算法一，使用三个超参数：最大聚类数、最大比例和最

小比例来自适应的确定聚类个数，使得每个类别的面积比例尽可能低于最大比例，高于最小比例。

算法 1: 自适应聚类算法

输入： 图片 I 、最大聚类数 N 、最大比例 $p_{\max}$ 、最小比例 $p_{\min}$

输出： 图片 I 对应的聚类图 R_I

1) 图片 I 输入 LSeg 模型得到具有 M 个类的分割图 $M_I(x,y)\in\{0,1,\cdots,M\}$

2) **for** m in range $(0,M+1)$ **do**

3) 对于 $M_I(x,y)=m$，令 $K=2$ 执行 K 均值聚类得到聚类图 R_m

4) **while** 最大类所占比例≥ $p_{\max}$ **or** 最小类所占比例≥ $p_{\min}$ **do**

5) **if** $K>N$ **or** 第二小类所占比例≤ $p_{\min}$ **then**

6) $R_m=K-1$ 聚类图

7) $R_I I_{M_I(x,y)=m}=R_m$

8) **return**

9) $K=K+1$ 执行聚类得到聚类图 R_m

10) $R_I I_{M_I(x,y)=m}=R_m$

11) **return** R_I

4.1.2. 自动化匹配算法

在 4.1.1 小节里，我们通过自适应聚类算法分别对内容图像 I_c 和风格图像 I_s 进行聚类，得到了含有不同的语义类别的聚类图编号 $R_c=\{1,2,\cdots,K_c\}$ 和 $R_s=\{1,2,\cdots,K_s\}$。在实际情况中 K_c 与 K_s 不一定相等，所以需要设计一个算法可以将 R_c 中的元素与 R_s 中的一个或多个元素进行自动化匹配。换句话说，我们需要构造一个从 R_c 到 R_s 幂集(不包括空集)的映射。

算法 2 展现了自动化匹配算法的具体流程，对于事先得到的分割图 $M_I(x,y)\in\{0,1,\cdots,M\}$，提取内容图像和风格图像语义相同的类别，然后首先对聚类中心进行[0, 1]归一化并排序，设置超参数最小距离阈值 β，如果两张聚类图的聚类中心距离小于 β，则建立对应关系。在特别情况下，若设置 $\beta=1$，则算法类似于 WCT 算法；若设置的 β 过小，接近 0 时，则可能出现找不到风格聚类中心进行匹配，针对这种异常情况，将设置其与所有的风格聚类中心进行匹配。

算法 2: 自动化匹配算法

输入： 内容聚类图 R_c 、风格聚类图 R_s 、最小距离阈值 β

输出： 聚类匹配字典 D

1：初始化 $D=\varnothing$

2：计算聚类中心的范数并进行升序排序，令排序后的内容聚类编号 $R_c=\{1,2,\cdots,K_c\}$，聚类中心范数列表 $L_c=\{c_1,c_2,\cdots,c_{K_c}\}$；排序后的风格聚类编号 $R_s=\{1,2,\cdots K_s\}$，聚类中心范数列表 $L_s=\{s_1,s_2,\cdots,s_{K_s}\}$:

3：对 L_c 和 L_s 分别进行[0, 1]标准化

4：for i in R_c do:

5：$D[i]=\varnothing$:

6：matched = False

7：for j in R_s do:

8：if abs $(c_i-s_j)<\beta$ then

9：$D[i]=\{D[i],j\}$:

10：matched = True

11：if not matched then:

12：$D[i]=R_s$

续表

13：return D

4.1.3. 迁移算法

在匹配完成后，对应的类别之间会正式进行风格的迁移过程，具体来说，会进行白化、着色与解码三个过程。白化就是对内容图像的特征进行去相关，使其去除原本的统计特性(如均值和协方差)。设特征 F_c 为一个 $d\times N$ 维矩阵(d 为通道数， N 为像素数)，其协方差矩阵为 $\Sigma_c=\frac{1}{N-1}F_cF_c^{\mathrm{T}}$，对其进行特征分解 $\Sigma_c=U_c\Lambda_cU_c^{\mathrm{T}}$，然后进行白化，使协方差矩阵变为单位矩阵，从而去除特征的相关性：

