Gaussian in the Wild: 3D Gaussian Splatting for Unconstrained Image Collections

Dongbin Zhang*, Chuming Wang*, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang†

Tsinghua Shenzhen International Graduate School, Tsinghua University

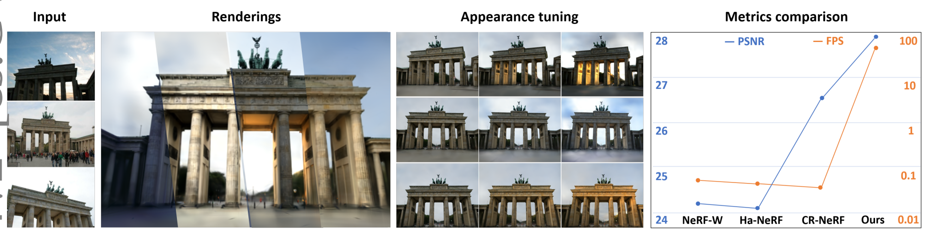

Fig. 1: With an unconstrained image collection input, GS-W can render novel views with appearance tuning, achieving state-of-the-art quality and faster rendering speed.

Abstract. Novel view synthesis from unconstrained in-the-wild images remains a meaningful but challenging task. The photometric variation and transient occluders in those unconstrained images make it difficult to reconstruct the original scene accurately. Previous approaches tackle the problem by introducing a global appearance feature in Neural Radiance Fields (NeRF). However, in the real world, the unique appearance of each tiny point in a scene is determined by its independent intrinsic material attributes and the varying environmental impacts it receives. Inspired by this fact, we propose Gaussian in the wild (GS-W), a method that uses 3D Gaussian points to reconstruct the scene and introduces separated intrinsic and dynamic appearance feature for each point, capturing the unchanged scene appearance along with dynamic variation like illumination and weather. Additionally, an adaptive sampling strategy is presented to allow each Gaussian point to focus on the local and detailed information more effectively. We also reduce the impact of transient occluders using a 2D visibility map. More experiments have demonstrated better reconstruction quality and details of GS-W compared to NeRF-based methods, with a faster rendering speed. Video results and code are available at https://eastbeanzhang.github.io/GS-W/.

Keywords: Novel view synthesis · 3D Gaussian Splatting · Unconstrained image collections

* Equal contributions.
† Corresponding author.

arXiv:2403.15704v2 [cs.CV] 14 Jul 2024

野外的高斯：无约束图像集合的三维高斯散射

张东斌*，王楚明*，王伟涛，李培豪，秦明翰，王浩谦†

清华大学深圳国际研究生院，清华大学

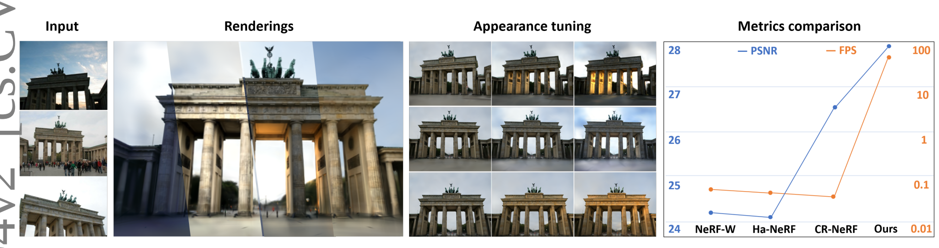

图1： 通过无约束图像集合输入，GS-W可以渲染具有外观调优的新视图，实现最先进的质量和更快的渲染速度。

摘要。 从无约束的野外图像中进行新视角合成仍然是一个有意义但具有挑战性的任务。这些无约束图像中的光度变化和瞬态遮挡物使得准确重建原始场景变得困难。以前的方法通过在神经辐射场（NeRF）中引入全局外观特征来解决这个问题。然而，在现实世界中，场景中每个微小点的独特外观由其独立的内在材料属性和接收到的不断变化的环境影响决定。受此启发，我们提出了野外高斯（GS-W），一种使用三维高斯点重建场景的方法，并为每个点引入了分离的内在和动态外观特征，捕捉不变的场景外观以及如照明和天气等动态变化。此外，还提出了一种自适应采样策略，使每个高斯点能够更有效地关注局部和详细信息。我们还使用二维可见性图来减少瞬态遮挡物的影响。更多实验表明，与基于NeRF的方法相比，GS-W具有更好的重建质量和细节，同时具有更快的渲染速度。视频结果和代码可在https://eastbeanzhang.github.io/GS-W/上获得。

关键词： 新视角合成 · 三维高斯散射 · 无约束图像集合

* 同等贡献。
† 通讯作者。

1 Introduction

Novel view synthesis has long been a high-profile and complicated task in computer vision which aims to recover the 3D structure of a scene from 2D image collections and plays a significant role in many applications like virtual reality (VR) and autonomous driving. Recently, implicit representations, especially Neural Radiance Field (NeRF) [26] and its subsequent work [1,2,27] have shown impressive progress in rendering photorealistic images from arbitrary viewpoints.

Meanwhile, explicit representations have also drawn increasing attention thanks to their real-time rendering speed. 3D Gaussian Splatting(3DGS) [16] introduces 3D Gaussian as a novel and flexible scene representation and designs a fast differentiable rendering approach, allowing real-time rendering while maintaining high-fidelity for view synthesis.

However, these aforementioned methods only focus on static scenes from which images are captured without transient occluders and dynamic appearance variation such as ever-changing sky, weather, and illumination. Unfortunately, in practice, input unconstrained image collections may be captured by cameras with different settings at different times, and typically include pedestrians or vehicles. The assumption of the previous methods that images must be captured in static scenes is severely violated, resulting in sharp performance degradation [23, 50].

Several recent attempts [6, 19, 23, 41] adopt a global latent embedding for each image, granting their model the ability to handle the appearance variations between images. CR-NeRF [50] proposes a new cross-ray paradigm that utilizes information from multiple rays to obtain colors of several pixels, achieving more realistic and efficient appearance modeling. Nonetheless, the methods above still suffer from three defects: **1) Global appearance representation:** The representation they use to control the dynamic appearance variation is shared by the whole scene, which may struggle to describe local high-frequency changes. By contrast, in the 3D real world, every single point of the scene has its appearance feature which brings them unique gloss and texture in different views; **2) Blurring intrinsic and dynamic appearance:** An object's intrinsic appearance is determined by its own material and surface properties while dynamic appearance is affected by environmental factors like highlight and shadow. Previous approaches blur the two appearances together, causing confusion in applications such as appearance tuning; **3) High time cost:** Similar to volume rendering based methods, most NeRF-based approaches above suffer from high training costs and low rendering speed due to a large amount of network evaluation.

To address these challenges, we propose Gaussian in the wild (GS-W), a method to achieve high-quality and flexible scene reconstruction for unconstrained image collections. Specifically, we first use 3D Gaussian points to represent the scene and introduce independent appearance features to each point, enabling their unique appearance expression. Second, the intrinsic and dynamic appearance is separated and an adaptive sampling strategy is presented to grant every point the freedom to focus on various detailed dynamic appearance information. Additionally, the scene rendering speed is significantly accelerated thanks to the tile-based rasterizer.

1 引言

新视角合成长期以来一直是计算机视觉中一个备受关注且复杂的任务，其目标是从二维图像集合中恢复场景的三维结构，并在虚拟现实（VR）和自动驾驶等许多应用中发挥重要作用。最近，隐式表示，特别是神经辐射场（NeRF） [26] 及其后续工作 [1,2,27] 在从任意视点渲染逼真图像方面取得了令人印象深刻的进展。

同时，显式表示由于其实时渲染速度也引起了越来越多的关注。三维高斯散射（3DGS） [16] 引入三维高斯作为一种新颖且灵活的场景表示，并设计了一种快速可微渲染方法，允许在保持高保真视图合成的同时进行实时渲染。

然而，上述方法仅关注静态场景，这些场景中的图像在没有瞬态遮挡物和动态外观变化（如不断变化的天空、天气和照明）的情况下拍摄。不幸的是，在实践中，输入的无约束图像集合可能由不同设置的相机在不同时间拍摄，并且通常包括行人或车辆。先前方法假设图像必须在静态场景中拍摄的假设被严重违反，导致性能急剧下降 [23, 50]。

最近的一些尝试 [6, 19, 23, 41] 采用每个图像的全局潜在嵌入，赋予其模型处理图像之间外观变化的能力。CR-NeRF [50] 提出了一种新的交叉射线范式，利用多条射线的信息来获取多个像素的颜色，实现了更真实和高效的外观建模。然而，上述方法仍然存在三个缺陷：**1) 全局外观表示：** 他们用来控制动态外观变化的表示是整个场景共享的，这可能难以描述局部高频变化。相比之下，在三维真实世界中，场景的每个点都有其外观特征，这使得它们在不同视图中具有独特的光泽和纹理；**2) 模糊内在和动态外观：** 物体的内在外观由其自身的材质和表面属性决定，而动态外观则受环境因素（如高光和阴影）的影响。先前的方法将这两种外观模糊在一起，导致在外观调优等应用中产生混淆；**3) 高时间成本：** 类似于基于体积渲染的方法，上述大多数基于NeRF的方法由于大量网络评估而遭受高训练成本和低渲染速度。

为了解决这些挑战，我们提出了野外高斯（GS-W），一种实现高质量和灵活场景重建的方法，适用于无约束图像集合。具体来说，我们首先使用3D高斯点来表示场景，并为每个点引入独立的外观特征，使其能够表达独特的外观。其次，分离了内在和动态外观，并提出了一种自适应采样策略，使每个点能够自由地关注各种详细的动态外观信息。此外，由于基于瓦片的光栅化器，场景渲染速度显著加快。

Our contribution can be summarized as follows:

- We propose a new framework GS-W, a 3D Gaussian Splatting based method, in which each Gaussian point is equipped with separated intrinsic and dynamic appearance features to enable more flexible varying appearance modeling from unconstrained image collections.
- To better incorporate environmental factors from the image into the scene, we propose adaptive sampling, allowing each point to sample dynamic appearance features more effectively from the feature maps, thereby focusing on more local and detailed information.
- Experimental results demonstrate that our method not only outperforms the state-of-the-art NeRF-based methods in terms of quality but also surpasses them in rendering speed by over 1000×.

2 Related Work

2.1 3D representations

Diverse 3D representations are developed to represent the geometric and appearance information of three-dimensional objects or scenes, among which implicit and explicit representation are two common methods for practical applications like 3D object generation and scene reconstruction. Implicit representation represents 3D data as continuous functions or fields, like occupancy fields [24], distance fields [29], color, and density. NeRF [26] is an outstanding work among them which models the scene as a continuous field of density and radiance. NeRF indicates that using MLP can represent complex scenes and render novel photo-realistic views with the help of volume rendering. On the contrary, explicit representation describes scenes by storing and manipulating 3D data using discrete structures like meshes [14, 15, 45], point clouds [31, 32, 40], voxels [38, 47, 48] and so on. More recently, methods represented by 3DGS [16] have entered researchers' vision by their real-time rendering speed while preserving high-resolution synthesis quality. An efficient tile-based rasterizer is introduced by 3DGS to splat Gaussians to the image plane and accelerate the rendering speed many times compared to previous methods. Many researchers are exploring its potential by extending it to various tasks [33, 46, 52]. Several hybrid representations [3–5, 10, 39] are also emerging and creating more possibilities.

2.2 Novel view synthesis

Synthesizing arbitrary views of a scene using a set of 2D images is a long-standing problem in computer vision. Many NeRF-based methods [11, 12, 27, 34, 53] have achieved expressive synthesis quality [1, 42, 49] along with good view-consistency [7, 28, 43]. Mip-NeRF [1] replaces the rays with a 3D conical frustum and proposes integrated position embedding to anti-aliasing. Instant-NGP [27] introduces multi-resolution hash encoding that permits the use of a smaller network to reduce the training cost. There are also some Gaussian-based methods [9, 22, 51, 54] contributing to this task by modifying 3DGS. For example,

我们的贡献可以总结如下：

我们提出了一种新的框架GS-W，这是一种基于三维高斯散射的方法，其中每个高斯点都配备了分离的内在和动态外观特征，以实现从无约束图像集合中更灵活的外观变化建模。

为了更好地将图像中的环境因素融入场景，我们提出了自适应采样，使每个点能够更有效地从特征图中采样动态外观特征，从而关注更多局部和详细的信息。

– 实验结果表明，我们的方法不仅在质量上优于最先进的基于NeRF的方法，而且在渲染速度上也超过了它们，超过 1000×。

2 相关工作

2.1 3D表示

为了表示三维物体或场景的几何和外观信息，开发了多种3D表示方法，其中隐式和显式表示是两种常见的方法，适用于3D物体生成和场景重建等实际应用。隐式表示将3D数据表示为连续函数或场，如占据场 [24], 距离场 [29], 颜色和密度。NeRF [26]是其中杰出的工作，它将场景建模为密度和辐射的连续场。NeRF表明，使用MLP可以表示复杂场景，并在体积渲染的帮助下渲染出新颖的逼真视图。相反，显式表示通过使用离散结构（如网格 [14, 15, 45], 点云 [31, 32, 40], 体素 [38, 47, 48]等）存储和操作3D数据来描述场景。最近，以3DGS [16]为代表的方法因其实时渲染速度而进入研究人员的视野，同时保持了高分辨率合成质量。3DGS引入了一种高效的基于瓦片的光栅化器，将高斯分布投影到图像平面，与之前的方法相比，渲染速度提高了许多倍。许多研究人员正在通过将其扩展到各种任务 [33, 46, 52]来探索其潜力。一些混合表示 [3–5, 10, 39] 也在不断涌现，创造了更多可能性。

2.2 新视角合成

使用一组2D图像合成场景的任意视图是计算机视觉中一个长期存在的问题。许多基于NeRF的方法 [11, 12, 27, 34, 53] 已经实现了富有表现力的合成质量 [1, 42, 49] ，同时保持了良好的视图一致性[7, 28, 43]。Mip-NeRF [1] 用3D锥形截头体替换光线，并提出集成位置嵌入以进行抗锯齿。Instant-NGP [27] 引入了多分辨率哈希编码，允许使用较小的网络来降低训练成本。还有一些基于高斯的方法 [9, 22, 51, 54] 通过修改3DGS来贡献于这项任务。例如，

to handle scenes with specular elements, Spec-Gaussian [51] departs from using spherical harmonics and instead adopts an anisotropic spherical Gaussian appearance field to model each point. Since these aforementioned methods all assume that input images are captured in a static scene, their performance declines intensely when reconstructing from unconstrained photo collections. Thus, several attempts [6,19,23,25,36,50] are proposed to address this challenging in-the-wild task by handling appearance variation and transient occluders. Other works [20,21] focus on scenes with time-varying appearances, while methods [8,18,55] use physical rendering models for diverse lighting conditions. This field still faces some remaining issues and looks forward to advancements.

As one of them, our proposed method tries to push the field one step forward by achieving a more delicate and flexible synthesis with higher rendering speed, through our modifications mentioned in Sec. 4.

