DentalGS: Pose-Free 3D Gaussian Splatting from Five Intraoral Images for Novel View Synthesis

Honghao Dai[1], Yuanfeng Zhou[1*], Guangshun Wei[1], Zhihao Li[1], Wenping Wang[2]

[1]School of Software, Shandong University, Jinan, China
[2]Texas A&M University, USA
dhhtang@163.com, yfzhou@sdu.edu.cn, guangshunwei@gmail.com, li15106431515@163.com, wenping@tamu.edu

Abstract

Orthodontic treatment needs regular tooth alignment checks, but current methods depend on clinic visits, limiting remote care. With the emergence of 3D Gaussian Splatting (3DGS), realistic novel views can be synthesized, making it possible for clinicians to remotely monitor orthodontic conditions. However, using only five intraoral images with unknown camera poses and dynamic lighting presents major challenges in dental applications. To address these challenges, we propose DentalGS, an enhanced 3DGS framework capable of synthesizing novel intraoral views from five post-orthodontic intraoral images and pre-orthodontic intraoral scan (IOS) data as prior, without camera poses. Our method initializes a Gaussian point cloud labeled with ISO-FDI tooth classes based on the patient's pre-orthodontic IOS data, then estimates camera poses through iterative optimization. We introduce a Progressive Pair Generation Strategy as a data augmentation method that generates damage–repair image pairs to train a RepairNet, aiming to restore degraded geometry and appearance caused by the limited number of intraoral images. Additionally, we introduce a Lighting-Aware 3DGS inspired by physical reflectance properties to mitigate the effects of dynamic lighting conditions. Experimental results show that our method produces high-quality novel views while preserving geometric structure even under extreme viewpoints, offering an efficient and reliable solution for 3D tooth visualization in remote orthodontic monitoring.

Introduction

Accurate monitoring is essential for effective orthodontic treatment. In clinical practice, dentists typically rely on intraoral photographs, Cone-Beam Computed Tomography (CBCT) (Kapila, Conley, and Harrell Jr 2011), and Intraoral Scanners (IOS) (Hong-Seok and Chintal 2015) to obtain detailed digital models of teeth and gingiva. However, such professional equipment is often bulky, expensive, and requires trained personnel to operate. Frequent usage not only increases the burden on patients but also limits the feasibility of remote dental care and at-home follow-up.

Many patients aim to use smartphones to capture intraoral photos for remote check-ups, reducing cost and improving access. However, 2D images lack spatial information, making this approach insufficient for accurate diagnosis. Recent advances in novel view synthesis (NVS), such as Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023), make it possible to synthesize realistic 3D views from 2D images, potentially supporting accurate occlusal evaluation at home. Despite this progress, applying NVS in dental scenarios remains difficult. We identify three key challenges in dental NVS: (1) Low overlap between images and the smooth surfaces of teeth make it difficult for Structure-from-Motion (SfM) using COLMAP (Schonberger and Frahm 2016) to extract stable features, leading to camera pose errors; (2) With only five-intraoral images, it's hard to build a complete 3D model, leading to overfitting and broken geometry that doesn't work well for NVS; (3) Dynamic lighting causes strong illumination inconsistencies across views, creating visual artifacts.

To address these challenges, we propose DentalGS, a 3D Gaussian Splatting-based framework specifically designed for dental NVS using five post-orthodontic intraoral images and pre-orthodontic IOS data as prior, without camera poses (Kerbl et al. 2023). Leveraging anatomical priors of dental structures, we design an Iterative Camera Pose Fitting Algorithm to estimate camera poses based on viewing frustum. To address the geometric degradation, we design a RepairNet to restore rendered images. By leveraging pre-orthodontic IOS as prior knowledge, RepairNet transforms fragmented outputs into structurally complete and texture-consistent results, while also simulating lighting throughout the restoration process. Due to the lack of a dedicated dataset for Gaussian restoration, we further propose a progressive pair generation strategy to simulate view restoration scenarios for training RepairNet. To address lighting inconsistencies across views, we propose a lighting-aware Gaussian based on a physics-inspired shading model. The repaired images from RepairNet are used as coarse supervision to guide the learning of appearance attributes in the Gaussian, enabling realistic reflectance modeling under limited lighting cues. Ultimately, our method achieves high-fidelity and structurally consistent 3D dental NVS from five intraoral images, enabling low-cost remote orthodontic monitoring. In summary, our contributions include:

- We propose **DentalGS**, a pose-free 3DGS framework specifically designed for novel view synthesis in den-

*Corresponding author: Yuanfeng Zhou

tal scenarios using five post-orthodontic intraoral images and pre-orthodontic IOS data. By leveraging anatomical priors, we develop an iterative camera pose fitting algorithm that robustly estimates camera poses under.

- We propose a **progressive pair generation strategy** to train **RepairNet**, a network designed to address geometric fragmentation caused by five intraoral views. This strategy enables the network to learn from simulated damage-repair pairs with increasing complexity.
- We introduce a physically inspired, reflectance-guided, **lighting-aware Gaussian**. By using restored views from RepairNet to jointly supervise 3D appearance learning, we achieve consistent reflectance modeling with only five images and inconsistent lighting, enabling high-fidelity novel view synthesis for remote orthodontic monitoring.

Related Work

In clinical intraoral photography, image acquisition typically relies on manual operation by experienced practitioners, lacking accurate camera pose information. To address 3D reconstruction from pose-less image collections, recent works such as (Wang et al. 2025a), (Wang et al. 2024) and (Dong et al. 2025) explore dense and unconstrained reconstruction via large-scale pre-trained models. However, these methods are not specifically trained for medical scenarios, resulting in noisy predictions and inaccurate pose estimates. (Guo et al. 2024) propose to jointly optimize camera poses and scene representations for surgical scene reconstruction, but this approach requires continuous video input and optical flow priors. (Chang, Wang, and Lai 2025) utilize structure-from-motion (SfM) to obtain coarse pose estimates, followed by refinement, yet the method still relies on sufficient feature correspondences, which are often lacking in intraoral views.