$$\hat{F}_c=U_c\Lambda_c^{-\frac{1}{2}}U_c^{\mathrm{T}}F_c \tag{8}$$

接下来，需要将内容特征调整为风格特征的统计分布。通过类似的方式得到风格图像的协方差矩阵并对其进行特征分解 $\Sigma_s=\frac{1}{N-1}F_sF_s^{\mathrm{T}}=U_s\Lambda_sU_s^{\mathrm{T}}$，然后将去相关后的内容图像特征调整到风格特征的统计分布，并加上风格特征的均值：

$$F_{cs}=U_s\Lambda_s^{\frac{1}{2}}U_s^{\mathrm{T}}\hat{F}_c+\mu_s \tag{9}$$

经过上述变换， F_{cs} 有了风格特征的统计特性。然后将其输入 VGG 解码器即可得到最终的风格迁移图像。令内容损失 $L_{\text{content}}=\left\|\phi\left(F_{\text{output}}\right)-\phi\left(F_c\right)\right\|_2^2$，风格损失 $L_{\text{style}}=\sum_l\left\|\phi_l\left(F_{\text{output}}\right)-\phi_l\left(F_s\right)\right\|_2^2$，其中 ϕ 和 ϕ_l 分别表示 VGG 编码器和其第 l 层提取的特征，则总损失函数如下：

$$L_{\text{image}}=\lambda_{\text{content}}L_{\text{content}}+\lambda_{\text{style}}L_{\text{style}}+\lambda_{\text{tv}}L_{\text{tv}} \tag{10}$$

4.2. 可控风格迁移任务

在二维图片风格迁移的基础上，可以对三维物体进行可控的风格迁移。给定一个相机视图的渲染图像 I_r，可以使用任何二维分割算法，如 LSeg [5]，得到具有 M 个类的分割掩码 M_r，其中每个像素 $M_r(x,y)\in\{0,1,\cdots,M\}$。定义三维模型的优化损失为：

$$L=\left(\frac{1}{N}\sum_{x,y}\sum_m\boldsymbol{I}\left[M_r(x,y)=m\right]L^m(x,y)\right)+\lambda_{\text{tv}}\cdot l_{\text{tv}} \tag{11}$$

其中 L^m 表示类别 m 对应像素的损失函数，$\boldsymbol{I}(\cdot)$ 为示性函数。这个损失函数是基于单视角图片优化三维物体的，在三维模型优化中引入 LSeg 模型有很多好处，本研究选择 LSeg 模型进行语义分割，主要基于以下几个原因：

1) 精度和鲁棒性

LSeg 模型在处理复杂图像时，能够有效地捕捉图像的语义信息，并在多种场景中展现出较高的精度和鲁棒性。其通过端到端训练，能够从大规模数据中学习到更为细粒度的语义特征，有效提升分割精度，尤其是在高复杂度的场景中。

2) 适应性强

LSeg 模型具有较强的适应性，能够根据输入图像的特征自动调整网络结构。尤其在多视角图像训练中，LSeg 模型能够对每个视角中的物体进行精确分割，确保不同视角下的语义一致性。这对于保持三维模型的结构和风格一致性至关重要，因为每个视角的语义信息都能够被准确提取并应用于风格迁移。

3) 跨视图训练的优势

在多视角图像训练中，LSeg 模型通过对多视角的语义信息进行综合分析，能够有效克服单一视角下

的分割精度低和不一致性问题。由于三维场景往往包含多个视角，LSeg 模型能够通过从不同角度获得的信息，减少视角间的误差，提升整体的分割精度。特别是当处理具有复杂几何结构和遮挡的三维场景时，使用 LSeg 可以更好地理解各个物体的语义边界，保证在风格迁移过程中，结构信息得到有效的保留。

4) 与风格迁移的结合

通过语义分割，LSeg 能够为后续的风格迁移过程提供精确的区域划分。这样一来，我们可以基于不同的语义区域进行风格迁移，从而提高风格的细腻度和一致性。多视角的训练不仅加强了语义分割的精度，也保证了在不同视角下风格的迁移具有一致性，避免了风格转化过程中出现的割裂感。