3 Preliminaries

3D Gaussian Splatting (3DGS) [16] is a method for reconstructing 3D scenes from static images with camera pose information. It uses explicit 3D Gaussian points GP to represent the scene and achieves real-time image rendering through a differentiable tile-based rasterizer. These Gaussian points' positions X are initialized with point clouds extracted by SFM [37] from the image set. Particularly, it uses 3D covariance Σ to model the impact of each Gaussian point on the color anisotropy of the surrounding area:

$$G(x-X,\Sigma)=e^{-\frac{1}{2}(x-X)^T\Sigma^{-1}(x-X)} \tag{1}$$

For ease of optimizing the covariance Σ while maintaining its positive semi-definiteness, the method decomposes the covariance of each Gaussian point into a scaling matrix S and a rotation matrix R, which are then stored as the Gaussian point attributes s and r respectively, using 3D vectors and quaternions.

$$\Sigma=RSS^TR^T \tag{2}$$

Additionally, each Gaussian point is equipped with two more attributes: opacity α and color c, with the color attribute represented by third-order spherical harmonic coefficients. When rendering, besides projecting each Gaussian point onto a grid of 16×16 tiles on the image plane, the 3D covariance Σ is projected to 2D Σ' using the viewing transformation W and the Jacobian of the affine approximation of the projective transformation J:

$$\Sigma'=JW\Sigma W^TJ^T \tag{3}$$

Then, based on the Gaussian points sorted by the rasterizer, the color of each pixel is aggregated using α-blending:

$$\sigma_i=G(px'-X_i,\Sigma_i') \tag{4}$$

$$C(px')=\sum_{i\in GP_{px'}}c_i\sigma_i\prod_{j=1}^{i-1}(1-\sigma_j) \tag{5}$$

为了处理具有镜面元素的场景，Spec-Gaussian [51] 放弃了使用球谐函数，而是采用各向异性球面高斯外观场来建模每个点。由于上述方法都假设输入图像是在静态场景中捕获的，因此在从无约束照片集合中重建时，它们的性能会显著下降。因此，提出了几种尝试 [6,19,23,25,36,50]来通过处理外观变化和瞬态遮挡物来解决这个具有挑战性的野外任务。其他工作 [20,21] 专注于具有时变外观的场景，而方法 [8,18,55] 使用物理渲染模型来处理不同的光照条件。该领域仍然面临一些剩余问题，并期待着进步。

作为其中之一，我们提出的方法试图通过我们在第4节中提到的修改，实现更精细和灵活的合成，以更高的渲染速度，将领域向前推进一步。

3 预备知识

3D高斯溅射 (3DGS) [16] 是一种从具有相机姿态信息的静态图像中重建3D场景的方法。它使用显式的3D高斯点 GP 来表示场景，并通过可微分基于瓦片的光栅化器实现实时图像渲染。这些高斯点的位置 X 通过从图像集中提取的点云 [37] 进行初始化。特别是，它使用3D协方差 Σ 来建模每个高斯点对周围区域颜色各向异性的影响：

$$G(x-X,\Sigma)=e^{-\frac{1}{2}(x-X)^T\Sigma^{-1}(x-X)} \tag{1}$$

为了在保持协方差 Σ 的正半定性的同时方便优化，该方法将每个高斯点的协方差分解为缩放矩阵 S 和旋转矩阵 R，然后分别使用3D向量和四元数将其存储为高斯点属性 s 和 r 。

$$\Sigma=RSS^TR^T \tag{2}$$

此外，每个高斯点还具有两个额外属性：不透明度 α 和颜色 c，其中颜色属性由三阶球谐系数表示。在渲染时，除了将每个高斯点投影到图像平面上的 16×16 瓦片网格上，3D 协方差 Σ 通过视图变换 W 和投影变换的仿射近似的雅可比矩阵 J 投影到 2D Σ' ：

$$\Sigma'=JW\Sigma W^TJ^T \tag{3}$$

然后，基于光栅化器排序的高斯点，使用 α-混合来聚合每个像素的颜色：

$$\sigma_i=G(px'-X_i,\Sigma_i') \tag{4}$$

$$C(px')=\sum_{i\in GP_{px'}}c_i\sigma_i\prod_{j=1}^{i-1}(1-\sigma_j) \tag{5}$$

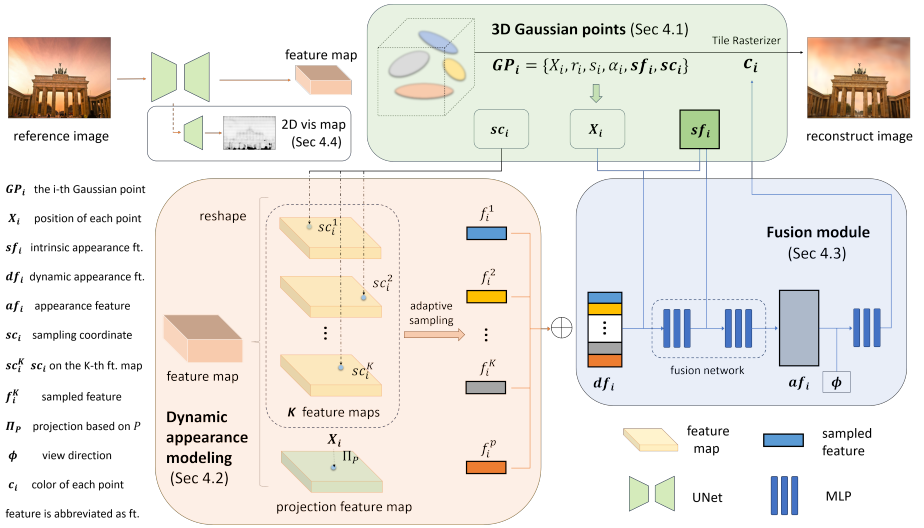

Fig. 2: An overview of the GS-W framework. We begin with a scene's reference image and its camera pose P. After extracting image features via a Unet model, we reshape them into K feature maps and one projection feature map. Each Gaussian point GP_i then samples features from these maps adaptively, capturing dynamic appearance feature df_i. These features are fused with the intrinsic appearance feature sf_i through a fusion network, decoded for Gaussian point color c_i. Finally, all Gaussian points are rendered using a tile rasterizer.

Where px' represents the position of a pixel, and $GP_{px'}$ denotes the sorted Gaussian points associated with that pixel. The final rendered image is then used to compute loss with reference images for training, jointly optimizing all Gaussian attributes. Moreover, it devises a strategy for point growth and pruning based on gradients and opacity.

4 Method

Based on previous analysis, aforementioned NeRF-based methods [6,23,50] lack enough attention to high-frequency and local detailed information in appearance, along with significant rendering costs due to the large number of sampling points. To address these issues, we utilize 3D Gaussian points to explicitly model the scene in Sec. 4.1 and introduce a new appearance modeling method for each Gaussian point in Sec. 4.2. Intrinsic and dynamic appearance features are separated and then fused in Sec. 4.3. Additionally, when calculating the losses in Sec. 4.5, a visibility map is employed to reduce the impact of transient objects in Sec. 4.4. The whole pipeline is visualized as Fig. 2.

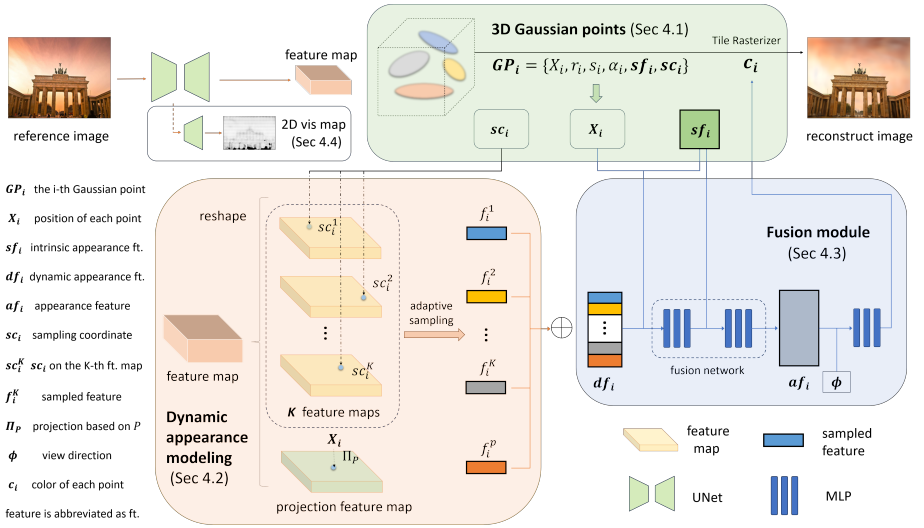

图2: GS-W框架概述。我们从场景的参考图像及其相机姿态 P开始。通过Unet模型提取图像特征后，我们将它们重塑为K特征图和一个投影特征图。每个高斯点 GP_i然后从这些图中自适应地采样特征，捕捉动态外观特征 df_i。这些特征通过融合网络与内在外观特征 sf_i 融合，解码为高斯点颜色 c_i。最后，所有高斯点使用瓦片光栅化器进行渲染。

其中 px' 表示像素的位置， $GP_{px'}$ 表示与该像素关联的排序后的高斯点。最终渲染的图像随后用于与参考图像计算损失以进行训练，联合优化所有高斯属性。此外，它还设计了一种基于梯度和不透明度的点增长和修剪策略。

4 方法

基于之前的分析，上述基于NeRF的方法 [6,23,50]缺乏对外观中高频和局部详细信息的足够关注，同时由于采样点数量众多，渲染成本显著。为了解决这些问题，我们在第4.1节中利用3D高斯点显式建模场景，并在第4.2节中引入了一种新的高斯点外观建模方法。内在和动态外观特征在第4.3节中被分离然后融合。此外，在第4.5节计算损失时，使用可见性图来减少第4.4节中瞬态物体的影响。整个流程如图2所示。

4.1 3D Gaussian Splatting in the Wild

Appearance. Since 3DGS [16] is designed for reconstructing static scenes, GS-W abandons the conventional color modeling approach using spherical harmonic coefficients. Instead, in the subsequent section, we introduce a new appearance feature af_i for each Gaussian point, adapting to variations in the reference image by fusing intrinsic appearance feature sf_i with dynamic appearance feature df_i extracted from the image.

Transient object. Handling transient objects is also challenging for 3DGS, in which Gaussian points around transient object regions may receive gradients and move or grow, bringing the emergence of meaningless floating points and rendering artifacts. Therefore, a visibility map is employed to mitigate this issue in Sec. 4.4. In addition, we maintain the pruning and growing strategies for points to reconstruct buildings that are absent in the initial point cloud.

4.2 Dynamic appearance features modeling

In the real physical world, even within the same scene, most object points experience varying environmental influences, such as light from different directions. Combined with unique intrinsic material properties, each point displays a different color, gloss, and texture. NeRF-W [23] and Ha-NeRF [6] employ a global feature embedding from the reference image for the entire scene's appearance, assuming uniform environmental information across all points. CR-NeRF [50] integrates reference image features into rendered image features at the 2D level, lacking dynamic environmental information in 3D. These methods tend to roughly restore the global color tone of the scene and struggle to capture local details such as highlight and shadow. To address this, we introduce diverse information for each point, better aligning with real-world scenarios.

Projection feature map. We extract 2D features from the image I_{gt} and then map each point to the 2D space for feature sampling. With the known camera pose of the image, the mapping relationship from 3D points to the 2D image can be determined. Therefore, from the image we extract a projection feature map F^P, onto which each 3D point is projected using a projection matrix P, followed by the bilinear interpolation feature sampling, as illustrated below:

$$f_i^P = BL(\Pi_P(X_i), F^P) \tag{6}$$

where X_i represents the position coordinate of the i-th Gaussian point, while $f_i^P \in \mathbb{R}^{16}$ denotes the features sampled from the projection feature map for this Gaussian point, constituting a portion of its dynamic appearance feature. The term BL indicates the bilinear interpolation sampling.

By employing this method, Gaussian points along different rays can effectively capture features at their corresponding positions on the reference image.

K feature maps. Limited by the single-view reference image I_{gt}, features sampled from the projection feature map are identical along the same ray, which is unreasonable under uneven lighting. Besides, since the reference image sometimes only contains part of the scene, many points may not be projected into

4.1 三维高斯溅射在野外

外观。 由于3DGS [16] 旨在重建静态场景，GS-W放弃了使用球谐系数的传统颜色建模方法。相反，在后续章节中，我们为每个高斯点引入了一种新的外观特征 af_i ，通过将内在外观特征 sf_i 与从图像中提取的动态外观特征 df_i融合，以适应参考图像的变化。

瞬态物体。 处理瞬态物体对3DGS也是一个挑战，其中瞬态物体区域周围的高斯点可能会接收梯度并移动或增长，导致出现无意义的浮动点和渲染伪影。因此，在第4.4节中使用可见性图来缓解这个问题。此外，我们保持了点的修剪和生长策略，以重建初始点云中缺失的建筑物。

4.2 动态外观特征建模

在真实的物理世界中，即使在同一场景内，大多数物体点也会经历不同的环境影响，例如来自不同方向的光线。结合独特的内在材料属性，每个点显示不同的颜色、光泽和纹理。NeRF-W [23]和Ha-NeRF [6] 使用参考图像的全局特征嵌入来表示整个场景的外观，假设所有点的环境信息均匀。CR-NeRF [50] 在2D级别将参考图像特征集成到渲染图像特征中，缺乏3D中的动态环境信息。这些方法倾向于粗略恢复场景的全局色调，难以捕捉高光和阴影等局部细节。为了解决这个问题，我们为每个点引入了多样的信息，更好地与现实世界场景对齐。

投影特征图。我们从图像 I_{gt} 中提取2D特征，然后将每个点映射到2D空间进行特征采样。通过已知的图像相机姿态，可以确定从3D点到2D图像的映射关系。因此，从图像中我们提取一个投影特征图 F^P，每个3D点使用投影矩阵 P 投影到该图上，然后进行双线性插值特征采样，如下所示：

$$f_i^P = BL(\Pi_P(X_i), F^P) \tag{6}$$

其中 X_i 表示第i个高斯点的位置坐标，而 $f_i^P \in \mathbb{R}^{16}$ 表示从投影特征图中为该高斯点采样的特征，构成其动态外观特征的一部分。术语 BL 表示双线性插值采样。

通过使用这种方法，不同射线上的高斯点可以有效地捕获参考图像上相应位置的特征。

K特征图。 由于单视图参考图像 I_{gt}的限制，从投影特征图中采样的特征在同一条射线上是相同的，这在不均匀照明下是不合理的。此外，由于参考图像有时只包含场景的一部分，许多点可能无法投影到

effective regions for obtaining valid feature samples, causing inconsistencies in new viewpoints. To address these issues and grant Gaussian points the freedom to focus on diverse information, we propose extracting additional K feature maps $(F^1, F^2...F^K)$ from I_{gt} to construct a high-dimensional sampling space. Each Gaussian point is mapped to these maps respectively for adaptive sampling, allowing them to focus on high-frequency features.