To overcome challenges posed by sparse-view reconstruction, existing methods generally follow two directions. The first line of work introduces auxiliary geometric priors. For instance, (Kim 2023), (Hou et al. 2025), (Xu et al. 2022), and (Wang et al. 2023) enhance scene realism and continuity through depth estimation, smoothness constraints, or regularization. (Han et al. 2025) further improves data diversity by generating virtual lane imagery. In contrast, (Chen et al. 2023) leverage statistical shape models for parameterized dental geometry reconstruction, which incorporates structural priors but often lacks photorealism. The second direction incorporates data-driven models into 3DGS to improve performance under sparse-view settings. For example, (Zhang et al. 2025) introduce semantic guidance into 3DGS to enhance geometric consistency, while (Xu et al. 2024) employs large models to reconstruct high-quality dental models from only five intraoral images. However, these large diffusion-based models are prone to hallucinations and may fail to produce anatomically reliable results.

Furthermore, lighting modeling—an essential factor for visual consistency—has been progressively integrated into the 3DGS framework. (Jiang et al. 2024) and (Gao et al. 2024) preliminarily explore illumination-aware modeling, but their assumptions of static lighting conditions limit their generality. (Zhang et al. 2024) introduces dynamic lighting into 3DGS and discusses reconstruction under moving light sources, yet it lacks detailed treatment of strong specular reflections typical in intraoral scenes. (Wang et al. 2025b) attempts to isolate reflective components from five-view intraoral images using only luminance cues, but does not explicitly model lighting effects. However, existing illumination-aware models typically rely on dense multi-view inputs to capture lighting variation accurately. In sparse-view scenarios, the lack of diverse lighting supervision makes it difficult to learn generalizable reflectance, often resulting in increased noise in NVS.

Method

Overview

As shown in Fig. 1, our pipeline consists of four stages. In Stage 1, we initialize the IOS data using five post-orthodontic intraoral images to obtain view-aligned point clouds, followed by an iterative fitting algorithm to estimate camera poses. In Stage 2, we propose a progressive pair generation strategy during 3DGS optimization to simulate diverse degradation scenarios, constructing a rich set of training image pairs. In Stage 3, we introduce RepairNet, a geometry-aware restoration network trained on the generated image pairs. It utilizes IOS priors and combines geometric and chromatic cues to restore fragmented renderings into structurally complete and texture-consistent images. In Stage 4, we present a lighting-aware 3DGS approach guided by a physics-inspired shading model, which jointly optimizes the restored views and the original input images to achieve illumination-consistent reconstruction.

Camera Pose Estimation

A reliable initial point cloud and accurate camera poses are vital for NVS. Traditional methods like 3DGS rely on COLMAP to reconstruct geometry and estimate poses from multi-view images. However, with only five intraoral views, COLMAP often fails due to low overlap and smooth tooth surfaces. To overcome this, we propose an Iterative Camera Pose Fitting algorithm that uses pre-orthodontic IOS data as a geometric prior for pose estimation.

As shown in Stage 1 of Fig. 1, we first apply Multi-view Contour Fitting (MCF) (Xie et al. 2024) to align the five post-orthodontic intraoral images with the pre-orthodontic IOS data, estimating a post-orthodontic point cloud. We then assign ISO-FDI labels to each tooth point cloud $\mathcal{P}_i$, compute per-tooth pixel centers for each image, and refine the camera poses using Algorithm 1.

Where PnP (Perspective-n-Point) (Lepetit, Moreno-Noguer, and Fua 2009) estimates camera poses using 2D–3D correspondences; Rays represent a viewing frustum originating from the camera and passing through each pixel. This iterative algorithm suppresses occlusion-induced errors and enhances the robustness and accuracy of multi-view camera pose estimation.

RepairNet Training Strategy

In five-view intraoral images, 3DGS often produces geometric distortions and appearance inconsistencies under novel

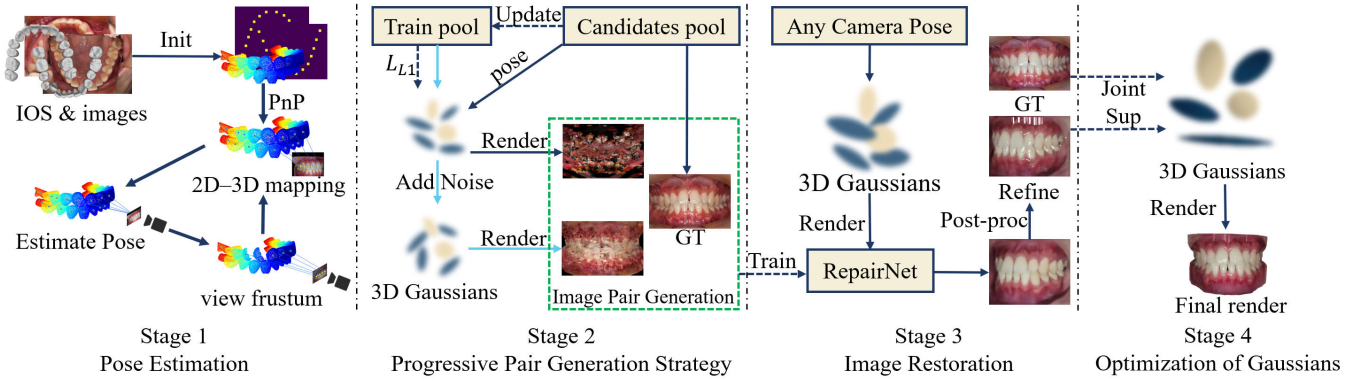

Figure 1: Overview of the proposed pipeline. The proposed framework consists of four main stages. **Stage 1**: Input data is initialized, and camera poses are estimated via Iterative Camera Pose Fitting. **Stage 2**: A progressive pair generation strategy is employed to generate image pairs for training RepairNet. **Stage 3**: Novel views are sampled from rendered degraded images, which are then restored by RepairNet. **Stage 4**: Both restored and GT images jointly supervise the lighting-aware 3DGS.