在不同的任务中，L^m 可以有不同选择，接下来我们将介绍三个可控风格迁移任务的具体实现。

4.2.1. 选择性风格迁移

为了在三维场景中选择物体进行单独的风格迁移，可以对多视角图片进行 0-1 掩码，其中 1 表示需要进行风格迁移的物体，0 表示保留原样的部分，则损失函数为：

$$L^m(x,y)=\begin{cases} l_2 & m=0 \\ L_{\text{image}}+\lambda\cdot l_2 & m=1 \end{cases} \tag{12}$$

其中 L_{image} 为 4.1.3 节里提到的风格迁移损失，l_2 损失则控制渲染图片和参考图像的一致性。

4.2.2. 组合式风格迁移

对于不同的物体类别，可以输入不同的风格图像进行迁移，属于类别 m 的区域的损失函数定义为：

$$L^m(x,y)=L_{\text{image}}\left(F_{\text{output}}(x,y),F_s^m\right)+\lambda\cdot l_2\left(F_{\text{output}}(x,y),F_c(x,y)\right) \tag{13}$$

其中 F_s^m 表示从类别 m 对应的风格图片中提取的特征图。

4.2.3. 语义感知风格迁移

VGG 网络编码的大多是纹理、结构与颜色信息，而非语义信息，在此基础上我们使用与文本编码器一起训练的 LSeg 算法，提取内容图像与风格图像的语义信息，并进行语义匹配，然后就可以针对语义相近的区域进行特征匹配和风格迁移。如图 3 所示，可以将风格图像中的白马的特征转移到内容图像中马的雕像上。具体的损失函数与公式(13)相同，但 F_s^m 表示的是风格图片里与类别 m 进行语义匹配区域的特征图。

4.3. 创新点及对比分析

在本研究中，我们提出了一种基于三维高斯模型的可控语义风格迁移方法，旨在解决现有风格迁移方法在三维场景中的应用问题。与传统的风格迁移方法不同，本研究不仅关注图像的风格变化，还强调了在三维高斯模型的框架下如何保持三维结构的完整性与一致性。下面将介绍具体的创新点，并与现有方法进行了对比分析。

4.3.1. 创新点

1、三维高斯模型与风格迁移的结合

现有的风格迁移方法大多数集中于二维图像的处理，尤其是在 VGG、AdaIN 和 WCT 等方法中，风格迁移仅限于平面图像。尽管这些方法在平面图像风格化上取得了显著成效，但它们未能有效解决三维场景中复杂的几何结构和多视角一致性问题。我们的方法通过将三维高斯模型与风格迁移技术相结合，提出了一种全新的思路来处理三维场景的风格迁移。

具体而言，我们首先通过多视角图像训练三维高斯模型，将其作为三维场景的表示方式。通过将风

格迁移过程直接嵌入到三维高斯点云的特征空间中，我们能够在保持三维结构的同时实现风格迁移。相比于传统的基于网格(Mesh)或点云(Point Cloud)的方法，三维高斯模型不仅具有更高的灵活性和计算效率，还能够在保证高质量渲染效果的同时，生成逼真的新视角图像。

2、语义分割与风格迁移的结合

在传统的风格迁移方法中，风格的融合通常是在图像的像素层面进行的，缺乏对内容和风格之间语义的深入理解。为了提升风格迁移的精度和效果，我们在本研究中引入 LSeg 模型来处理内容图像和风格图像，将 LSeg 语义分割与自适应 K 均值聚类和匹配算法相结合，以提高三维高斯模型中的风格迁移精度和一致性。

首先，我们通过 LSeg 模型对内容图像和风格图像进行语义分割，提取出具有语义意义的区域。这一分割结果帮助我们明确了图像中的不同语义区域，从而能够对不同区域进行精细化处理。接下来，在这些分割区域的基础上，我们应用自适应 K 均值聚类进行颜色空间的划分。与传统的 K 均值聚类不同，我们的方法能够根据图像的复杂度自适应地确定聚类数目，确保在处理不同风格的图像时，能够有效地划分出有意义的风格区域。聚类后，我们进一步使用自动化匹配算法来精确对齐内容图像与风格图像中对应的语义区域，从而保证风格迁移过程中语义区域的一致性和结构保持性。