Adaptive sampling. It is necessary to map the Gaussian points to different positions on K feature maps, to better focus on diverse information. Inspired by the motive of each Gaussian point independently learning its own attributes, we consider it an efficient way to allow every point to determine its sampling positions through self-learning. Therefore, we assign each Gaussian point with K learnable sampling coordinate attributes $(sc_i^1, sc_i^2...sc_i^K)$, enabling them to adaptively select the information they need to focus on. The sampling process using these K sampling coordinates in K feature maps is as follows:

$$(f_i^1, f_i^2...f_i^K) = BL((sc_i^1, sc_i^2...sc_i^K), (F^1, F^2...F^K)) \tag{7}$$

where $(f_i^1, f_i^2...f_i^K) \in \mathbb{R}^{K\times 16}$ represents the features sampled from K feature maps for the i-th Gaussian point. Next, we concatenate the sampled features with f_i^P to jointly represent the dynamic appearance feature of the Gaussian point as follows:

$$df_i = f_i^P \oplus f_i^1 \oplus f_i^2 \oplus ...f_i^K \tag{8}$$

During training, to prevent sampling coordinates from deviating beyond the effective sampling range, we introduce a regularization term:

$$L_{sc} = \frac{1}{N}\sum_i^N max\{0, |sc_i| - 1\} \tag{9}$$

where N represents the total number of Gaussian points, and $|.|$ denotes the absolute value.

Feature maps extraction. We utilize a Unet [35] model with ResNet [13] backbone to generate both the projection feature map and K feature maps. The Unet model takes the reference image $I_{gt} \in \mathbb{R}^{3\times H\times W}$ as input and produces a 2D feature map $F \in \mathbb{R}^{16(K+1)\times H\times W}$ of the same spatial size. Subsequently, this feature map is evenly divided along the channel dimension into $(K+1)$ feature maps, each serving as one of the K feature maps $(F^1, F^2...F^K)$, and the projection feature map F^P. We choose Unet as the feature extractor due to its simplicity and effectiveness in extracting features from images at the same scale.

4.3 Intrinsic and dynamic appearance

Separation of intrinsic and dynamic appearance. As mentioned before, an object's appearance is influenced by both its intrinsic material as well as surface properties, and dynamic environmental factors. However, previous methods like Ha-NeRF primarily rely on image features, positional data, and viewing direction to decode appearance using MLPs. This implicit modeling of the scene's

有效区域以获取有效的特征样本，导致新视点的不一致性。为了解决这些问题并赋予高斯点关注多样信息的自由，我们提出从 I_{gt} 中提取额外的 K 特征图（F^1，$F^2...F^K$）以构建高维采样空间。每个高斯点分别映射到这些图进行自适应采样，使它们能够关注高频特征。

自适应采样。有必要将高斯点映射到K特征图上的不同位置，以更好地关注多样化信息。受到每个高斯点独立学习其自身属性的启发，我们认为允许每个点通过自学习确定其采样位置是一种有效的方法。因此，我们为每个高斯点分配K个可学习的采样坐标属性（sc_i^1，$sc_i^2...sc_i^K$），使它们能够自适应地选择需要关注的信息。使用这些K个采样坐标在K特征图中的采样过程如下：

$$(f_i^1, f_i^2...f_i^K) = BL((sc_i^1, sc_i^2...sc_i^K), (F^1, F^2...F^K)) \tag{7}$$

其中（f_i^1，$f_i^2...f_i^K$）$\in \mathbb{R}^{K\times 16}$表示从K特征图中为第i个高斯点采样的特征。接下来，我们将采样的特征与 f_i^P拼接，共同表示高斯点的动态外观特征，如下所示：

$$df_i = f_i^P \oplus f_i^1 \oplus f_i^2 \oplus ...f_i^K \tag{8}$$

在训练过程中，为了防止采样坐标偏离有效采样范围，我们引入了一个正则化项：

$$L_{sc} = \frac{1}{N}\sum_i^N max\{0, |sc_i| - 1\} \tag{9}$$

其中N表示高斯点的总数，$|.|$ 表示绝对值。

特征图提取。我们使用一个带有ResNet [13]主干的Unet [35] 模型来生成投影特征图和K特征图。Unet模型以参考图像 $I_{gt} \in \mathbb{R}^{3\times H\times W}$ 作为输入，并生成与相同空间大小的2D特征图 $F \in \mathbb{R}^{16(K+1)\times H\times W}$。随后，该特征图沿通道维度均匀分为$(K+1)$个特征图，每个特征图作为K特征图之一$(F^1, F^2...F^K)$，以及投影特征图 F^P。我们选择Unet作为特征提取器，因为它在从相同尺度的图像中提取特征时简单且有效。

4.3 固有和动态外观

内在和动态外观的分离。如前所述，物体的外观受到其内在材料和表面属性以及动态环境因素的影响。然而，之前的方法如Ha-NeRF主要依赖图像特征、位置数据和视图方向，使用MLPs解码外观。这种通过MLPs对场景的

intrinsic attributes through MLPs makes it difficult to express the high-frequency features solely with small MLPs, thus relying on the extracted dynamic features to capture comprehensive information. Consequently, such blurring of features hinders the model's ability to distinguish between intrinsic and dynamic appearance accurately, particularly in scenarios involving changes in illumination and weather conditions. To address this, we explicitly separate the scene appearance into two forms: intrinsic and dynamic appearance features.

Similar to modeling the static Gaussian points' positions, we assign a new learnable intrinsic appearance attribute sf_i to each Gaussian point. Meanwhile, the dynamic appearance features df_i are obtained by features extracted from the reference image.

Fusion of intrinsic and dynamic appearance features. After acquiring independent dynamic appearance features for each Gaussian point through adaptive sampling, it's essential to combine them with the corresponding intrinsic appearance features to generate a comprehensive appearance feature af_i. To accomplish this, we design a fusion network M_f composed of two MLPs, which takes two appearance features and position information as inputs and produces a holistic appearance feature influenced by both. Specifically:

$$af_i = M_f(sf_i, df_i, X_i) \tag{10}$$

When rendering images, the fused appearance feature af_i, along with the view direction ϕ, are jointly decoded by an MLP M_c to obtain the color c_i of the i-th Gaussian point, as in Eq. (11). Finally, the Gaussian points are rendered using a differentiable tile rasterizer and colors are aggregated according to Eq. (5), producing image I_r with the appearance features of the reference image.

$$c_i = M_c(af_i, \phi) \tag{11}$$

4.4 Transient objects handling

To mitigate the impact of transient objects and prevent the occurrence of artifacts like floating points, we employ a 2D visibility map $VM \in \mathbb{R}^{1\times H\times W}$ obtained from a Unet model, facilitating accurate segmentation between transient and static objects. Leveraging the visibility map, we weight the loss calculation between the reference image I_{gt} and the rendered image I_r, as Eq. (13). During training, the model often struggles to reconstruct the geometry and appearance of transient objects, leading to larger losses in regions containing such objects. Therefore, in unsupervised scenarios, the 2D visibility map tends to decrease the visibility of transient objects to minimize the training loss. As a result, regions with higher visibility receive more emphasis, while those with lower visibility are disregarded. Additionally, to prevent the visibility map from marking all pixels invisible, we introduce a regularization loss term for the visibility map:

$$L_{vm} = L_2(VM, 1) \tag{12}$$

内在属性的隐式建模使得仅用小MLPs难以表达高频特征，因此依赖提取的动态特征来捕捉全面信息。因此，这种特征的模糊性阻碍了模型准确区分内在和动态外观的能力，特别是在涉及光照和天气条件变化的场景中。为了解决这个问题，我们明确地将场景外观分为两种形式：内在和动态外观特征。

类似于建模静态高斯点的位置，我们为每个高斯点分配一个新的可学习的内在外观属性 sf_i。同时，动态外观特征 df_i 通过从参考图像中提取的特征获得。

内在和动态外观特征的融合。 在通过自适应采样为每个高斯点获取独立的动态外观特征后，必须将它们与相应的内在外观特征结合起来，以生成全面的外观特征 af_i。为此，我们设计了一个由两个 MLP 组成的融合网络 M_f，该网络以两个外观特征和位置信息作为输入，并生成受两者影响的整体外观特征。具体来说：

$$af_i = M_f(sf_i, df_i, X_i) \tag{10}$$

在渲染图像时，融合的外观特征 af_i 与视图方向 ϕ 一起由 MLP M_c 解码，以获得第 i 个高斯点的颜色 c_i，如方程 (11) 所示。最后，使用可微分的瓦片光栅化器渲染高斯点，并根据方程 (5) 聚合颜色，生成具有参考图像外观特征的图像 I_r。

$$c_i = M_c(af_i, \phi) \tag{11}$$

4.4 瞬态物体处理

为了减轻瞬态物体的影响并防止出现浮动点等伪影，我们使用从Unet模型获得的二维可见性图 $VM \in \mathbb{R}^{1\times H\times W}$，以促进瞬态和静态物体之间的准确分割。利用可见性图，我们对参考图像 I_{gt} 和渲染图像 I_r之间的损失计算进行加权，如公式（13）所示。在训练过程中，模型通常难以重建瞬态物体的几何和外观，导致包含此类物体的区域损失较大。因此，在无监督场景中，二维可见性图倾向于降低瞬态物体的可见性以最小化训练损失。结果，可见性较高的区域得到更多重视，而可见性较低的区域被忽略。此外，为了防止可见性图将所有像素标记为不可见，我们引入了可见性图的正则化损失项：

$$L_{vm} = L_2(VM, 1) \tag{12}$$

Table 1: Quantitative results on the test set of three PhotoTourism scenes. The **bold** and the underline represent the best and second-best results, respectively. Our method outperforms the previous methods across all scenes on PSNR, SSIM, and LPIPS.

	Brandenburg Gate			Sacre Coeur			Trevi Fountain		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS	19.33	0.8838	0.1317	17.70	0.8454	0.1761	17.08	0.7139	0.2413
NeRF-W	24.17	0.8905	0.1670	19.20	0.8076	0.1915	18.97	0.6984	0.2652
Ha-NeRF	24.04	0.8873	0.1391	20.02	0.8012	0.1710	20.18	0.6908	0.2225
CR-NeRF	26.53	0.9003	0.1060	22.07	0.8233	0.1520	21.48	0.7117	0.2069
GS-W (Ours)	**27.96**	**0.9319**	**0.0862**	**23.24**	**0.8632**	**0.1300**	**22.91**	**0.8014**	**0.1563**

4.5 Optimization

Similar to [16], we apply two types of the loss function, L_1 and L_{SSIM} [44], to calculate the pixel error between the rendered image I_r and the reference image I_{gt}. Differently, a visibility map VM is incorporated to guide the supervision of rendered images by reference images. Moreover, we introduce the perceptual loss L_{LPIPS} [56]. Therefore, the overall image loss function is represented as in Eq. (13), where $\odot$ denotes the Hadamard product. Combining the regularization loss terms mentioned in Eq. (9) and Eq. (12), the total loss function is formulated as in Eq. (14), where λ_1, λ_{SSIM}, λ_{LPIPS}, λ_{sc} and λ_{vm} are 0.8, 0.2, 0.005, 0.001, 0.15, respectively.

$$L_c = \lambda_1 L_1(VM \odot I_r, VM \odot I_{gt}) + \lambda_{SSIM} L_{SSIM}(VM \odot I_r, VM \odot I_{gt}) + \lambda_{LPIPS} L_{LPIPS}(I_r, I_{gt}) \tag{13}$$

$$L = L_c + \lambda_{sc} L_{sc} + \lambda_{vm} L_{vm} \tag{14}$$

5 Experiments

5.1 Implementation details

We implement our method using Pytorch [30] and train our networks with Adam optimizer [17]. We set $K = 3$ in our experiment as the increasing K brings no performance improvement with meaningless computational cost. We train the full model on a single Nvidia RTX 3090 GPU for $70k$ steps and downsample all the images 2 times during training and evaluation, which takes approximately 2 hours. We also perform the adaptive control of the 3D Gaussians and follow other hyperparameter settings similar to 3DGS [16].

表 1: 在三个 PhotoTourism 场景的测试集上的定量结果。**粗体** 和下划线分别表示最佳和次佳结果。我们的方法在所有场景的 PSNR、SSIM 和 LPIPS 上均优于先前的方法。

	勃兰登堡门			圣心大教堂			特雷维喷泉		
	峰值信噪比↑	SSIM↑	LPIPS↓	峰值信噪比↑	SSIM↑	LPIPS↓	峰值信噪比↑	SSIM↑	LPIPS↓
3DGS	19.33	0.8838	0.1317	17.70	0.8454	0.1761	17.08	0.7139	0.2413
NeRF-W	24.17	0.8905	0.1670	19.20	0.8076	0.1915	18.97	0.6984	0.2652
Ha-NeRF	24.04	0.8873	0.1391	20.02	0.8012	0.1710	20.18	0.6908	0.2225
CR-NeRF	26.53	0.9003	0.1060	22.07	0.8233	0.1520	21.48	0.7117	0.2069
GS-W（我们的）	**27.96**	**0.9319**	**0.0862**	**23.24**	**0.8632**	**0.1300**	**22.91**	**0.8014**	**0.1563**

4.5 优化

类似于 [16], ，我们应用两种类型的损失函数， L_1 和 L_{SSIM} [44]，以计算渲染图像 I_r 和参考图像I_{gt}之间的像素误差。不同之处在于，引入了可见性图 VM 以指导参考图像对渲染图像的监督。此外，我们引入了感知损失 L_{LPIPS} [56]。因此，整体图像损失函数表示为公式（13），其中 $\odot$ 表示哈达玛积。结合公式（9）和公式（12）中提到的正则化损失项，总损失函数表示为公式（14），其中 λ_1、 λ_{SSIM}、 λ_{LPIPS}、 λ_{sc}和 λ_{vm}分别为0.8、0.2、0.005、0.001、0.15。

$$L_c = \lambda_1 L_1(VM \odot I_r, VM \odot I_{gt}) + \lambda_{SSIM} L_{SSIM}(VM \odot I_r, VM \odot I_{gt}) + \lambda_{LPIPS} L_{LPIPS}(I_r, I_{gt}) \tag{13}$$

$$L = L_c + \lambda_{sc} L_{sc} + \lambda_{vm} L_{vm} \tag{14}$$

5 实验

5.1 实现细节

我们使用Pytorch [30] 实现我们的方法，并使用Adam优化器 [17]训练我们的网络。在实验中，我们设置 $K = 3$ ，因为增加K不会带来性能提升，反而会增加无意义的计算成本。我们在单个Nvidia RTX 3090 GPU上训练完整模型，共进行 $70k$ 步，并在训练和评估期间将所有图像下采样2倍，大约需要2小时。我们还对3D高斯进行自适应控制，并遵循与3DGS [16]类似的其他超参数设置。

Table 2: Comparison of rendering speed with one RTX 3090 GPU on three scenes with a resolution of 800 × 800, measured by FPS(Frames Per Second). Ours-cache means the appearance feature is cached for each point when synthesizing novel views.

	Brandenburg Gate	Sacre Coeur	Trevi Fountain
3DGS	221	268	198
NeRF-W	0.0518	0.0514	0.0485
Ha-NeRF	0.0489	0.0497	0.0498
CR-NeRF	0.0445	0.0447	0.0446
Ours	55.8	58.3	38
Ours-cache	221	301	197

5.2 Evaluation

Dataset, metrics, baseline. We evaluate our proposed method on three scenes from the PhotoTourism dataset: Brandenburg Gate, Sacre Coeur, and Trevi Fountain, which include varying appearances and transient objects. For quantitative comparison, we use PSNR, SSIM [44], and LPIPS [56] as metrics to assess the performance of our method. We also present rendered images generated from the same pose as the input view for visual inspection. To demonstrate the superiority of our method, we evaluate our proposed method against 3DGS [16], NeRF-W [23], Ha-NeRF [6], and CR-NeRF [50].