Algorithm 1: Iterative Camera Pose Fitting

Input: Point-cloud set $\{\mathcal{P}_i\}$, per teeth pixel centers set $\{u_i\}$, intrinsics $\mathbf{K}$, number of iterations $\mathbf{N}$, tooth class i
Output: extrinsics $(\mathbf{R}^{(N)}, \mathbf{t}^{(N)})$

Compute initial centroid $\mathbf{p}_i^{(0)}$ from $\mathcal{P}_i$ of each tooth.
$(\mathbf{R}^{(0)}, \mathbf{t}^{(0)}) = PnP((\mathbf{p}_i^{(0)}, \mathbf{u}_i), \mathbf{K})$
for $n = 0$ **to** $N - 1$ **do**
  $\mathcal{D} = \{\}$
  Build view frustum $\mathcal{F}^{(H\times W\times 3)} = Rays(\mathbf{R}^{(n)}, \mathbf{t}^{(n)})$
  for f **in** $\mathcal{F}$ **do**
    Find the first intersection point $p_i^n \in \mathcal{P}_i$ with f
    $\mathcal{D}.append(p_i^n)$
  end for
  Recompute centroid $\mathbf{p}_i^{(n+1)}$ from $\mathcal{D}$
  $(\mathbf{R}^{(n+1)}, \mathbf{t}^{(n+1)}) = PnP((\mathbf{p}_i^{(n+1)}, \mathbf{u}_i), \mathbf{K})$
end for
return $(\mathbf{R}^{(N)}, \mathbf{t}^{(N)})$

viewpoints. To address this, inspired by (Yang et al. 2024), we introduce a RepairNet R that leverages structured cues to refine degraded renderings. However, training R requires diverse supervision pairs, which are scarce in this setting. To overcome this limitation, as shown in Stage 2 of Fig.1 and Fig.2 for details, we propose a Progressive Pair Generation Strategy, which incrementally constructs rich training pairs from limited data, enabling R to learn structural refinement under various challenging conditions.

Progressive Pair Generation Strategy. We divide the view $(v = (\mathbf{R}^{(N)}, \mathbf{t}^{(N)}, \mathbf{K}), x^{\text{gt}})$ into two subsets: the Train Pool Π_{train}, initially containing only upper and lower views, and the Candidates Pool $\Pi_{\text{candidates}}$ with the rest. After optimizing 3DGS G_0 solely with Π_{train}, we initiate the Progressive Pair Generation Strategy.

Based on the model G_0, as shown in Fig. 2, our algorithm simulates four representative degradation scenarios to construct image pairs. By progressively optimizing G_0 and

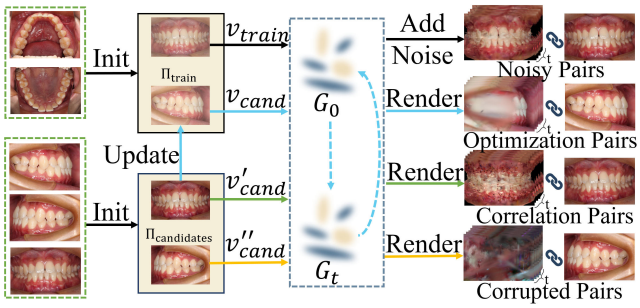

Figure 2: Overview of the four scenarios: Noisy Pairs; Optimization Pairs; Correlation Pairs; Corrupted Pairs.

providing different v, G and x^{gt}, we generate corresponding rendered images using $x_t = G_t(v), \{x_t, x_{\text{gt}}\}$, which are then paired with the ground truth to simulate common challenges encountered in image restoration tasks.

Specifically, they include: Optimization Pairs, simulating the full restoration trajectory from severe degradation to full recovery and capturing a fine-grained difficulty spectrum; Correlation Pairs, generated from spatially overlapping views during optimization, reflect moderate difficulty as they benefit from cross-view consistency; Corrupted Pairs, created from unrelated viewpoints with no spatial overlap, are the most challenging to restore; Noisy Pairs, produced by adding noise to fully optimized views, are used to enhance the model's robustness to noise perturbations. Details on how to choose v, t, and x^{gt} are provided below.

1. **Optimization Pairs** (Blue Arrow): We select a view $(v_{\text{cand}}, x_{\text{gt}}^{\text{cand}}) \in \Pi_{\text{candidates}}$ as supervision and perform a single optimization on G_0, producing intermediate models G_t over iterations $t \in [1, t_{\max}]$. At each step, we render an image using v, t, and x^{gt}. During this process, a sequence of rendered images from fully degraded to fully optimized. Finally, once $t = t_{\max}$, the view $(v_{\text{cand}}, x_{\text{gt}}^{\text{cand}})$ is moved from $\Pi_{\text{candidates}}$ to Π_{train} and $G_0 = G_t$.
2. **Correlation Pairs** (Green Arrow): At optimization step

t, we select the other views $(v'_{\text{cand}}, x'_{\text{gt}}) \in \Pi_{\text{candidates}}$ that overlap spatially with v_{cand}.