通过这种方式，不仅使风格迁移在细节层面更加精确，也有效避免了传统方法中风格区域不一致的问题，提升了多视角风格迁移的整体质量。

4.3.2. 对比分析

传统的风格迁移方法，如基于 VGG 网络的风格提取和 AdaIN 等方法，通常只能处理二维图像，且风格迁移过程中容易失去结构信息，特别是在多视角的情况下，可能会产生风格不一致的问题。对于三维场景，现有方法难以有效地保留三维结构的完整性，风格迁移效果往往偏向于平面图像的效果，导致渲染结果的层次感和真实感不足。

三维点云和网格方法能够较好地保留三维结构信息，但它们在风格迁移方面的处理仍然有限。点云方法虽然可以表示三维场景，但缺乏对细节和局部结构的精确建模，难以进行高质量的风格迁移。而网格方法虽然能够在局部区域进行风格转换，但往往存在模型重建精度不高、渲染效果不一致的问题。

本文方法使用的白化和着色过程来源于 WCT，它在传统的风格迁移中已经取得了广泛的应用，其核心思想是通过对内容图像进行白化和着色变换，将图像的特征空间与风格图像进行匹配，从而实现风格迁移。在二维图像中，WCT 方法能够较为有效地将风格信息传递到内容图像中，且在一些简单的场景中，能够达到较好的视觉效果。然而，将 WCT 方法应用于 三维高斯模型 中时，面临着一些特有的挑战和局限性。

首先，WCT 方法通常基于图像的特征空间进行变换，而在三维场景的表示中，三维高斯模型以点云或体素的形式表示三维结构信息，这就要求我们在风格迁移时，不仅需要考虑图像的颜色和纹理信息，还需要保持三维结构的一致性。然而，WCT 方法的线性变换并未专门考虑三维结构信息，尤其是在处理复杂的三维形状时，可能导致在三维场景中风格迁移时出现结构损失或失真。

其次，WCT 方法在风格迁移过程中会将图像的颜色特征进行线性变换，这使得其在处理具有强烈几何结构的三维场景时，容易忽略三维物体的空间关系，导致不同视角下的风格不一致或结构错位。而我们提出的方法通过 LSeg 语义分割和自适应 K 均值聚类，在分割出图像的语义区域后，基于这些区域进行风格迁移，有效避免了 WCT 方法在结构保持上的局限性。

最后，WCT 方法在多视角场景中应用时，由于特征变换的方式较为统一，可能无法处理不同视角之间的风格一致性问题。不同于 WCT 的单一特征变换，我们的方法通过自动化匹配算法精确地对齐不同

视角中的语义区域，确保了多视角下的风格迁移一致性。具体而言，在多视角下，我们的方法能够根据每个视角的图像特征和语义分割结果，自动调整风格迁移的参数，保证了风格的统一性和结构的完整性。与 WCT 的具体对比结果在第五章里有所展示。

因此，尽管 WCT 方法在某些简单的二维场景中表现优秀，但在应用于三维高斯模型时，其线性变换的局限性使得其难以有效处理三维结构的保持和多视角的一致性问题。相比之下，我们的方法通过结合语义分割、自适应聚类和匹配，能够更好地解决三维场景中的风格迁移挑战，保持更高的风格一致性和结构精度。

4.4. 实施细节

在自适应聚类算法中，我们将超参数设置为最大聚类数 $N=5$、最大比例 $p_{\max}=0.7$、最小比例 $p_{\min}=0.1$。在自动化匹配算法，设置最小距离阈值 $\beta=0.25$。对于二维图片风格迁移损失函数 L_{image}，设置 $\lambda_{\text{content}}=\lambda_{\text{style}}=1$，$\lambda_{tv}=0.1$。对于三维高斯模型的优化过程，设置内容损失权重 $\lambda=0.005$。我们使用预训练的 VGG 编码器与 LSeg 模型，然后使用来自 MS-COCO [36]和 WikiArt [37]的数据分别作为内容图像和风格图像来训练解码器。在学习过程中，内容图像和风格图像的分辨率大小均调整为 512×512，每批次输入四张图片进行训练，使用 Adam [38]学习器，学习率设置为 10^{-5}。

5. 实验

我们进行实验来评估 4.2 节里的三个任务：选择性风格迁移、组合式风格迁移和语义感知风格迁移任务，与本文方法相近的 WCT 和 MST 进行对比从而验证自适应聚类和匹配算法模块的有效性。使用的数据集来自[20]和[39]。