Quantitative comparison. Quantitative results are shown in Tab. 1. 3DGS performs poorly on both PSNR and SSIM metrics, as it does not explicitly model the appearance variation and transient objects. NeRF-W and Ha-NeRF achieve moderate performance by introducing a global appearance embedding and anti-transient module. It's worth noting that NeRF-W needs to optimize the appearance embedding of test images, thus the comparison with NeRF-W is unfair. CR-NeRF achieves competitive performance due to cross-ray manner. Utilizing an adaptive sampling strategy that allows each point to focus on local details, our method outperforms the baselines on three scenes in terms of PSNR, SSIM, and LPIPS, which verifies that we can capture more details and render higher-quality images.

Render speed. To compare the render speed of different methods during inference, we experiment on the three scenes by setting the image resolution to 800 × 800 and calculating the average rendering time per image using a single RTX 3090 GPU. The time taken for feature extraction from the reference images in Ha-NeRF, CR-NeRF, and our method is included in the overall inference time. As shown in Tab. 2, our method achieves a significant improvement in rendering speed, which is 1000× faster than previous NeRF-based methods.

Since our method only requires one feature extraction step and can cache appearance feature af_i for each Gaussian point, only a small MLP decoder M_c is needed for color decoding when synthesizing a novel view. This enables GS-W to achieve a 200 FPS rendering speed, comparable to 3DGS.

表 2: 使用一个 RTX 3090 GPU 在三个分辨率为 800 × 800 的场景上比较渲染速度，以 FPS（每秒帧数）测量。Ours-cache 表示在合成新视图时为每个点缓存外观特征。

	勃兰登堡门	圣心大教堂	特雷维喷泉
3DGS	221	268	198
NeRF-W	0.0518	0.0514	0.0485
Ha-NeRF	0.0489	0.0497	0.0498
CR-NeRF	0.0445	0.0447	0.0446
Ours	55.8	58.3	38
Ours-cache	221	301	197

5.2 评估

数据集、指标、基线。 我们在PhotoTourism数据集的三个场景上评估我们提出的方法：勃兰登堡门、圣心大教堂和特雷维喷泉，这些场景包括不同的外观和瞬态物体。为了进行定量比较，我们使用PSNR、SSIM [44], 和LPIPS [56] 作为指标来评估我们方法的性能。我们还展示了从与输入视图相同姿态生成的渲染图像，以便进行视觉检查。为了展示我们方法的优越性，我们评估了我们提出的方法与3DGS [16],NeRF-W [23], Ha-NeRF [6], 和CR-NeRF [50]的对比。

定量比较。 定量结果如表1所示。3DGS在PSNR和SSIM指标上表现不佳，因为它没有显式建模外观变化和瞬态物体。NeRF-W和Ha-NeRF通过引入全局外观嵌入和抗瞬态模块实现了中等性能。值得注意的是，NeRF-W需要优化测试图像的外观嵌入，因此与NeRF-W的比较是不公平的。CR-NeRF由于跨射线方式实现了竞争性能。利用允许每个点关注局部细节的自适应采样策略，我们的方法在三个场景的PSNR、SSIM和LPIPS方面优于基线，这验证了我们可以捕捉更多细节并渲染更高品质的图像。

渲染速度。 为了比较不同方法在推理过程中的渲染速度，我们在三个场景上进行实验，将图像分辨率设置为800 × 800 ，并使用单个RTX 3090 GPU计算每张图像的平均渲染时间。Ha-NeRF、CR-NeRF和我们的方法中从参考图像提取特征的时间包含在整体推理时间中。如表2所示，我们的方法在渲染速度上实现了显著提升，比之前的基于NeRF的方法快 1000× 。

由于我们的方法只需要一个特征提取步骤，并且可以为每个高斯点缓存外观特征 af_i ，因此在合成新视图时，只需要一个小的MLP解码器 M_c进行颜色解码。这使得GS-W能够达到200 FPS的渲染速度，与3DGS相当。

Fig. 3: Qualitative results on the test set of three PhotoTourism scenes. GS-W recovers finer details of appearance(*e.g.* the horse sculpture in Brandenburg, the sky and clouds in Sacre, the light on columns, and the color of windows in Trevi). Moreover, GS-W reconstructs more consistent and detailed scenes (*e.g.* the distant tower in Brandenburg, the cavities in Sacre, and the distant building in Trevi).

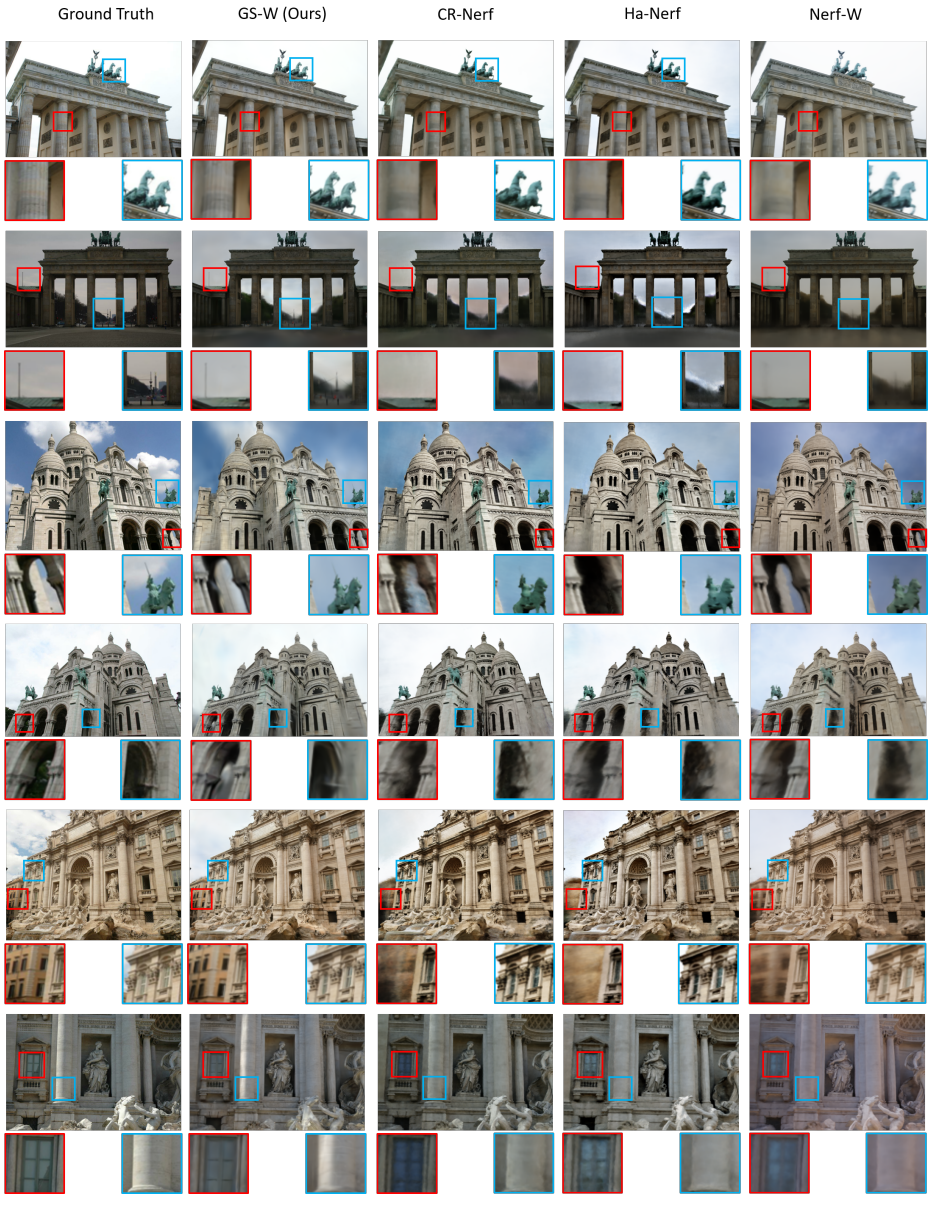

图3: 在三个PhotoTourism场景的测试集上的定性结果。GS-W恢复了更精细的外观细节(例如勃兰登堡的马雕塑，圣心的天空和云，特雷维的柱子上的光和窗户的颜色)。此外，GS-W重建了更一致和详细的场景(例如勃兰登堡的远处的塔，圣心的空腔，和特雷维的远处的建筑)。

Table 3: Ablation studies on three scenes. The **bold** and the underline represent the best and second-best results, respectively. See Sec. 5.3 for detailed descriptions.

	Brandenburg Gate			Sacre Coeur			Trevi Fountain		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o visibility map	**28.64**	**0.9324**	0.0867	**23.76**	**0.8723**	**0.1234**	23.03	0.8025	**0.1532**
w/o K feature maps	26.37	0.9210	0.1018	21.59	0.8521	0.1417	22.27	0.7859	0.1700
w/o Projection map	26.61	0.9285	0.0955	22.99	0.8635	0.1320	22.33	0.7980	0.1580
w/o adaptive sampling	26.44	0.9219	0.0961	21.45	0.8479	0.1439	22.18	0.7866	0.1674
w/o separation	28.22	0.9291	0.0942	22.94	0.8520	0.1469	**23.27**	0.7907	0.1714
full (Ours)	27.96	0.9319	**0.0862**	23.24	0.8632	0.1300	22.91	**0.8014**	0.1563

Qualitative comparison. Fig. 3 presents the qualitative result for all methods. NeRF-W and Ha-NeRF can model varying appearances from the reference image by introducing a global appearance embedding. CR-NeRF can reconstruct better geometry and model appearance variation compared with Ha-NeRF and NeRF-W. However, they all struggle to reconstruct the details of scenes in the distance and the intricate textures of the scenes, *e.g.* the door pillars and distant tower in Brandenburg, the cavity in Sacre, and the distant buildings in Trevi. In contrast, thanks to the higher-frequency dynamic appearance features, our method recovers more accurate appearance details, *e.g.* the horse sculpture in Brandenburg, the sky and clouds in Sacre, and the light on columns in Trevi.

5.3 Ablation studies

We summarize the ablation studies of our method on Brandenburg, Sacre, and Trevi datasets in Tab. 3 and produce qualitative results in Fig. 4 and Fig. 5 to validate the effectiveness of each component.

Without the visibility map. Removing the transient object handling module leads to higher metric performances but introduces artifacts in the rendered image due to the influence of transient objects, as depicted in Fig. 4. Since most test images lack dynamic objects, these artifacts may not significantly impact the metrics.

Without K feature maps or projection feature map. The removal of K feature maps or the projection feature map results in performance degradation. Especially, synthesizing novel views without the K feature maps not only reduces the ability to capture information from the reference image but also produces view-inconsistent appearances, as shown in Fig. 4.

Without adaptive sampling. Retaining K feature maps while keeping the sampling coordinates fixed for each point significantly decreases performance. This highlights the importance of our adaptive sampling strategy for K feature maps, enabling Gaussian points to adaptively focus on local and detailed features.

Without separation. We eliminate the intrinsic feature and predict the color of Gaussian points by the dynamic appearance feature only. The results in Tab. 3 show a noticeable decrease in both LPIPS and SSIM, which are more in

表3: 在三个场景上的消融研究。**粗体** 和下划线分别表示最佳和次佳结果。详见第5.3节的详细描述。

	勃兰登堡门			圣心大教堂			特雷维喷泉		
	峰值信噪比↑	SSIM↑	LPIPS↓	峰值信噪比↑	SSIM↑	LPIPS↓	峰值信噪比↑	SSIM↑	LPIPS↓
无可见性图	**28.64**	**0.9324**	0.0867	**23.76**	**0.8723**	**0.1234**	23.03	0.8025	**0.1532**
无K特征图	26.37	0.9210	0.1018	21.59	0.8521	0.1417	22.27	0.7859	0.1700
w/o 投影图	26.61	0.9285	0.0955	22.99	0.8635	0.1320	22.33	0.7980	0.1580
无自适应采样	26.44	0.9219	0.0961	21.45	0.8479	0.1439	22.18	0.7866	0.1674
无分离	28.22	0.9291	0.0942	22.94	0.8520	0.1469	**23.27**	0.7907	0.1714
full (Ours)	27.96	0.9319	**0.0862**	23.24	0.8632	0.1300	22.91	**0.8014**	0.1563

定性比较。 图3展示了所有方法的定性结果。NeRF-W和Ha-NeRF通过引入全局外观嵌入，可以建模参考图像中变化的外观。CR-NeRF可以重建比Ha-NeRF和NeRF-W更好的几何结构并建模外观变化。然而，它们在重建远处场景的细节和场景的复杂纹理方面都存在困难，例如勃兰登堡的门柱和远处的塔，圣心的空腔，以及特雷维的远处建筑。相比之下，由于高频动态外观特征，我们的方法恢复了更准确的外观细节，例如勃兰登堡的马雕塑，圣心的天空和云，以及特雷维柱子上的光。

5.3 消融研究

我们在勃兰登堡、圣心和特雷维数据集上总结了我们方法的消融研究，见表3，并在图4和图5中生成定性结果，以验证每个组件的有效性。

没有可见性图。移除瞬态物体处理模块会导致更高的指标性能，但由于瞬态物体的影响，在渲染图像中引入了伪影，如图4所示。由于大多数测试图像缺乏动态对象，这些伪影可能不会显著影响指标。

没有K特征图或投影特征图。移除K特征图或投影特征图会导致性能下降。特别是，没有K特征图合成新视图不仅降低了从参考图像捕获信息的能力，还产生了视图不一致的外观，如图4所示。

没有自适应采样。保留K特征图，同时为每个点保持采样坐标固定，会显著降低性能。这突显了我们自适应采样策略对K特征图的重要性，使高斯点能够自适应地关注局部和详细特征。

Without separation. 我们消除内在特征，仅通过动态外观特征预测高斯点的颜色。表3中的结果显示LPIPS和SSIM均有明显下降，这更符合

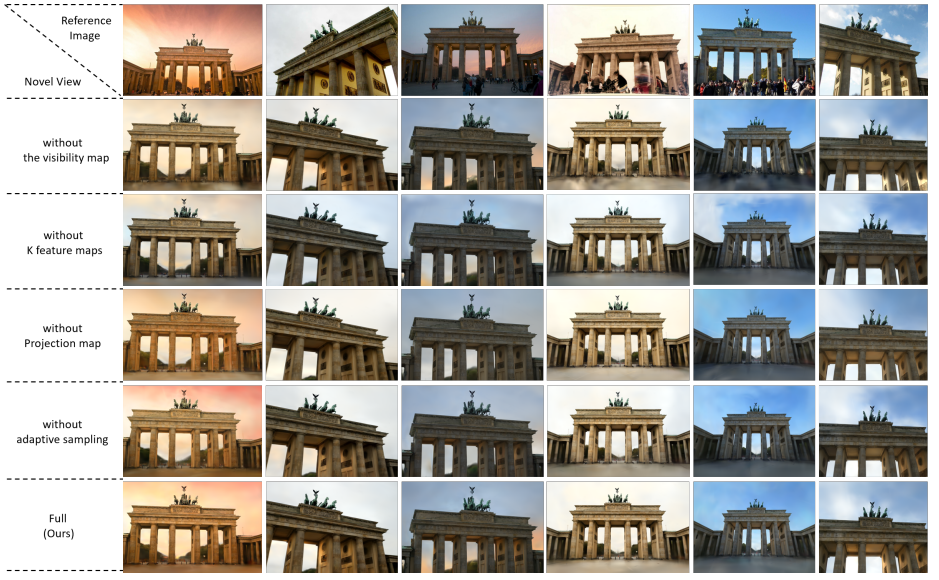

Fig. 4: Ablation studies by visualization. The first row represents the reference images and the corresponding rows represent the rendered images from a novel view. Our full method is capable of performing view-consistent appearance and reducing artifacts.

line with human visual perception. In Fig. 5, the absence of dynamic appearance features results in incomplete scene appearance. This emphasizes the significance of intrinsic feature in retaining essential scene characteristics. Both illustrate that separation is beneficial for the model to accurately comprehend and reconstruct sharp appearances.