3. **Corrupted Pairs** (Yellow Arrow): At optimization step t, we randomly select view $(v''_{\text{cand}}, x''_{\text{gt}}) \in \Pi_{\text{candidates}}$, which has not yet been optimized or observed as Corrupted Pairs.
4. **Noisy Pairs** (Black Arrow): We randomly sample a view $(v_{\text{train}}, x^{\text{train}}_{\text{gt}}) \in \Pi_{\text{train}}$ and add noise ϵ into the attributes (position, opacity, color) of G_ϵ.

This progressive pair generation strategy enhances data diversity and realism, enabling the RepairNet R to robustly handle noisy, intermediate, correlated, and degraded views in five-view scenarios.

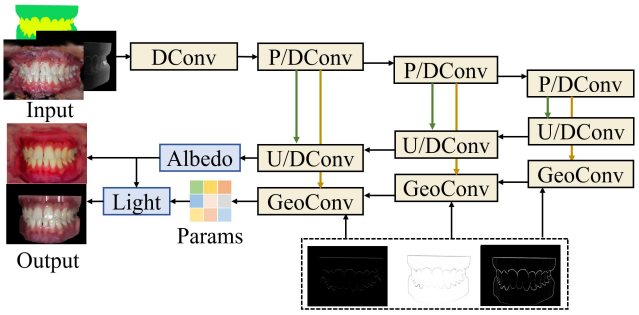

Figure 3: Architecture of the RepairNet: takes degraded renderings, depth and semantics as input, and produces refined images guided by geometric cues.

RepairNet for Rendered Image Restoration. The RepairNet R is a encoder-decoder U-Net with dual branches for illumination-aware and albedo reconstruction (as shown in Stage 3 of Fig.1 and Fig.3 for details). For each view (v, x_{gt}), the R takes $\mathbf{x}_{\text{in}}$ as input the degraded image $x = G(v)$, depth map $Dep(v)$, and semantic segmentation map $Cls(v)$, which provide appearance, geometry, and anatomical context, respectively. To incorporate geometric priors $\mathbf{x}_{\text{cond}}$, we compute pseudo-normal maps $\mathcal{N}$ and depth gradient κ from the depth gradients using Sobel filters:

$$\mathcal{N} = \text{normalize}(-\nabla_x Dep(v), -\nabla_y Dep(v), 1), \quad (1)$$

$$\kappa = |\nabla_x Dep(v)| + |\nabla_y Dep(v)|, \quad (2)$$

$$\mathbf{x}_{\text{in}} = \text{concat}(G(v), Dep(v), Cls(v)) \in \mathbb{R}^{5\times H\times W}, \quad (3)$$

$$\mathbf{x}_{\text{cond}} = (\mathcal{N}, \kappa) \in \mathbb{R}^{4\times H\times W}. \quad (4)$$

The encoder comprises three downsampling stages, each built with a modified Double Conv-group (DConv) and residual connections. The first decoder branch incorporates geometric priors $\mathbf{x}_{\text{cond}}$ through a geometry-aware upsampling module using bilinear DConv and attention fusion. The second branch, designed for albedo prediction, excludes geometric cues.

The two branches output a 2-channel specular parameter map and a 3-channel albedo map, both passed through sigmoid activation. The final shaded result is rendered using predicted reflectance components feature and geometry, based on a physically-inspired formulation combining ambient, diffuse, subsurface, and Cook-Torrance specular terms.

The output is a restored RGB image $\mathbf{r} \in \mathbb{R}^{H\times W\times 3}$ that corrects structural distortions and restores surface details:

$$\mathbf{r} = R(\mathbf{x}_{\text{in}}, \mathbf{x}_{\text{cond}}). \quad (5)$$

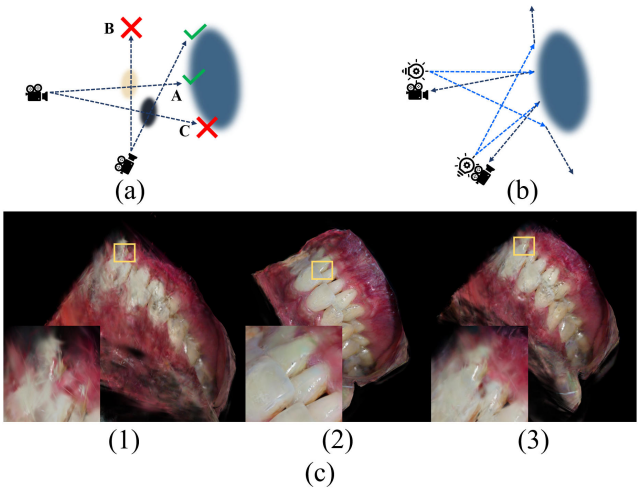

Figure 4: Limitations of traditional Gaussian Splatting under dynamic lighting: (a) causes of failure with changing light sources; (b) need for lighting-aware modeling; (c) examples (1–3) of inconsistent visuals under varying illumination

Lighting-Aware Gaussian

Due to the dim oral environment and reliance on external light sources, there are strong lighting inconsistencies across viewpoints. However, as shown in Fig. 4, traditional 3DGS methods typically assign each Gaussian a set of Spherical Harmonics (SH) coefficients to approximate low-frequency directional radiance. Since radiance is predicted solely as a function of view direction—independent of actual scene lighting or geometry—complex lighting effects cannot be captured. For example, a Gaussian encoding a specular highlight may only be visible from a specific viewpoint (ray A), but due to its physical presence in 3D space, it may incorrectly appear in unrelated views (ray B). Similarly, a Gaussian representing a shadowed region may remain visible even without an occluder (ray C). This illumination-agnostic design causes structural inconsistencies that often result in floating artifacts (ghost Gaussians) reducing overall rendering quality.