5.1. 选择性风格迁移

我们选择物体进行风格迁移，如图 2(a)所示。针对三维场景的多视角图片，我们使用 LSeg 得到多视角掩模，对选择的物体进行自适应聚类并与风格图像进行匹配，从而进行局部风格化。如结果所示，我们的方法可以正确地将样式转换应用于所选区域，而其他区域则保持不变。

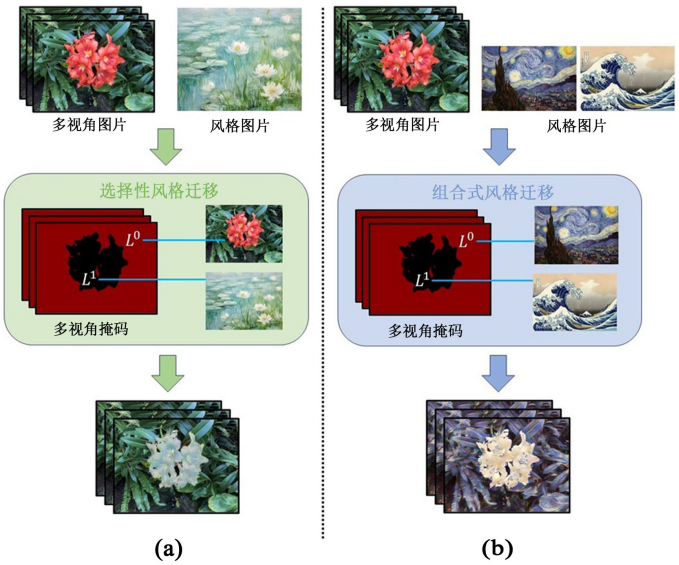

Figure 2. Illustration of selective style transfer and compositional style transfer
图 2. 选择性风格迁移与组合式风格迁移示意图

5.2. 组合式风格迁移

由于使用 LSeg 模型对多视角图片进行语义分割，就可以针对不对分割区域，将其与不同的风格图像进行匹配，从而可以将多种风格组合在一起，进行组合式风格迁移，结果如图 2(b)所示。我们将场景中的花与背景分离，分别与两张不同的风格图像进行匹配，生成组合样式场景。

由于本文方法是基于对三维场景的优化，因此可以进一步推广到其他的三维场景表示方法(如神经辐射场等)，以提高渲染质量或加快风格化收敛。

5.3. 语义感知风格迁移

在图 3 中，我们展示了语义感知风格迁移的效果，并将其与现有的 WCT 和 MST 方法进行了对比。从结果中可以明显看出，由于采用了自适应聚类和自动化匹配算法，我们的方法在三维场景的结构化保持方面表现得更加出色。与 WCT 方法相比，我们的方法能够更精确地保持三维场景中的几何结构，生成的风格迁移图像更具清晰度和细节。此外，得益于自适应聚类的特性，我们能够在风格迁移过程中有效避免了结构失真和风格不一致的问题，从而提供了更自然的风格过渡效果。

Figure 3. Comparison of 3D semantic-aware style transfer
图 3. 三维语义感知风格迁移对比图

与 MST 方法的对比进一步突显了我们方法在语义信息匹配上的优势。以图 3 中的第二列卡车场景为例，采用我们的方法后，风格图像中的车身颜色和纹理得以准确地转移到三维卡车模型上，并且背景的颜色成功地转移到了后方街道中。相反，MST 方法的风格迁移结果存在颜色转换不准确和部分区域风格丢失的情况。例如，MST 在处理卡车的纹理时，未能有效地将纹理特征转移到三维场景中，导致风格化效果明显缺乏一致性。

这种差异的根本原因在于我们引入了 LSeg 模型进行语义分割，该模型能够精确地捕捉到图像中的语义特征，如物体的边界、形状和语义标签。通过这种方式，我们能够在风格迁移时根据语义区域进行精细化控制，从而实现更高精度的风格匹配。而 MST 方法仅依赖于 VGG 特征编码，它主要关注图像的颜色和纹理信息，缺乏对语义信息的深入理解和处理，这使得其在复杂场景中的风格迁移效果有所限制。

综上所述，本文提出的语义感知风格迁移方法，在保持三维结构的同时，能够更准确地匹配和迁移语义信息，解决了传统方法在多视角和复杂场景中的不足，提供了更加自然和一致的风格迁移效果。