5.4 Appearance tuning experiment

Since we explicitly model the scene's appearance as unchanged intrinsic and varying dynamic features, we can adjust the impact of dynamic appearance features on intrinsic appearance by multiplying df_i by a proportional weight. We also compare it with Ha-NeRF and CR-NeRF by applying the same weights to extracted image features. Qualitative results are shown in Fig. 5. Ha-NeRF and CR-NeRF show odd coloration in the sky and buildings at low weights, while at high weights, they fail to capture the detailed highlights and enhance the style of sufficient illumination, resulting in darker building colors. In contrast, our method aligns more closely with human understanding of the physical world. As the weight increases from small to large, our method gradually applies extracted environmental impacts to the scene, *e.g.* appearing highlights on pillars and enhancing illumination. This demonstrates the importance of dynamic features in capturing environmental information and explicitly separating intrinsic and dynamic appearances aid the model in clearly learning and distinguishing between the two, thereby achieving more flexible tuning over the scene's appearance.

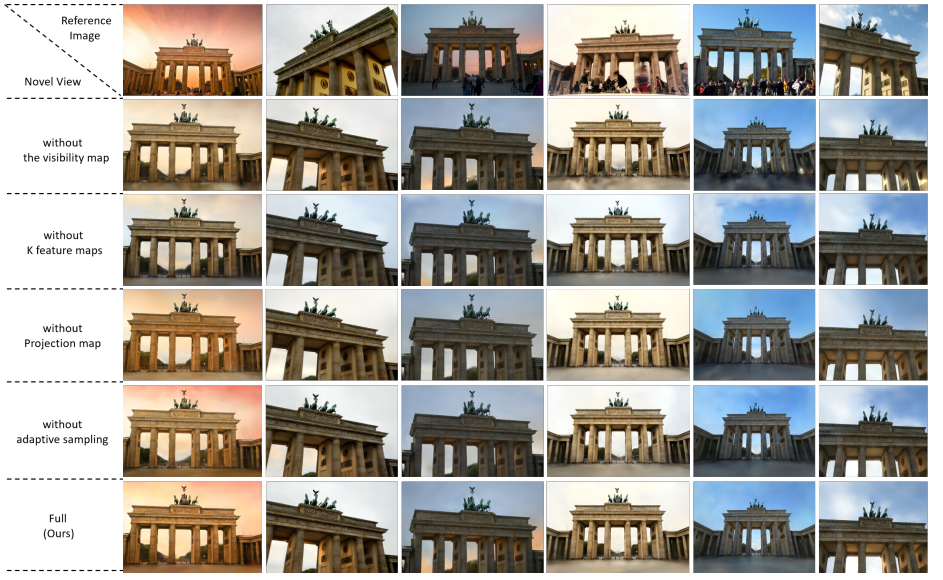

图4: 通过可视化进行的消融研究。第一行表示参考图像，对应的行表示从新视角渲染的图像。我们的完整方法能够实现视图一致的外观并减少伪影。

人类视觉感知。在图5中，缺少动态外观特征导致场景外观不完整。这强调了内在特征在保留基本场景特征方面的重要性。两者都表明分离有助于模型准确理解和重建清晰的外观。

5.4 外观调优实验

由于我们明确地将场景的外观建模为不变的内在特征和变化的动态特征，我们可以通过将 df_i 乘以一个比例权重来调整动态外观特征对内在外观的影响。我们还通过将相同的权重应用于提取的图像特征，将其与Ha-NeRF和CR-NeRF进行比较。定性结果如图5所示。Ha-NeRF和CR-NeRF在低权重时在天空和建筑物中显示出异常的着色，而在高权重时，它们无法捕捉详细的高光并增强充足照明的风格，导致建筑物颜色变暗。相比之下，我们的方法更符合人类对物理世界的理解。随着权重从小到大增加，我们的方法逐渐将提取的环境影响应用于场景，例如柱子上出现的高光和增强的照明。这表明动态特征在捕捉环境信息中的重要性，明确分离内在和动态外观有助于模型清晰地学习和区分两者，从而实现对场景外观的更灵活调优。

Fig. 5: Images are rendered at the same camera pose with increasing weight of features extracted from the image. Our method incorporates environmental factors, like highlights on pillars and enhancing illumination, in a manner that is closer to human understanding.

5.5 Limitations

While GS-W outperforms previous methods, it still has limitations. It struggles with complex lighting variations, specular reflections, and accurately reconstructing textures in frequently occluded scenes, like the floor texture in the Brandenburg Gate scene. Additionally, it assumes known image poses when incorporating appearance information from reference images. Future research may focus on developing new appearance modeling techniques to address these issues.

6 Conclusion

In this paper, we introduce GS-W, a method for reconstructing scenes from unconstrained image collections. Using 3D Gaussian points as 3D representation, we introduce separated intrinsic and dynamic appearance features for each point to effectively model scene appearance. We propose an adaptive sampling strategy to capture local environmental factors like highlights and utilize a visibility map to handle transient objects. Our approach outperforms previous NeRF-based methods by providing better extraction of dynamic environmental impacts from images and addressing slow rendering speeds. Experimental results demonstrate the superiority and efficiency of our method compared to previous approaches.

图5: 图像在相同的相机姿态下渲染，随着从图像中提取的特征权重增加。我们的方法以更接近人类理解的方式纳入了环境因素，如柱子上的高光和增强照明。

5.5 局限性

虽然GS-W优于以前的方法，但它仍然存在局限性。它在处理复杂的光照变化、镜面反射以及准确重建频繁遮挡场景中的纹理（如勃兰登堡门场景中的地板纹理）时遇到困难。此外，它在从参考图像中整合外观信息时假设已知图像姿态。未来的研究可能集中在开发新的外观建模技术以解决这些问题。

6 结论

本文中，我们介绍了GS-W，一种从无约束图像集合中重建场景的方法。使用3D高斯点作为三维表示，我们为每个点引入了分离的内在和动态外观特征，以有效建模场景外观。我们提出了一种自适应采样策略，以捕捉局部环境因素（如高光），并利用可见性图处理瞬态对象。我们的方法通过从图像中更好地提取动态环境影响并解决渲染速度慢的问题，优于之前的基于NeRF的方法。实验结果表明，与之前的方法相比，我们的方法具有优越性和效率。

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
3. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
5. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision. pp. 333–350. Springer (2022)
6. Chen, X., Zhang, Q., Li, X., Chen, Y., Feng, Y., Wang, X., Wang, J.: Hallucinated neural radiance fields in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12943–12952 (2022)
7. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)
8. Engelhardt, A., Raj, A., Boss, M., Zhang, Y., Kar, A., Li, Y., Sun, D., Brualla, R.M., Barron, J.T., Lensch, H., et al.: Shinobi: Shape and illumination using neural object decomposition via brdf optimization in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19636–19646 (2024)
9. Fan, Z., Wang, K., Wen, K., Zhu, Z., Xu, D., Wang, Z.: Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. arXiv preprint arXiv:2311.17245 (2023)
10. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12479–12488 (2023)
11. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5501–5510 (2022)
12. Garbin, S.J., Kowalski, M., Johnson, M., Shotton, J., Valentin, J.: Fastnerf: High-fidelity neural rendering at 200fps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14346–14355 (2021)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 371–386 (2018)

参考文献

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srini- vasan, P.P.: Mip-nerf: 一种用于抗锯齿神经辐射场的多尺度表示。在：IEEE/CVF国际计算机视觉会议论文集。第5855–5864页（2021） 2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip- nerf 360: 无界抗锯齿神经辐射场。在：IEEE/CVF计算机视觉与模式识别会议论文集。第5470–5479页（2022） 3. Cao, A., Johnson, J.: Hexplane: 一种用于动态场景的快速表示。在：IEEE/CVF计算机视觉与模式识别会议论文集。第130–141页（2023） 4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., 等：高效的几何感知3D生成对抗网络。在：IEEE/CVF计算机视觉与模式识别会议论文集。第16123–16133页（2022） 5. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: 张量辐射场。在：欧洲计算机视觉会议。第333–350页。Springer（2022） 6. Chen, X., Zhang, Q., Li, X., Chen, Y., Feng, Y., Wang, X., Wang, J.: 在野外的幻觉神经辐射场。在：IEEE/CVF计算机视觉与模式识别会议论文集。第12943–12952页（2022） 7. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: 深度监督的nerf：更少的视图和更快的训练。在：IEEE/CVF计算机视觉与模式识别会议论文集。第12882–12891页（2022） 8. Engelhardt, A., Raj, A., Boss, M., Zhang, Y., Kar, A., Li, Y., Sun, D., Brualla, R.M., Barron, J.T., Lensch, H., 等：Shinobi: 通过BRDF优化在野外使用神经对象分解进行形状和照明。在：IEEE/CVF计算机视觉与模式识别会议论文集。第19636–19646页（2024） 9. Fan, Z., Wang, K., Wen, K., Zhu, Z., Xu, D., Wang, Z.: Lightgaussian: 无界3D高斯压缩，减少15倍和 200+ fps。arXiv预印本 arXiv:2311.17245（2023） 10. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: 空间、时间和外观中的显式辐射场。在：IEEE/CVF计算机视觉与模式识别会议论文集。第12479–12488页（2023） 11. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: 没有神经网络的辐射场。在：IEEE/CVF计算机视觉与模式识别会议论文集。第5501–5510页（2022） 12. Garbin, S.J., Kowalski, M., Johnson, M., Shotton, J., Valentin, J.: Fastnerf: 200fps的高保真神经渲染。在：IEEE/CVF国际计算机视觉会议论文集。第14346–14355页（2021） 13. He, K., Zhang, X., Ren, S., Sun, J.: 深度残差学习用于图像识别。在：IEEE计算机视觉与模式识别会议论文集。第770–778页（2016） 14. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: 从图像集合中学习类别特定的网格重建。在：欧洲计算机视觉会议（ECCV）论文集。第371–386页（2018）

15. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3907–3916 (2018)
16. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Kuang, Z., Olszewski, K., Chai, M., Huang, Z., Achlioptas, P., Tulyakov, S.: Neroic: Neural rendering of objects from online image collections. ACM Transactions on Graphics (TOG) **41**(4), 1–12 (2022)
19. Li, P., Wang, S., Yang, C., Liu, B., Qiu, W., Wang, H.: Nerf-ms: Neural radiance fields with multi-sequence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18591–18600 (2023)
20. Li, Z., Xian, W., Davis, A., Snavely, N.: Crowdsampling the plenoptic function. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 178–196. Springer (2020)
21. Lin, H., Wang, Q., Cai, R., Peng, S., Averbuch-Elor, H., Zhou, X., Snavely, N.: Neural scene chronology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20752–20761 (2023)
22. Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., Dai, B.: Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. arXiv preprint arXiv:2312.00109 (2023)
23. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7210–7219 (2021)
24. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4460–4470 (2019)
25. Meshry, M., Goldman, D.B., Khamis, S., Hoppe, H., Pandey, R., Snavely, N., Martin-Brualla, R.: Neural rerendering in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6878–6887 (2019)
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
27. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
28. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022)
29. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019)
30. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-

15. Kato, H., Ushiku, Y., Harada, T.: 神经3D网格渲染器。在：会议记录中IEEE计算机视觉与模式识别会议。第3907–3916页（2018）16. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 实时辐射场渲染的3D高斯散射。ACM Transactions on Graphics **42(4)**（2023）17. Kingma, D.P., Ba, J.: Adam: 随机优化方法。arXiv预印本arXiv:1412.6980（2014）18. Kuang, Z., Olszewski, K., Chai, M., Huang, Z., Achlioptas, P., Tulyakov, S.: Neroic: 从在线图像集合中神经渲染对象。ACM Transactions on Graphics (TOG) **41(4)**, 1–12（2022）19. Li, P., Wang, S., Yang, C., Liu, B., Qiu, W., Wang, H.: Nerf-ms: 多序列神经辐射场。在：IEEE/CVF国际计算机视觉会议论文集。第18591–18600页（2023）20. Li, Z., Xian, W., Davis, A., Snavely, N.: 众采光场函数。在：计算机视觉–ECCV 2020: 第16届欧洲会议，英国格拉斯哥，2020年8月23–28日，论文集，第一部分16。第178–196页。Springer（2020）21. Lin, H., Wang, Q., Cai, R., Peng, S., Averbuch-Elor, H., Zhou, X., Snavely, N.: 神经场景时间线。在：IEEE/CVF计算机视觉与模式识别会议论文集。第20752–20761页（2023）22. Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., Dai, B.: Scaffold-gs: 结构化3D高斯用于视图自适应渲染。arXiv预印本arXiv:2312.00109（2023）23. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: 野外的NeRF: 无约束照片集合的神经辐射场。在：IEEE/CVF计算机视觉与模式识别会议论文集。第7210–7219页（2021）24. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: 占用网络: 在函数空间中学习3D重建。在：IEEE/CVF计算机视觉与模式识别会议论文集。第4460–4470页（2019）25. Meshry, M., Goldman, D.B., Khamis, S., Hoppe, H., Pandey, R., Snavely, N., Martin-Brualla, R.: 野外的神经重渲染。在：IEEE/CVF计算机视觉与模式识别会议论文集。第6878–6887页（2019）26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: 将场景表示为神经辐射场以进行视图合成。ACM通讯 **65(1)**, 99–106（2021）27. Müller, T., Evans, A., Schied, C., Keller, A.: 具有多分辨率哈希编码的即时神经图形基元。ACM Transactions on Graphics (ToG) **41(4)**, 1–15（2022）28. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: 从稀疏输入中正则化神经辐射场以进行视图合成。在：IEEE/CVF计算机视觉与模式识别会议论文集。第5480–5490页（2022）29. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: 学习连续符号距离函数以进行形状表示。在：IEEE/CVF计算机视觉与模式识别会议论文集。第165–174页（2019）30. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., 等: Pytorch: 一种命令式风格，高

performance deep learning library. Advances in neural information processing systems **32** (2019)
31. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
32. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
33. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. arXiv preprint arXiv:2312.16084 (2023)
34. Reiser, C., Peng, S., Liao, Y., Geiger, A.: Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14335–14345 (2021)
35. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
36. Rudnev, V., Elgharib, M., Smith, W., Liu, L., Golyanik, V., Theobalt, C.: Nerf for outdoor scene relighting. In: European Conference on Computer Vision. pp. 615–631. Springer (2022)
37. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
38. Schwarz, K., Sauer, A., Niemeyer, M., Liao, Y., Geiger, A.: Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. Advances in Neural Information Processing Systems **35**, 33999–34011 (2022)
39. Shao, R., Zheng, Z., Tu, H., Liu, B., Zhang, H., Liu, Y.: Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16632–16642 (2023)
40. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10529–10538 (2020)
41. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8248–8258 (2022)
42. Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-nerf: Structured view-dependent appearance for neural radiance fields. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5481–5490. IEEE (2022)
43. Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. arXiv preprint arXiv:2303.16196 (2023)
44. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
45. Wen, C., Zhang, Y., Li, Z., Fu, Y.: Pixel2mesh++: Multi-view 3d mesh generation via deformation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1042–1051 (2019)