To address these issues, inspired by (Zhang et al. 2024), we propose a physically-inspired reflectance model for 3DGS $\mathbf{G}$. Instead of using SH to encode view-dependent color, as shown in Stage 4 of Fig.1, we assign each Gaussian material attributes feature. Given a Gaussian kernel $\mathbf{g}_i$ with attributes parameters position po_i, covariance Σ_i opacity α_i, albedo $\mathbf{c}_i$, normal $\mathbf{n}_i$, roughness ro_i, metallicity m_i and bias feature s_i, that $\mathbf{g}_i = \{po_i, \Sigma_i, \alpha_i, \mathbf{c}_i, \mathbf{n}_i, ro_i, m_i, s_i\} \in \mathbf{G}$. According to the assumption proposed in (Zwicker et al. 2002), we assume that the material properties of a single

Gaussian kernel can represent those of all 3D points within its spatial extent. With Num ordered points for pixel (u, v), the rendering equation then becomes:

$$\hat{L}_{u,v} = \sum_{i \in Num} (Ig_i + s_i) f_r(\mathbf{c}_i, \mathbf{n}_i, r_i, m_i, v, \mathbf{l}) \alpha_i \prod_{j}^{i-1} (1 - \alpha_j) \tag{6}$$

where $\mathbf{l}$ denotes the light directions, I_g denotes the input radiance, and f_r is the reflectance function used to compute $\mathbf{g}_i$, serving as an alternative to SH.

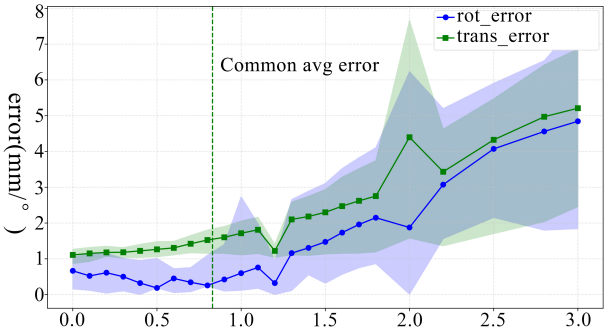

Figure 5: Evaluation of pose estimation accuracy under different levels of initialization noise.

Loss Function

Current lighting models need many images for training, but with only five images, there isn't enough lighting information to learn consistent surface reflections, which hurts the quality of NVS. To address this, we introduce a joint supervision loss that combines sparse GT views x_{gt} with dense restored renderings $\mathbf{r}$ from RepairNet. This design compensates for the limited lighting coverage in GTs and enables the 3DGS to learn more consistent reflectance behavior.

Where GT views x_{gt} are supervised with pixel-wise L1 loss to preserve fine details:

$$\mathcal{L}_{\text{L1}} = \|\hat{x} - x_{\text{gt}}\|_1. \tag{7}$$

For dense restored renderings $\mathbf{r}$, we emphasize structural and perceptual consistency rather than pixel-level accuracy. A structure-preserving loss is applied using local average pooling:

$$\mathcal{L}_{\text{struct}} = \left\| \text{AvgPool}_{5\times5}(\hat{x}) - \text{AvgPool}_{5\times5}(\mathbf{r}) \right\|_1. \tag{8}$$

and high-level feature alignment is enforced via VGG-based perceptual loss:

$$\mathcal{L}_{\text{perc}} = \sum_{l} \|\phi_l(\hat{x}) - \phi_l(\mathbf{r})\|_1, \tag{9}$$

where $\phi_l(\cdot)$ denotes the activation from the l-th layer of a pretrained VGG network.

Additionally, both $\mathbf{r}$ and x_{gt} is encouraged using the SSIM loss to perceptual similarity:

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(\hat{x}, \mathbf{r} \text{ or } x_{\text{gt}}) \tag{10}$$

The total loss is a weighted combination:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{L1}} + \lambda_2 \mathcal{L}_{\text{struct}} + \lambda_3 \mathcal{L}_{\text{ssim}} + \lambda_4 \mathcal{L}_{\text{perc}}. \tag{11}$$

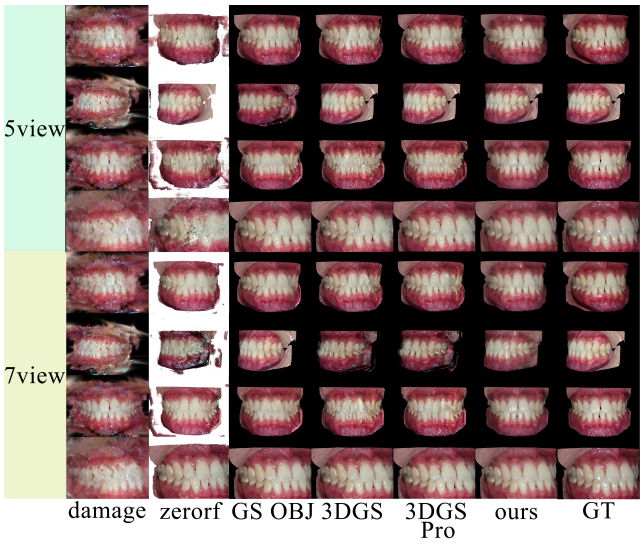

Figure 6: Qualitative results on our dataset under 5-view and 7-view input settings. Many baseline methods fail to capture clear geometric structures, resulting in noticeable blurring and artifacts.