6. 总结

本文提出了针对三维场景的可控式风格迁移方法，包括选择风格迁移、组合式风格转换和语义感知风格迁移。我们引入多视图掩码，并对不懂区域施加不同的损失函数进行优化。本文的方法也可以推广到其他的三维表示方法上，如 CLIPNeRF [40]，以实现用户定义的可控性。实验表明，我们的方法不仅具有更好的效果和鲁棒性，并且可以根据任务不同，进行更好的语义感知把控。

参考文献

[1] Kerbl, B., Kopanas, G., Leimkuehler, T. and Drettakis, G. (2023) 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, **42**, 1-14. https://doi.org/10.1145/3592433

[2] Huang, X. and Belongie, S. (2017) Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 1501-1510. https://doi.org/10.1109/iccv.2017.167

[3] Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X. and Yang, M.H. (2017) Universal Style Transfer via Feature Transforms. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 386-396.

[4] Zhang, Y., Fang, C., Wang, Y., Wang, Z., Lin, Z., Fu, Y., *et al*. (2019) Multimodal Style Transfer via Graph Cuts. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 5942-5950. https://doi.org/10.1109/iccv.2019.00604

[5] Li, B., Weinberger, K.Q., Belongie, S., Koltun, V. and Ranftl, R. (2022) Language-Driven Semantic Seg Mentation.

[6] Kyprianidis, J.E., Collomosse, J., Wang, T. and Isenberg, T. (2013) State of the "Art": A Taxonomy of Artistic Stylization Techniques for Images and Video. *IEEE Transactions on Visualization and Computer Graphics*, **19**, 866-885. https://doi.org/10.1109/tvcg.2012.160

[7] Portilla, J. and Simoncelli, E.P. (2000) A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, **40**, 49-70.

[8] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition.

[9] Risser, E., Wilmot, P. and Barnes, C. (2017) Stable and Controllable Neural Texture Synthesis and Style Transfer Using Histogram Losses.

[10] Gu, S., Chen, C., Liao, J. and Yuan, L. (2018). Arbitrary Style Transfer with Deep Feature Reshuffle. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8222-8231. https://doi.org/10.1109/cvpr.2018.00858

[11] Kolkin, N., Salavon, J. and Shakhnarovich, G. (2019) Style Transfer by Relaxed Optimal Transport and Self-Similarity. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 10051-10060. https://doi.org/10.1109/cvpr.2019.01029

[12] Liao, J., Yao, Y., Yuan, L., Hua, G. and Kang, S.B. (2017) Visual Attribute Transfer through Deep Image Analogy. *ACM Transactions on Graphics*, **36**, 1-15. https://doi.org/10.1145/3072959.3073683

[13] An, J., Huang, S., Song, Y., Dou, D., Liu, W. and Luo, J. (2021) Artflow: Unbiased Image Style Transfer via Reversible Neural Flows. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 862-871. https://doi.org/10.1109/cvpr46437.2021.00092

[14] Chen, C. (2020) Structure-Emphasized Multimodal Style Transfer. Zenodo.

[15] Park, D.Y. and Lee, K.H. (2019) Arbitrary Style Transfer with Style-Attentional Networks. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 5880-5888. https://doi.org/10.1109/cvpr.2019.00603

[16] Liu, K., Zhan, F., Chen, Y., Zhang, J., Yu, Y., Saddik, A.E., *et al*. (2023) StyleRF: Zero-Shot 3D Style Transfer of Neural Radiance Fields. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 8338-8348. https://doi.org/10.1109/cvpr52729.2023.00806

[17] Martin-Brualla, R., Radwan, N., Sajjadi, M.S.M., Barron, J.T., Dosovitskiy, A. and Duckworth, D. (2021) Nerf in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 7210-7219. https://doi.org/10.1109/cvpr46437.2021.00713

[18] Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P.P. and Barron, J.T. (2022) NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 16190-16199. https://doi.org/10.1109/cvpr52688.2022.01571

[19] Mishra, S. and Granskog, J. (2022) Clip-Based Neural Neighbor Style Transfer for 3D Assets.