性能深度学习库。神经信息处理系统进展**32** (2019) 31. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: 深度学习在点集上的3D分类和分割。在：IEEE计算机视觉与模式识别会议论文集。第652–660页 (2017) 32. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: 在度量空间中点集上的深度层次特征学习。神经信息处理系统进展**30** (2017) 33. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3D语言高斯溅射。arXiv预印本 arXiv:2312.16084 (2023) 34. Reiser, C., Peng, S., Liao, Y., Geiger, A.: Kilonerf: 使用数千个小型MLP加速神经辐射场。在：IEEE/CVF国际计算机视觉会议论文集。第14335–14345页 (2021) 35. Ronneberger, O., Fischer, P., Brox, T.: U-net: 用于生物医学图像分割的卷积网络。在：医学图像计算与计算机辅助干预–MICCAI 2015: 第18届国际会议，德国慕尼黑，2015年10月5-9日，会议论文集，第三部分18。第234–241页。Springer (2015) 36. Rudnev, V., Elgharib, M., Smith, W., Liu, L., Golyanik, V., Theobalt, C.: 用于户外场景重新照明的NeRF。在：欧洲计算机视觉会议。第615–631页。Springer (2022) 37. Schonberger, J.L., Frahm, J.M.: 结构从运动的重新审视。在：IEEE计算机视觉与模式识别会议论文集。第4104–4113页 (2016) 38. Schwarz, K., Sauer, A., Niemeyer, M., Liao, Y., Geiger, A.: Voxgraf: 使用稀疏体素网格的快速3D感知图像合成。神经信息处理系统进展**35**，第33999–34011页 (2022) 39. Shao, R., Zheng, Z., Tu, H., Liu, B., Zhang, H., Liu, Y.: Tensor4d: 高保真动态重建和渲染的高效神经4D分解。在：IEEE/CVF计算机视觉与模式识别会议论文集。第16632–16642页 (2023) 40. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: 用于3D对象检测的点-体素特征集抽象。在：IEEE/CVF计算机视觉与模式识别会议论文集。第10529–10538页 (2020) 41. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: 可扩展的大场景神经视图合成。在：IEEE/CVF计算机视觉与模式识别会议论文集。第8248–8258页 (2022) 42. Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-nerf: 神经辐射场的结构化视图依赖外观。在：2022 IEEE/CVF计算机视觉与模式识别会议(CVPR)。第5481–5490页。IEEE (2022) 43. Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: 提取深度排名以进行少量新视角合成。arXiv预印本 arXiv:2303.16196 (2023) 44. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: 图像质量评估：从错误可见性到结构相似性。IEEE图像处理交易**13**(4)，第600–612页 (2004) 45. Wen, C., Zhang, Y., Li, Z., Fu, Y.: Pixel2mesh++: 通过变形生成多视图3D网格。在：IEEE/CVF国际计算机视觉会议论文集。第1042–1051页 (2019)

46. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv preprint arXiv:2310.08528 (2023)
47. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
48. Xu, Y., Peng, S., Yang, C., Shen, Y., Zhou, B.: 3d-aware image synthesis via learning structural and textural representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18430–18439 (2022)
49. Yang, J., Pavone, M., Wang, Y.: Freenerf: Improving few-shot neural rendering with free frequency regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8254–8263 (2023)
50. Yang, Y., Zhang, S., Huang, Z., Zhang, Y., Tan, M.: Cross-ray neural radiance fields for novel-view synthesis from unconstrained image collections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15901–15911 (2023)
51. Yang, Z., Gao, X., Sun, Y., Huang, Y., Lyu, X., Zhou, W., Jiao, S., Qi, X., Jin, X.: Spec-gaussian: Anisotropic view-dependent appearance for 3d gaussian splatting. arXiv preprint arXiv:2402.15870 (2024)
52. Yi, T., Fang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. arXiv preprint arXiv:2310.08529 (2023)
53. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenoctrees for real-time rendering of neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5752–5761 (2021)
54. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. arXiv preprint arXiv:2311.16493 (2023)
55. Zhang, J., Yang, G., Tulsiani, S., Ramanan, D.: Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. Advances in Neural Information Processing Systems **34**, 29835–29847 (2021)
56. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)

46. 吴, G., 易, T., 方, J., 谢, L., 张, X., 魏, W., 刘, W., 田, Q., 王, X.: 4D高斯溅射用于实时动态场景渲染。arXiv预印本arXiv:2310.08528 (2023) 47. 吴, Z., 宋, S., 科斯拉, A., 于, F., 张, L., 唐, X., 肖, J.: 3D形状网络：体积形状的深度表示。在：IEEE计算机视觉与模式识别会议论文集。第1912–1920页 (2015) 48. 徐, Y., 彭, S., 杨, C., 沈, Y., 周, B.: 通过学习结构和纹理表示实现3D感知图像合成。在：IEEE/CVF计算机视觉与模式识别会议论文集。第18430–18439页 (2022) 49. 杨, J., 帕沃内, M., 王, Y.: Freenerf：通过自由频率正则化改进少样本神经渲染。在：IEEE/CVF计算机视觉与模式识别会议论文集。第8254–8263页 (2023) 50. 杨, Y., 张, S., 黄, Z., 张, Y., 谭, M.: 通过无约束图像集合实现新视角合成的交叉射线神经辐射场。在：IEEE/CVF国际计算机视觉会议论文集。第15901–15911页 (2023) 51. 杨, Z., 高, X., 孙, Y., 黄, Y., 吕, X., 周, W., 焦, S., 齐, X., 金, X.: Spec-Gaussian：3D高斯溅射的各向异性视图依赖外观。arXiv预印本arXiv:2402.15870 (2024) 52. 易, T., 方, J., 吴, G., 谢, L., 张, X., 刘, W., 田, Q., 王, X.: Gaus-高斯梦想家：通过点云先验从文本到3D高斯溅射的快速生成。arXiv预印本arXiv:2310.08529 (2023) 53. 于, A., 李, R., 坦西克, M., 李, H., 吴, R., 金泽, A.: 用于神经辐射场实时渲染的Plenoctrees。在：IEEE/CVF国际计算机视觉会议论文集。第5752–5761页 (2021) 54. 于, Z., 陈, A., 黄, B., 萨特勒, T., 盖格, A.: Mip-溅射：无别名3D高斯溅射。arXiv预印本arXiv:2311.16493 (2023) 55. 张, J., 杨, G., 图尔萨尼, S., 拉马南, D.: Ners：用于野外稀疏视图3D重建的神经反射表面。神经信息处理系统进展 **34**, 29835–29847 (2021) 56. 张, R., 伊索拉, P., 埃夫罗斯, A.A., 谢赫特曼, E., 王, O.: 深度特征作为感知度量的不合理有效性。在：IEEE计算机视觉与模式识别会议论文集。第586–595页 (2018)

Supplementary material for "Gaussian in the Wild"

A Video demo

We strongly recommend readers to watch the video demo provided in webpage. The video showcases the novel view synthesis achieved by GS-W across various scenes, based on different reference images. It also illustrates the variations in scene appearance attained through the interpolation of different dynamic features df. Additionally, we provide a visual comparison between GS-W and CR-NeRF in terms of novel view synthesis and appearance tuning. The video illustrates GS-W's superior reconstruction of scene geometry and precise capture of environmental factors from the reference images, maintaining multi-view consistency. Furthermore, it highlights GS-W's ability to adjust scene appearance in a manner more aligned with human perception by weighting the features extracted from the reference images.

B More implementation details

B.1 Model parameters

Below, we provide a detailed overview of the network parameters and other hyperparameter settings in GS-W. The Unet's encoder utilizes the first 8 sub-convolution modules from a pre-trained ResNet-18 to extract image features. On the other hand, the decoder for generating $K+1$ feature maps comprises 4 up-sampling convolution modules, along with a decoder including 3 up-sampling convolution modules for producing the visibility map. The two decoder modules are both equipped with batch normalization and ReLU activation functions. To reduce computation, there are no skip connections between the encoder and the visibility map generating decoder during the forward process. Transpose convolution is employed for up-sampling, with the number of channels for each feature map fixed at 16. The learning rate for the Unet network gradually decreases from 2×10^{-3} to 2×10^{-5}.

The Fusion network M_f consists of two MLP modules, each with 3 and 2 hidden layers, and each with hidden unit counts of $[128, 96, 64]$ and $[48, 48]$, respectively. Meanwhile, the MLP M_c for color decoding comprises only one hidden layer with 48 units. Xavier initialization is used for initializing the weight parameters of these MLPs. During the forward process, we use 10 frequencies to encode position. For training, random dropout with 0.1 probability is applied to the dynamic appearance features df_i to prevent overfitting, and the learning rate for training these MLPs is set to 5×10^{-4}. The dimension of the intrinsic feature sf_i is set to 48.

《野外的高斯》的补充材料

A 视频演示

我们强烈建议读者观看提供的视频演示 网页。视频展示了GS-W基于不同参考图像在各种场景中实现的新视角合成。它还展示了通过插值不同动态特征获得的场景外观变化 df。此外，我们提供了GS-W和CR-NeRF在新视角合成和外观调优方面的视觉比较。视频展示了GS-W在重建场景几何和从参考图像中精确捕捉环境因素方面的优越性，保持了多视图一致性。此外，它突出了GS-W通过加权从参考图像中提取的特征，以更符合人类感知的方式调整场景外观的能力。

B 更多实现细节

B.1 模型参数

下面，我们详细介绍了GS-W中的网络参数和其他超参数设置。Unet的编码器使用预训练的ResNet-18的前8个子卷积模块来提取图像特征。另一方面，生成 $K+1$ 特征图的解码器包括4个上采样卷积模块，以及一个包含3个上采样卷积模块的解码器，用于生成可见性图。这两个解码器模块都配备了批量归一化和ReLU激活函数。为了减少计算量，在前向过程中，编码器和可见性图生成解码器之间没有跳跃连接。使用转置卷积进行上采样，每个特征图的通道数固定为16。Unet网络的学习率从2×10^{-3} 逐渐降低到 2×10^{-5}。

融合网络 M_f 由两个MLP模块组成，每个模块分别有3层和2层隐藏层，隐藏单元数量分别为 $[128, 96, 64]$ 和 $[48, 48]$,。同时，用于颜色解码的MLP M_c 仅包含一个隐藏层，有48个单元。这些MLP的权重参数使用Xavier初始化进行初始化。在前向过程中，我们使用10个频率来编码位置。在训练过程中，对动态外观特征 df_i 应用0.1概率的随机丢弃以防止过拟合，这些MLP的训练学习率设置为 5×10^{-4}。内在特征 sf_i 的维度设置为48。

Table 1: Average quantitative results on the Brandenburg, Sacre, and Trevi scenes.

	PSNR↑	SSIM↑	LPIPS↓	FPS↑	Size(MB)
global feature	22.43	0.8555	0.1349	**56.4**	123.0
K-planes-hybrid	22.40	0.7604	0.2731	0.67	133.3
Ours	**24.70**	**0.8655**	**0.1242**	50.7	122.7

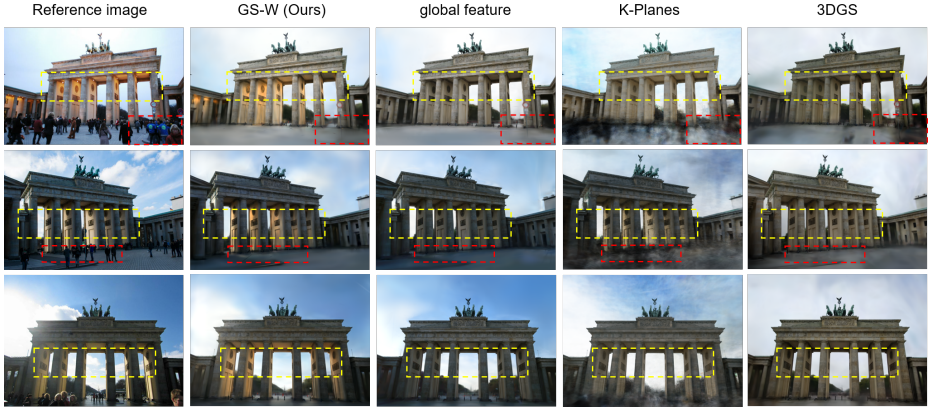

Fig. 1: More qualitative comparison results on Brandenburg. Our method captures local highlights and avoids artifacts caused by transient objects, as observed in K-Planes and 3DGS.

Furthermore, the learning rate for Gaussian point positions decreases from 1.6×10^{-4} to 1.6×10^{-7}. Gaussian points are densified from 500 iterations to 15k iterations during training, with a gradient threshold set to 4×10^{-4}. Other hyperparameters are set according to the guidelines of 3DGS.