Experiments

DataSets and Implementation Details

To support research objectives, we construct a dataset, which contains 20 complete 3D jaw models (upper and lower) collected from multiple hospitals, acquired via intraoral scanners and stored as surface meshes. Tooth regions are segmented using the method from (Zhuang et al. 2023), and corresponding five-view photographs (5616×3744) are semantically segmented using SAM (Kirillov et al. 2023). To simulate remote healthcare scenarios, we collect intraoral images captured by smartphones two years after the initial scans (4096×3072). For geometric initialization, we preprocess the original meshes using MCF (Xie et al. 2024).

Evaluation

Camera pose. To evaluate the robustness of our camera pose estimation method, we conducted a stress test by adding controlled perturbations to the initial tooth point cloud. Each perturbation level corresponds to a 0.5mm increase in translation error and a $15/\pi$degree increase in rotation error, with all rotation values reported in degrees. This mapping was calibrated using the empirical distribution of initial misalignment in clinical post-orthodontic cases, ensuring that Levels 1–3 cover the typical clinical range.

In such typical scenarios, the initial deviation reported by (Xie et al. 2024) (0.415 mm and 3.52°) falls at approximately Level 0.83 under this calibration. Within this range, our method achieves an average rotation error of 0.4294° and an average translation error of 1.2625 mm. Even under substantially larger perturbations, translation error remains around 5 mm and rotation error within 4°, demonstrating the robustness of our approach.

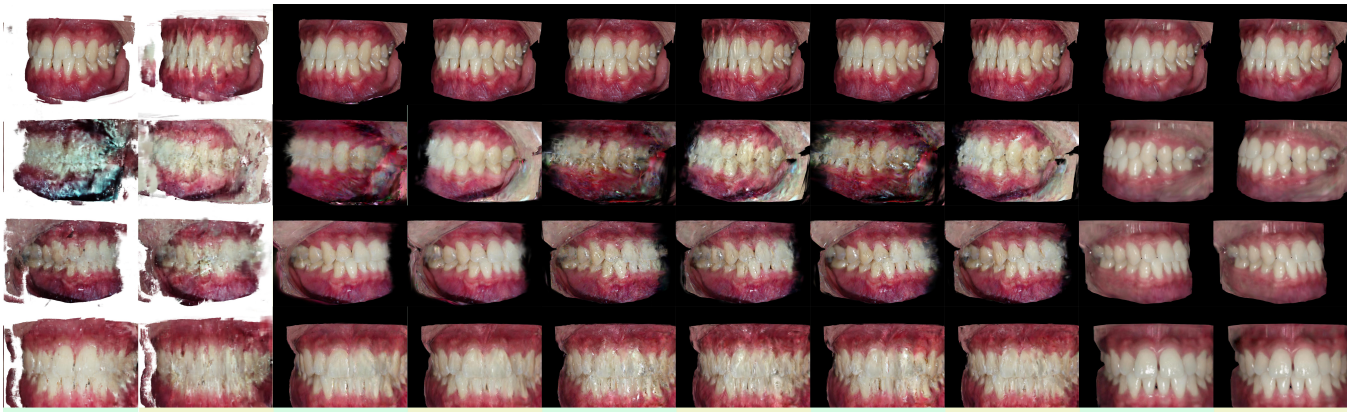

Figure 7: Qualitative results on our dataset under 5-view and 7-view input settings, including several extreme viewpoints that are difficult to capture.

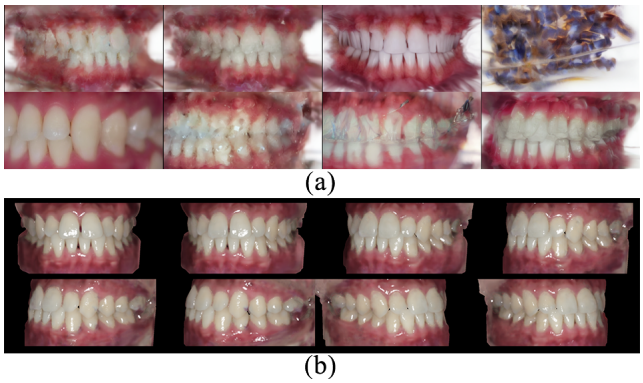

Figure 8: Restoration results using SDM(a) and RepairNet(b) on Gaussian renderings from multiple viewpoints.

Comparative experiment. We compare our DentalGS with several sparse-view and depth-based 3DGS baselines, including 3DGS (Kerbl et al. 2023), 3DGSPro (Cheng et al. 2024), Gaussian Object(GSOBJ) (Yang et al. 2024), and ZeroRF (Shi et al. 2024). All baselines use our estimated camera poses and are trained with the same settings for reproducibility.

Table 1 compares DentalGS with other methods on NVS. PSNR measures pixel-level reconstruction error, SSIM assesses structural similarity in luminance, contrast, and geometry, and LPIPS evaluates perceptual differences using deep features. GSOBJ leverages a large diffusion model (SDM) for image restoration, while $GSOBJ_{\text{base}}$ denotes its results before repair. Under extremely sparse input views, DentalGS consistently outperforms all methods—especially in SSIM—highlighting its superior capability in preserving geometric structure.

Fig. 6 and 7 present NVS results under different input conditions. Fig. 6 illustrates results from common views, while Fig. 7 includes extreme angles that are difficult to capture physically. Our method consistently delivers higher visual quality—preserving tooth geometry, reducing Gaussian artifacts, and producing more natural highlights.

Figure 9: The impact of various input cues on the RepairNet.

Interestingly, Fig. 6 shows that increasing the number of input views may degrade performance. As discussed in Section 3, additional views can introduce inconsistent lighting (e.g., specularities, shadows), which disrupts Gaussian coherence and reduces rendering quality. Unlike other methods that exhibit noticeable quality drops with more inputs, our approach remains stable. Even with limited front-facing views, DentalGS reconstructs complete geometry and coherent textures, demonstrating strong robustness and generalization under five-view settings.