[20] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R. and Ng, R. (2021) NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, **65**, 99-106. https://doi.org/10.1145/3503250

[21] Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., de Mello, S., *et al*. (2022) Efficient Geometry-Aware 3D Generative Adversarial Networks. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 16123-16133. https://doi.org/10.1109/cvpr52688.2022.01565

[22] Chen, A., Xu, Z., Geiger, A., Yu, J. and Su, H. (2022) TensoRF: Tensorial Radiance Fields. 17*th European Conference on Computer Vision*, Tel Aviv, 23-27 October 2022, 333-350. https://doi.org/10.1007/978-3-031-19824-3_20

[23] Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B. and Kanazawa, A. (2023) K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 12479-12488. https://doi.org/10.1109/cvpr52729.2023.01201

[24] Kato, H., Ushiku, Y. and Harada, T. (2018) Neural 3D Mesh Renderer. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 3907-3916. https://doi.org/10.1109/cvpr.2018.00411

[25] Michel, O., Bar-On, R., Liu, R., Benaim, S. and Hanocka, R. (2022) Text2mesh: Text-Driven Neural Stylization for Meshes. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 13492-13502. https://doi.org/10.1109/cvpr52688.2022.01313

[26] Yin, K., Gao, J., Shugrina, M., Khamis, S. and Fidler, S. (2021) 3DStyleNet: Creating 3D Shapes with Geometric and Texture Style Variations. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 12456-12465. https://doi.org/10.1109/iccv48922.2021.01223

[27] Huang, H., Tseng, H., Saini, S., Singh, M. and Yang, M. (2021) Learning to Stylize Novel Views. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 13869-13878. https://doi.org/10.1109/iccv48922.2021.01361

[28] Zhang, K., Kolkin, N., Bi, S., Luan, F., Xu, Z., Shechtman, E., *et al*. (2022) ARF: Artistic Radiance Fields. 17*th European Conference on Computer Vision*, Tel Aviv, 23-27 October 2022, 717-733. https://doi.org/10.1007/978-3-031-19821-2_41

[29] Gatys, L., Ecker, A. and Bethge, M. (2016) A Neural Algorithm of Artistic Style. *Journal of Vision*, **16**, Article No. 326. https://doi.org/10.1167/16.12.326

[30] Huang, Y., He, Y., Yuan, Y., Lai, Y. and Gao, L. (2022) StylizedNeRF: Consistent 3D Scene Stylization as Stylized Nerf via 2D-3D Mutual Learning. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 18342-18352. https://doi.org/10.1109/cvpr52688.2022.01780

[31] Wang, C., Jiang, R., Chai, M., He, M., Chen, D. and Liao, J. (2024) NeRF-Art: Text-Driven Neural Radiance Fields Stylization. *IEEE Transactions on Visualization and Computer Graphics*, **30**, 4983-4996. https://doi.org/10.1109/tvcg.2023.3283400

[32] Xu, S., Li, L., Shen, L. and Lian, Z. (2023) DeSRF: Deformable Stylized Radiance Field. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Vancouver, 17-24 June 2023, 709-718. https://doi.org/10.1109/cvprw59228.2023.00078

[33] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., *et al*. (2021) Learning Transferable Visual Models from Natural Language Supervision. *International Conference on Machine Learning*, 18-24 July 2021, 8748-8763.

[34] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2020) An Image Is Worth 16 x 16 Words: Transformers for Image Recognition at Scale.

[35] Gal, R., Patashnik, O., Maron, H., Bermano, A.H., Chechik, G. and Cohen-Or, D. (2022) StyleGAN-NADA: Clip Guided Domain Adaptation of Image Generators. *ACM Transactions on Graphics*, **41**, 1-13. https://doi.org/10.1145/3528223.3530164

[36] Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., *et al*. (2014) Microsoft COCO: Common Objects in Context. 13*th European Conference on Computer Vision*, Zurich, 6-12 September 2014, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[37] Nichol, K. (2016) Painter by Numbers. Wikiart. https://github.com/inejc/painters

[38] Kingma, D.P. and Ba, J.L. (2014) Adam: A Method for Stochastic Optimization.

[39] Knapitsch, A., Park, J., Zhou, Q. and Koltun, V. (2017) Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics*, **36**, 1-13. https://doi.org/10.1145/3072959.3073599

[40] Wang, C., Chai, M., He, M., Chen, D. and Liao, J. (2022) CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields. 2022 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 3835-3844. https://doi.org/10.1109/cvpr52688.2022.00381