B.2 Initialization of sampling coordinates

In Sec. 4.2, we equip each Gaussian point with K learnable sampling coordinate attributes $(sc_i^1, sc_i^2, ..., sc_i^K) \in \mathbb{R}^{K\times 2}$ to enable it to adaptively sample features on K feature maps. Specifically, here we describe how to initialize them before training. To ensure the points close to each other have similar initial sampling coordinates, we utilize linear transformation to convert the position $X_i \in \mathbb{R}^3$ of each Gaussian point into K sampling coordinates. First, we randomly generate K matrices $(M^1, M^2, ..., M^K) \in \mathbb{R}^{K\times 2\times 3}$, while ensuring that the sum of each matrix's rows is 1. Then, we multiply these matrices with each Gaussian point's position to obtain the initial values of these K sampling coordinates, as follows:

$$(sc_i^1, sc_i^2, ..., sc_i^K)_{init} = (M^1 X_i, M^2 X_i, ..., M^K X_i) \quad (1)$$

表1： A　　勃兰登堡、圣心和特雷维场景的平均定量结果。

	峰值信噪比↑	SSIM↑	LPIPS↓FPS（每秒帧数）↑		大小(MB)
全局特征	22.43	0.8555	0.1349	**56.4**	123.0
K-平面混合	22.40	0.7604	0.2731	0.67	133.3
Ours	**24.70**	**0.8655**	**0.1242**	50.7	122.7

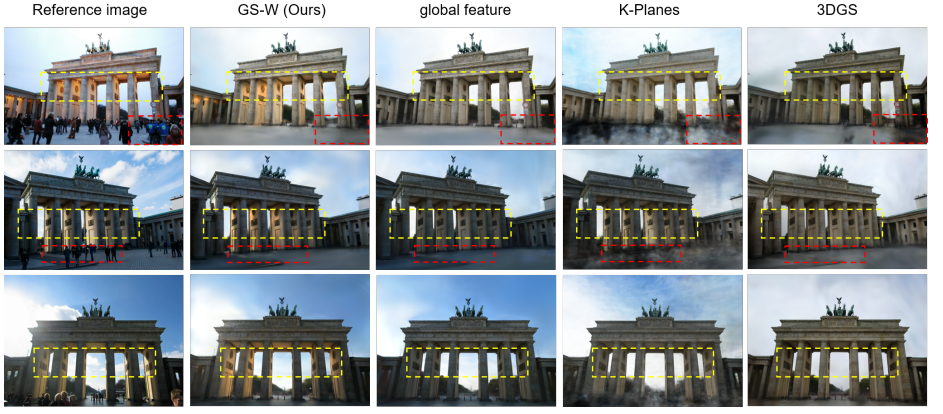

图1: 在勃兰登堡的更多定性比较结果。我们的方法捕捉局部高光并避免了由瞬态对象引起的伪影，如在K-平面和3DGS中观察到的那样。

此外，高斯点位置的学习率从1.6×10^{-4}降低到 1.6×10^{-7}。在训练过程中，高斯点从500次迭代到15k次迭代进行密集化，梯度阈值设置为 4×10^{-4}。其他超参数根据3DGS的指南进行设置。

B.2 采样坐标的初始化

在第4.2节中，我们为每个高斯点配备K个可学习的采样坐标属性（$sc_i^1, sc_i^2, ..., sc_i^K$）$\in \mathbb{R}^{K\times 2}$，使其能够自适应地在K个特征图上采样特征。具体来说，这里我们描述如何在训练前初始化它们。为了确保彼此靠近的点具有相似的初始采样坐标，我们利用线性变换将每个高斯点的位置 $X_i \in \mathbb{R}^3$ 转换为K个采样坐标。首先，我们随机生成K个矩阵（$M^1, M^2, ..., M^K$）$\in \mathbb{R}^{K\times 2\times 3}$，同时确保每个矩阵的行和为1。然后，我们将这些矩阵与每个高斯点的位置相乘，以获得这些K个采样坐标的初始值，如下所示：

$$(sc_i^1, sc_i^2, ..., sc_i^K)_{init} = (M^1 X_i, M^2 X_i, ..., M^K X_i) \quad (1)$$

Table 2: Quantitative results on the Sacre Coeur for different K values, where K represents the number of feature maps. We Choose K = 3 to achieve efficiency and effectiveness.

	Sacre Coeur				
	PSNR↑	SSIM↑	LPIPS↓	FPS↑	Size(MB)
K=1	23.08	0.8598	0.1318	**60.3**	93.8
K=2	23.13	0.8590	0.1334	59.6	96.1
K=3	**23.24**	**0.8632**	0.1300	58.3	97.2
K=4	23.02	0.8616	0.1303	58.0	98.2
K=8	22.94	0.8618	**0.1296**	55.8	101.8
K=16	22.98	0.8611	0.1321	51.8	113.9

Table 3: Quantitative results on the synthetic dataset with introduced perturbations (colors & occluders). GS-W outperforms the others in all evaluation metrics.

	Synthetic Lego Dataset		
	PSNR↑	SSIM↑	LPIPS↓
3DGS	23.73	0.9250	0.0748
NeRF-W	26.76	0.9224	0.0552
Ha-NeRF	26.51	0.9416	0.0339
GS-W (Ours)	**29.64**	**0.9522**	**0.0323**

C Further experiments

C.1 More comparison

Global appearance feature. To highlight the superiority of independently sampled dynamic appearance features over a single global feature, We replace each point's dynamic appearance feature with a global one extracted from the reference image. As shown in Tab. 1 and Fig. 1, the global feature method achieves an inferior result and fails to recover local appearance details on the pillars and sky.

K-Planes. Aiming to further demonstrate the advantages of our method in rendering quality and speed, we compared it with K-Planes [10]. Unlike NeRF-based methods, K-Planes uses a hybrid 3D representation that achieves lower storage and faster rendering. Quantitative results, including metrics for storage and rendering efficiency, along with qualitative results, are presented in Tab. 1,Tab. 4,Fig. 1 and Fig. 4. Our method continues to achieve superior rendering quality and speed in comparison.

3DGS. In Fig. 1 and Fig. 4, we provide visual comparisons between 3DGS and our method. It's clear that 3DGS fails to capture scene appearance variations and handle transient objects, leading to noticeable flaws and artifacts.

C.2 K value selection

In our method statement, K represents the number of feature maps used for sampling. We conducted experiments under the Sacre Coeur scene to determine

表 2: 在圣心大教堂上使用不同 K 值的定量结果，其中 K 表示特征图的数量。我们选择 K = 3 以实现效率和有效性。

	圣心大教堂				
	PSNR↑	SSIM↑	LPIPS↓	FPS↑	大小(MB)
K=1	23.08	0.8598	0.1318	**60.3**	93.8
K=2	23.13	0.8590	0.1334	59.6	96.1
K=3	**23.24**	**0.8632**	0.1300	58.3	97.2
K=4	23.02	0.8616	0.1303	58.0	98.2
K=8	22.94	0.8618	**0.1296**	55.8	101.8
K=16	22.98	0.8611	0.1321	51.8	113.9

表3: 在引入扰动（颜色和遮挡物）的合成数据集上的定量结果。GS-W在所有评估指标上均优于其他方法。

	合成乐高数据集		
	峰值信噪比↑	SSIM↑	LPIPS↓
3DGS	23.73	0.9250	0.0748
NeRF-W	26.76	0.9224	0.0552
Ha-NeRF	26.51	0.9416	0.0339
GS-W（我们的）	**29.64**	**0.9522**	**0.0323**

C 进一步实验

C.1 更多比较

全局外观特征。为了突出独立采样的动态外观特征相对于单一全局特征的优越性，我们将每个点的动态外观特征替换为从参考图像中提取的全局特征。如表1和图1所示，全局特征方法取得了较差的结果，并且无法恢复柱子和天空上的局部外观细节。

K-Planes. 为了进一步展示我们方法在渲染质量和速度上的优势，我们将其与K-Planes [10]进行了比较。与基于NeRF的方法不同，K-Planes使用了一种混合的三维表示，实现了更低的存储和更快的渲染。定量结果，包括存储和渲染效率的指标，以及定性结果，呈现在表1、表4、图1和图4中。我们的方法在比较中继续实现了更优越的渲染质量和速度。

3DGS. 在图1和图4中，我们提供了3DGS与我们方法的视觉比较。显然，3DGS无法捕捉场景外观的变化并处理瞬态物体，导致明显的缺陷和伪影。

C.2 K值选择

在我们的方法描述中，K表示用于采样的特征图数量。我们在圣心大教堂场景下进行了实验以确定

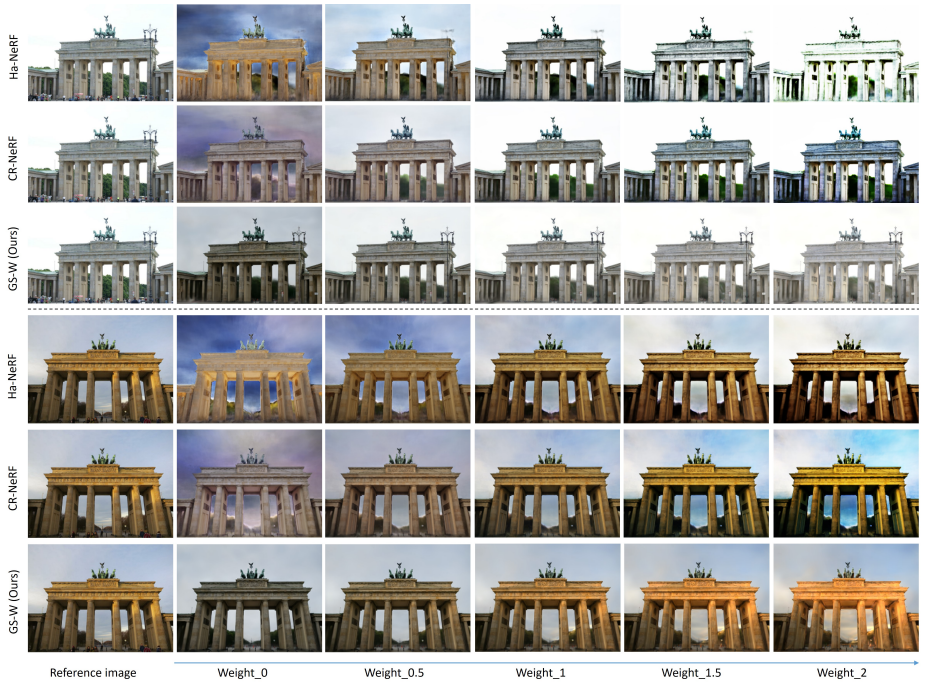

Fig. 2: More qualitative comparison experiments on appearance tuning. Similar to Fig. 5, images are rendered at the same camera pose with increasing weight of features extracted from the image. Our method not only captures environmental information better but also naturally adjusts its influence on the scene.

an effective K, setting K to 1, 2, 3, 4, 8, and 16 respectively. The experimental results are presented in Tab. 2. We observe that the performance on the test set does not improve when K exceeds 3 while bringing more computation cost. Thus, we reasonably set K to 3.

C.3 More appearance tuning results

In Fig. 2, we present additional qualitative comparison results for appearance tuning. From the figure, it's fair to conclude that our method applies the environmental factors from the reference images to the scene more reasonably, gradually increasing their influence. In the first example from top to bottom, the colors of the buildings become strange as the weight increases in the other two methods which is against practical human understanding. In the second example, our method captures the highlights on the pillars better and enhances them as the weight increases.

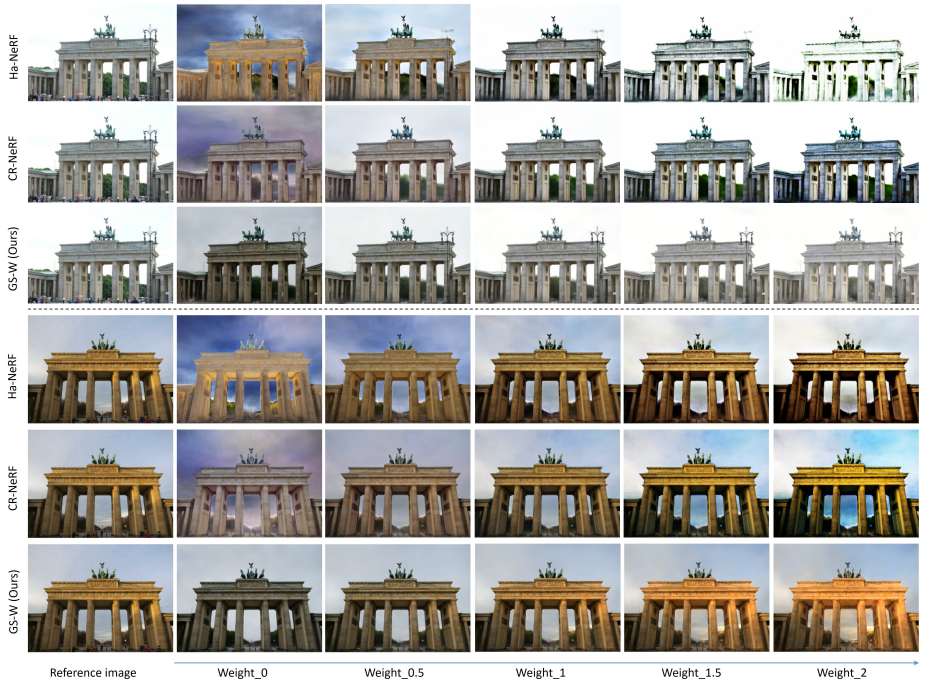

图2: 关于外观调优的更多定性比较实验。与图5类似，图像在相同的相机姿态下渲染，特征提取的权重逐渐增加。我们的方法不仅更好地捕捉了环境信息，还自然地调整了其对场景的影响。

一个有效的K值，分别将K设置为1、2、3、4、8和16。实验结果见表2。我们观察到当K超过3时，测试集上的性能没有提升，但计算成本增加。因此，我们合理地将K设置为3。

C.3 更多外观调整结果

在图2中，我们展示了外观调优的额外定性比较结果。从图中可以合理地得出结论，我们的方法更合理地将参考图像中的环境因素应用于场景，并逐渐增加其影响。在从上到下的第一个示例中，随着权重的增加，其他两种方法中建筑物的颜色变得奇怪，这与实际的人类理解相悖。在第二个示例中，我们的方法更好地捕捉了柱子上的高光，并随着权重的增加增强了它们。

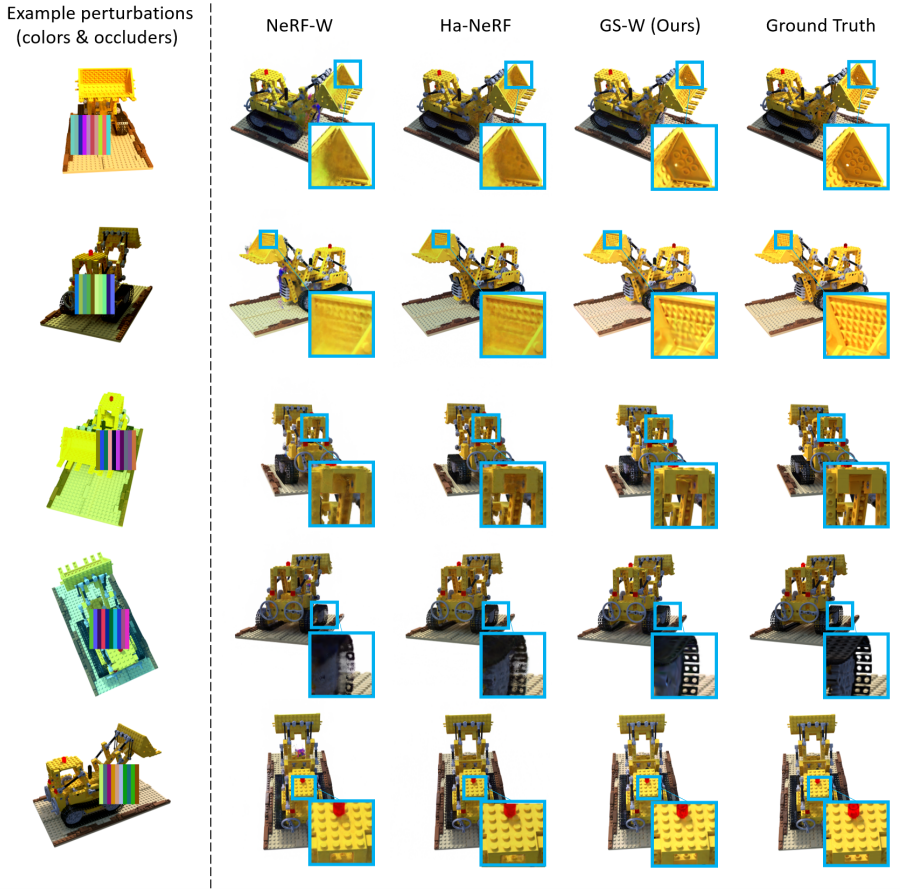

Fig. 3: Experiment results on synthetic Lego dataset with perturbations (colors & occluders). Our method excels in modeling finer details and achieving more accurate color.