Unet and Diffusion. While GSOBJ achieves favorable LPIPS due to SDM's strong perceptual prior in latent space,as shown in Fig.8(a), replacing the UNet with SDM in the image restoration module introduces noticeable structural inconsistencies across different input viewpoints. . This is further supported by the PSNR drop from 21.43 (pre-restoration) to 20.91 (post-restoration), indicating SDM's limitations in rigid-object modeling like teeth. In contrast, our UNet-based restoration (Fig. 8(b)) better maintains geometric consistency and achieves higher SSIM, producing visually plausible and structurally reliable results.

In summary, although pretrained diffusion models like SDM excel in photorealism, their stochastic nature and lack of domain-specific semantic understanding make them less suitable for high-geometric 3D reconstruction tasks without

Method-5view	PSNR↑	SSIM↑	LPIPS↓	Method-7view	PSNR↑	SSIM↑	LPIPS↓
$GSpro_5$	19.42	0.8136	0.2877	$GSpro_7$	24.26	0.8528	0.2622
GS_5	19.71	0.8083	0.2935	GS_7	24.02	0.8480	0.2699
$ZeroRF_5$	15.06	0.7147	0.3691	$ZeroRF_7$	15.76	0.6901	0.4316
$GSOBJ_{base}$	21.43	0.7929	0.2566	$GSOBJ_{base}$	21.43	0.7929	0.2566
$GSOBJ_5$	20.91	0.8044	0.2438	$GSOBJ_7$	24.71	0.8559	0.1938
$ours_5$	*26.20*	*0.9358*	0.2689	$ours_7$	**26.86**	**0.9381**	*0.2281*

Table 1: Quantitative Evaluation of DentalGS Compared to Other Methods

Config	$\mathcal{L}_{\text{L1}}$	$\mathcal{L}_{\text{perc}}$	$\mathcal{L}_{\text{struct}}$	$\mathcal{L}_{\text{ssim}}$	PSNR↑	SSIM↑	LPIPS↓
L_L	✓	×	×	×	24.12	0.8710	0.2770
L_H	✓	×	×	×			
L_L	✓	✓	×	×	24.45	0.8819	0.2470
L_H	✓	✓	×	×			
L_L	✓	✓	✓	×	24.54	0.8822	0.2461
L_H	✓	✓	✓	×			
L_L	✓	✓	✓	✓	24.84	0.8880	0.2454
L_H	✓	✓	✓	✓			
L_L	×	✓	✓	✓	26.20	0.9358	0.2689
L_H	✓	×	×	✓			
L_L	×	✓	✓	✓	26.89	0.9027	0.2294
L_H	✓	✓	×	✓			

Table 2: Ablation Study on the Effect of Different Loss Functions

targeted fine-tuning.

Ablation Studies

We conducted a series of ablation studies to evaluate the effectiveness of the proposed components. In all experiments, the number of input views was fixed at 5.

Loss Function. We progressively removed different loss terms to assess their individual contributions to overall performance. Here, L_L denotes the loss applied to low-quality images, while L_H corresponds to the loss used for high-quality images.

The results show that introducing spatial-aware losses significantly improves the LPIPS metric. Low-quality images contribute more effectively to structural similarity, enhancing the model's perception of global geometry. However, fine-grained noise and artifacts in these low-quality images may interfere with the Gaussian representation and negatively impact reconstruction quality. In contrast, spatial-aware losses from high-quality images can harm structural similarity, which is consistent with earlier observations of geometry distortion introduced by SDM.

Therefore, adaptively selecting different loss functions based on input image quality proves beneficial. This strategy helps preserve structural integrity while mitigating the adverse effects of noisy details, ultimately leading to significantly improved synthesis performance.

RepairNet. We conduct ablation studies by removing different inputs to RepairNet, as shown in Fig. 9. Using the damage image(c) significantly simplifies the restoration task and enhances realism. The depth map(d) provides effective geometric cues, while the class map(b) acts as a bridge between the damage and depth maps, reducing the impact of noise. Together, these inputs complement each other and collectively improve the performance of RepairNet. Similar to camera pose errors, discrepancies between the predicted and ground-truth point clouds may impact RepairNet's performance. We evaluate this through experiments, showing that within acceptable error bounds (Xie et al. 2024), restored images remain high-quality, with an average max L1 of 0.0499, minimum PSNR of 23.38, and minimum SSIM of 0.8959.

Conclusion

In summary, we propose **DentalGS**, an enhanced 3D Gaussian Splatting framework tailored for novel view synthesis using five post-orthodontic intraoral images and pre-orthodontic IOS data as a prior, without requiring camera poses, enabling remote orthodontic monitoring. The method initializes the Gaussian point cloud using pre-orthodontic IOS and post-orthodontic images, and estimates camera poses through iterative optimization. Next, a progressive pair generation strategy is used to train RepairNet, which restores rendering. A reflection-based, physics-inspired 3DGS is further introduced to mitigate illumination variations. Experimental results show that DentalGS effectively preserves geometric structure and handles extreme viewpoints, offering a robust and efficient solution for 3D dental visualization in remote orthodontic monitoring.

Limitations and Future Work

Although DentalGS performs well in NVS with sparse intraoral views, several challenges remain. First, the reliance on predicted IOS data introduces errors compared to real teeth, affecting camera parameter estimation and depth cues, which in turn impacts novel view synthesis accuracy.

In the future, as the dataset expands, the restoration model could evolve into a pretrained feedforward network, enabling high-quality Gaussian reconstruction without retraining or with only minimal fine-tuning. Furthermore, combining DentalGS with methods such as (Qian et al. 2024) may enable highly realistic, dynamic 3DGS of the orthodontic process, offering a promising solution for visualizing orthodontic treatment in a clinically meaningful way.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62572284, 62502273 (Youth Project), and 62172257; and by the Natural Science Foundation of Shandong Province (Major Basic Research)

under Grant ZR2024ZD12, and the Natural Science Foundation of Shandong Province (Youth Project) under Grant ZR2025QC1558.

References

Chang, K.-W.; Wang, Z.-M.; and Lai, S.-H. 2025. KeyGS: A Keyframe-Centric Gaussian Splatting Method for Monocular Image Sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1989–1997.

Chen, Y.; Gao, S.; Tu, P.; and Chen, X. 2023. Automatic 3d teeth reconstruction from five intra-oral photos using parametric teeth model. *IEEE Transactions on Visualization and Computer Graphics*, 30(8): 4780–4791.

Cheng, K.; Long, X.; Yang, K.; Yao, Y.; Yin, W.; Ma, Y.; Wang, W.; and Chen, X. 2024. Gaussianpro: 3d gaussian splatting with progressive propagation. In *Forty-first International Conference on Machine Learning*.

Dong, S.; Wang, S.; Liu, S.; Cai, L.; Fan, Q.; Kannala, J.; and Yang, Y. 2025. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 16739–16752.

Gao, J.; Gu, C.; Lin, Y.; Li, Z.; Zhu, H.; Cao, X.; Zhang, L.; and Yao, Y. 2024. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*, 73–89. Springer.

Guo, J.; Wang, J.; Kang, D.; Dong, W.; Wang, W.; and Liu, Y.-h. 2024. Free-surgs: Sfm-free 3d gaussian splatting for surgical scene reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 350–360. Springer.

Han, H.; Zhou, K.; Long, X.; Wang, Y.; and Xiao, C. 2025. Ggs: Generalizable gaussian splatting for lane switching in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3329–3337.

Hong-Seok, P.; and Chintal, S. 2015. Development of high speed and high accuracy 3D dental intra oral scanner. *Procedia Engineering*, 100: 1174–1181.

Hou, X.; Li, M.; Yang, D.; Chen, J.; Qian, Z.; Zhao, X.; Jiang, Y.; Wei, J.; Xu, Q.; and Zhang, L. 2025. BloomScene: Lightweight structured 3D gaussian splatting for crossmodal scene generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3536–3544.

Jiang, Y.; Tu, J.; Liu, Y.; Gao, X.; Long, X.; Wang, W.; and Ma, Y. 2024. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5322–5332.

Kapila, S.; Conley, R.; and Harrell Jr, W. 2011. The current status of cone beam computed tomography imaging in orthodontics. *Dentomaxillofacial Radiology*, 40(1): 24–34.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kim, S. 2023. DäRF: Boosting Radiance Fields from Sparse Inputs with Monocular Depth Adaptation. In *Neural Information Processing Systems (NeurIPS)*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Lepetit, V.; Moreno-Noguer, F.; and Fua, P. 2009. EP n P: An accurate O (n) solution to the P n P problem. *International journal of computer vision*, 81(2): 155–166.

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(1): 99–106.

Qian, S.; Kirschstein, T.; Schoneveld, L.; Davoli, D.; Giebenhain, S.; and Nießner, M. 2024. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20299–20309.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Shi, R.; Wei, X.; Wang, C.; and Su, H. 2024. Zerorf: Fast sparse view 360deg reconstruction with zero pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21114–21124.

Wang, G.; Chen, Z.; Loy, C. C.; and Liu, Z. 2023. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9065–9076.

Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; and Novotny, D. 2025a. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.

Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; and Revaud, J. 2024. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

Wang, Y.; Sun, X.; Jia, J.; Jin, Z.; and Ma, Y. 2025b. High-precision 3D teeth reconstruction based on five-view intra-oral photos. *Displays*, 87: 102988.

Xie, J.; Zhang, C.; Wei, G.; Wang, P.; Wei, G.; Liu, W.; Gu, M.; Luo, P.; and Wang, W. 2024. Tooth Motion Monitoring in Orthodontic Treatment by Mobile Device-based Multi-view Stereo. *IEEE Transactions on Visualization and Computer Graphics*.

Xu, C.; Liu, Z.; Liu, Y.; Dou, Y.; Wu, J.; Wang, J.; Wang, M.; Shen, D.; and Cui, Z. 2024. TeethDreamer: 3D teeth reconstruction from five intra-oral photographs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 712–721. Springer.

Xu, D.; Jiang, Y.; Wang, P.; Fan, Z.; Shi, H.; and Wang, Z. 2022. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vision*, 736–753. Springer.

Yang, C.; Li, S.; Fang, J.; Liang, R.; Xie, L.; Zhang, X.; Shen, W.; and Tian, Q. 2024. Gaussianobject: High-quality 3d object reconstruction from four views with gaussian splatting. *arXiv preprint arXiv:2402.10259*.

Zhang, T.; Huang, K.; Zhi, W.; and Johnson-Roberson, M. 2024. Darkgs: Learning neural illumination and 3d gaussians relighting for robotic exploration in the dark. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 12864–12871. IEEE.

Zhang, W.; Zhang, L.; Hu, P.; Ma, L.; Zhuge, Y.; and Lu, H. 2025. Bootstraping clustering of gaussians for view-consistent 3d scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10166–10175.

Zhuang, S.; Wei, G.; Cui, Z.; and Zhou, Y. 2023. Robust hybrid learning for automatic teeth segmentation and labeling on 3D dental models. *IEEE Transactions on Multimedia*, 27: 792–803.

Zwicker, M.; Pfister, H.; Van Baar, J.; and Gross, M. 2002. EWA splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(3): 223–238.