C.4 Synthetic Lego dataset

Following NeRF-W and Ha-NeRF, we further compared GS-W with them on the synthetic Lego dataset to validate our method's effectiveness. To ensure a fair comparison, we followed their setup by manually introducing perturbations such as colors and occluders in the training set of the synthetic Lego dataset to simulate scenarios that might be encountered in the wild, as the first column in Fig. 3. The training set comprises 100 images, while the test set consists of 200 images. During testing, only a single unperturbed training image is used as the reference image to extract features. Subsequently, novel views of the test images are rendered, and metrics are calculated against the test images.

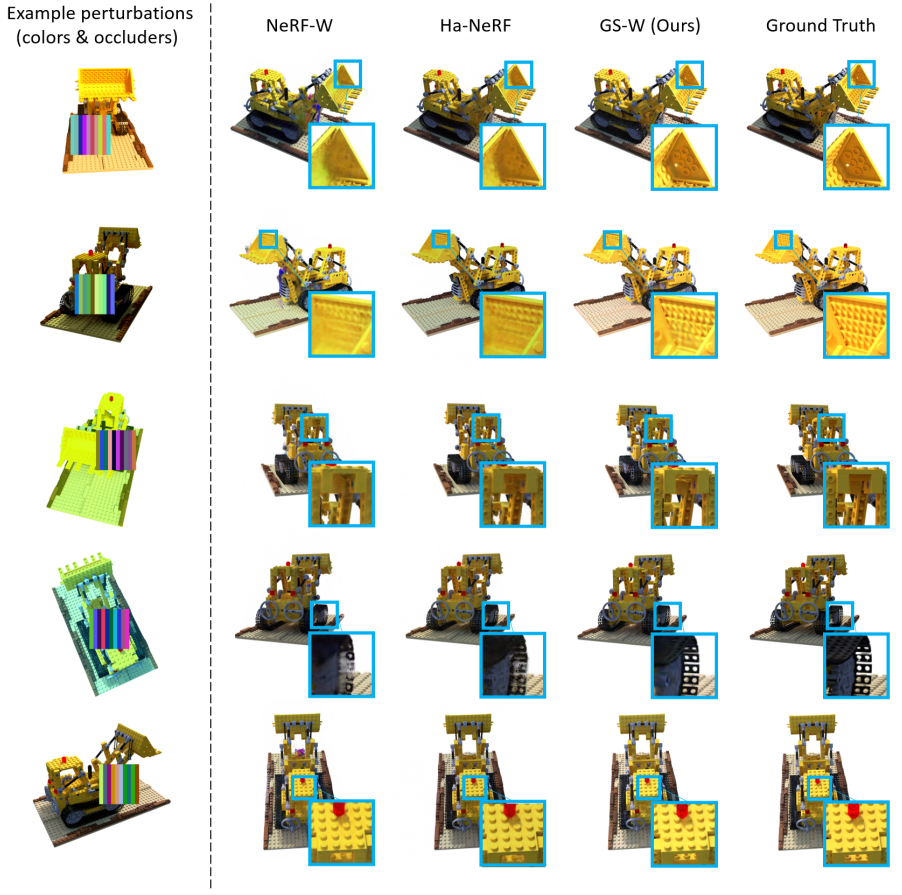

图3： 在具有扰动（颜色和遮挡物）的合成乐高数据集上的实验结果。我们的方法在建模更精细的细节和实现更准确的颜色方面表现出色。

C.4 合成乐高数据集

遵循NeRF-W和Ha-NeRF，我们进一步在合成乐高数据集上将GS-W与它们进行比较，以验证我们方法的有效性。为了确保公平比较，我们遵循它们的设置，通过在合成乐高数据集的训练集中手动引入颜色和遮挡物等扰动来模拟可能在野外遇到的场景，如图3的第一列所示。训练集包含100张图像，而测试集包含200张图像。在测试过程中，仅使用一张未扰动的训练图像作为参考图像来提取特征。随后，渲染测试图像的新视图，并计算与测试图像的指标。

Table 4: Quantitative results on the test set of four NeRF-OSR scenes. Our method outperforms other methods across all scenes on all metrics.

	stjohann			lwp			st			europa		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS	17.32	0.7430	0.313	15.43	0.7000	0.3360	15.23	0.5920	0.4240	16.32	0.6520	0.3270
K-planes-hybrid	20.39	0.7548	0.3366	21.65	0.7482	0.3397	19.66	0.6151	0.4361	18.75	0.6487	0.4549
NeRF-W	21.38	0.8200	0.1940	21.29	0.7610	0.2950	19.68	0.6310	0.4010	19.55	0.6870	0.3470
Ha-NeRF	19.93	0.7870	0.2100	21.32	0.7560	0.2780	20.56	0.6360	0.3780	18.76	0.6610	0.3480
CR-NeRF	22.27	0.8350	0.1750	22.61	0.7850	0.2580	21.67	0.6610	0.3600	19.92	0.6960	0.3100
GS-W (Ours)	**26.23**	**0.8927**	**0.1133**	**24.44**	**0.8272**	**0.2082**	**22.45**	**0.6900**	**0.3147**	**22.04**	**0.7577**	**0.2309**

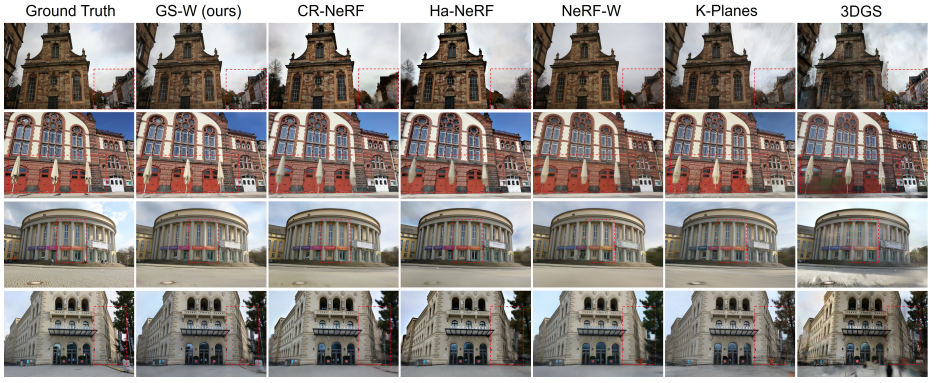

Fig. 4: Qualitative comparison results on the NeRF-OSR dataset test set are shown for the stjohann, lwp, st, and europa scenes. Our method reconstructs the scenes with greater detail compared to other methods.

As CR-NeRF does not provide hyperparameter settings for this dataset, the attempts to use its default parameters from other datasets result in training crashes and produce poor and peculiar results. Therefore, we only compare with NeRF-W, Ha-NeRF, and 3DGS on this dataset. Due to the limited size of the Lego training dataset and each view being influenced by a fixed color, we decode color using position instead of view direction to prevent dependency on the latter in appearance modeling. During training, without initialized point clouds, we adjust the gradient threshold for Gaussian point densification to 1.5×10^{-4} to allow the model to generate more valid points. Other hyperparameters remain the same, with training 20k iterations.

Quantitative and qualitative results are presented in Tab. 3 and Fig. 3, respectively. We can observe that 3DGS performs poorly while our method performs the best under scenarios with added color and occluder perturbations. This testing process, conducted on novel viewpoints rather than on the original reference image viewpoints, also partially validates the multi-view consistency of GS-W and its ability to generalize single-image features to new viewpoints.

表4: 在四个NeRF-OSR场景的测试集上的定量结果。我们的方法在所有场景和所有指标上都优于其他方法。

	stjohann			lwp			st			欧洲		
	峰值信噪比↑	SSIM↑	LPIPS↓	峰值信噪比↑	SSIM↑	LPIPS↓	峰值信噪比↑	SSIM↑	LPIPS↓	峰值信噪比↑	SSIM↑	LPIPS↓
3DGS	17.32	0.7430	0.313	15.43	0.7000	0.3360	15.23	0.5920	0.4240	16.32	0.6520	0.3270
K-平面混合	20.39	0.7548	0.3366	21.65	0.7482	0.3397	19.66	0.6151	0.4361	18.75	0.6487	0.4549
NeRF-W	21.38	0.8200	0.1940	21.29	0.7610	0.2950	19.68	0.6310	0.4010	19.55	0.6870	0.3470
Ha-NeRF	19.93	0.7870	0.2100	21.32	0.7560	0.2780	20.56	0.6360	0.3780	18.76	0.6610	0.3480
CR-NeRF	22.27	0.8350	0.1750	22.61	0.7850	0.2580	21.67	0.6610	0.3600	19.92	0.6960	0.3100
GS-W（我们的）	**26.23**	**0.8927**	**0.1133**	**24.44**	**0.8272**	**0.2082**	**22.45**	**0.6900**	**0.3147**	**22.04**	**0.7577**	**0.2309**

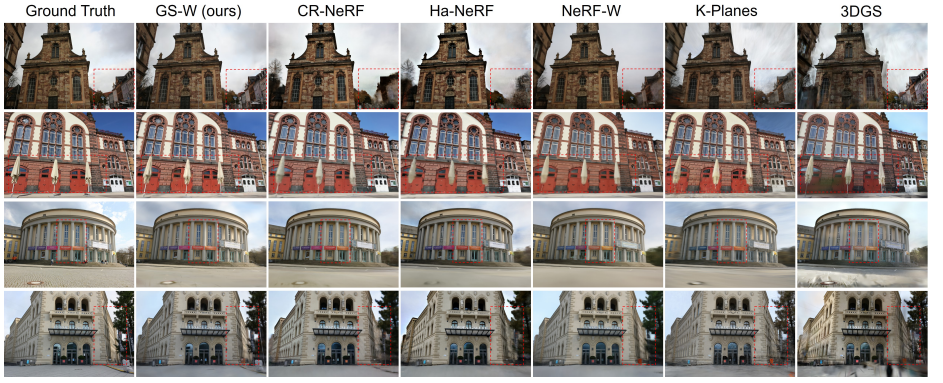

图4: 在NeRF-OSR数据集测试集上展示了stjohann、lwp、st和europa场景的定性比较结果。我们的方法与其他方法相比，重建的场景具有更多细节。

由于CR-NeRF未提供此数据集的超参数设置，尝试使用其默认参数从其他数据集进行训练会导致训练崩溃并产生较差和奇特的结果。因此，我们仅在此数据集上与NeRF-W、Ha-NeRF和3DGS进行比较。由于乐高训练数据集的规模有限且每个视图受固定颜色影响，我们使用位置而不是视图方向来解码颜色，以防止外观建模依赖于后者。在训练过程中，由于没有初始化的点云，我们将高斯点密度化的梯度阈值调整为 1.5×10^{-4}，以允许模型生成更多有效点。其他超参数保持不变，训练20k次迭代。

定量和定性结果分别在表3和图3中呈现。我们可以观察到，在添加颜色和遮挡扰动的场景下，3DGS表现不佳，而我们的方法表现最佳。这个测试过程是在新的视点上进行的，而不是在原始参考图像视点上进行的，这也部分验证了GS-W的多视图一致性和其将单图像特征泛化到新视点的能力。

Fig. 5: Style transfer from unseen images or across scenes for the three scenes. The first row represents the reference images used for extracting appearance features, while the subsequent three rows depict images rendered based on the views of the content images and the corresponding appearance features of the reference images. This demonstrates that our method can effectively perform style transfer.

C.5 NeRF-OSR dataset

We further evaluate the robustness of GS-W using the NeRF-OSR dataset [36]. We experiment with four scenes - europa, lwp, st, and stjohann - with 12.5% of images from each sequence as the test set. Each training set contains approximately 350 images, and each test set contains about 50 images.

We compared our method with NeRF-W, Ha-NeRF, CR-NeRF, 3DGS, and K-Planes, using their default settings. Evaluation results on the test set, shown in Tab. 4 and Fig. 4, demonstrate that our method consistently achieves the best performance. Other methods struggle to capture local details, and 3DGS fails to handle variations in scene appearance and transient occlusions. In contrast, our method accurately captures local details and effectively handles transient occlusions.

Despite our method achieving good performance, this dataset, not sourced from the internet, features limited appearance variations per scene. Consequently, the reconstructed scenes may exhibit inconsistencies in appearance across different viewpoints. In the future, training a more generalized dynamic appearance extractor across multiple scenes could effectively address and improve this issue.

D Style transfer

Similar to CR-NeRF, our method can also transfer style from images to the scene, as shown in Fig. 5. However, for style transfer across scenes or from unseen images, the absence of camera poses from the provided reference images makes it difficult to map Gaussian points to the projection feature map for

图5： 针对三个场景的未见图像或跨场景的风格迁移。第一行表示用于提取外观特征的参考图像，而接下来的三行描绘了基于内容图像的视图和参考图像的相应外观特征渲染的图像。这表明我们的方法可以有效地执行风格迁移。

C.5 NeRF-OSR数据集

我们进一步使用NeRF-OSR数据集 [36]评估GS-W的鲁棒性。我们实验了四个场景 - 欧洲、lwp、st和stjohann - 每个序列的12.5%图像作为测试集。每个训练集包含大约350张图像，每个测试集包含大约50张图像。

我们使用默认设置将我们的方法与NeRF-W、Ha-NeRF、CR-NeRF、3DGS和K-平面进行了比较。在测试集上的评估结果，如表4和图4所示，表明我们的方法始终实现了最佳性能。其他方法难以捕捉局部细节，而3DGS无法处理场景外观的变化和瞬时遮挡。相比之下，我们的方法准确捕捉了局部细节，并有效处理了瞬时遮挡。

尽管我们的方法表现良好，但该数据集并非来自互联网，每个场景的外观变化有限。因此，重建的场景在不同视点之间可能会出现外观不一致。未来，训练一个更通用的动态外观提取器跨多个场景可以有效解决并改善这一问题。

D 风格迁移

类似于CR-NeRF，我们的方法也可以将图像的风格迁移到场景中，如图5所示。然而，对于跨场景或从未见图像进行风格迁移时，由于提供的参考图像中缺乏相机姿态，使得将高斯点映射到投影特征图变得困难。

sampling. Meanwhile, in style transfer, there lack of physical significance to using the projection sampling method. Therefore, when performing style transfer from unseen images, we set all features f^P sampled from the projection feature map to 0. The results demonstrate that our method can effectively transfer the style provided by reference images to the scene.

26

采样。同时，在风格迁移中，使用投影采样方法缺乏物理意义。因此，在从未见图像进行风格迁移时，我们将从投影特征图中采样的所有特征 f^P 设置为0。结果表明，我们的方法可以有效地将参考图像提供的风格迁移到场景中。