DENTAL 3D RECONSTRUCTION WITH ADVANCED KEYPOINT DETECTION AND SURFACE-ALIGNED GAUSSIAN SPLATTING

Bohdan Vodianyk [1], Enrique Nava Baro [2], Alfonso Ariza Quintana [2], Anton Popov [3,4]

[1] Escuela de Ingenierías Industriales, Universidad de Málaga, Arquitecto Francisco Penalosa, 6, Malaga, 29071, Spain
[2] ETSI Telecomunicación, Universidad de Málaga, Blvr. Louis Pasteur, 35, Malaga, 29010, Spain
[3] Department of Electronic Engineering, Igor Sikorsky Kyiv Polytechnic Institute, Polytekhnichna Street, 16, Kyiv, 03056, Ukraine
[4] Faculty of Applied Sciences, Ukrainian Catholic University, Kozelnytska Street, 2a, Lviv, 79026, Ukraine

ABSTRACT

Accurate 3D reconstruction of dental structures is crucial for orthodontic assessment and surgical planning, yet traditional methods such as SIFT and ORB often struggle to capture fine details in complex dental textures. In this paper, we present a 3D reconstruction pipeline that combines KeyNetAffNetHardNet for feature detection and matching with Surface-Aligned Gaussian Splatting (SuGaR) for high-quality mesh reconstruction. By leveraging KeyNet for robust keypoint identification, AffNet for affine normalization, and HardNet for discriminative descriptors, our approach achieves a 25% reduction in computation time compared to advanced deep learning methods like LoFTR and DISK + LightGlue. To further optimize the 3D meshes, SuGaR aligns surface Gaussians to actual geometry, improving both structural accuracy and rendering fidelity. A new pipeline was evaluated using a set of high-resolution video frames from a single participant's dental panorama, achieving peak SSIM and PSNR scores of 0.9538 and 28.98, respectively — improvements of approximately 10% and 15% over conventional approaches. Our findings highlight how integrating learned feature matching and surface-aligned reconstruction can yield high-fidelity 3D dental models while maintaining efficiency, ultimately advancing diagnostic precision and treatment outcomes in dentistry.

KEYWORDS

3D Reconstruction, Keypoint Matching, Gaussian Splatting, Dental Imaging, Deep Learning, Computer Vision

1. INTRODUCTION

Accurate 3D reconstruction of dental structures plays a vital role in modern dentistry, assisting in orthodontic assessment, implant planning, and maxillofacial surgery. For instance, clinicians can employ the resulting 3D models for orthodontic bracket placement or planning complex implant surgeries with improved accuracy. High-fidelity 3D models provide precise insights into dental anatomies, enabling better treatment decisions and improved patient outcomes. However,

David C. Wyld et al. (Eds): Signal, CDKP, AdNLP, MLT, SAI – 2025
pp. 01-16, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.150901

achieving high-quality reconstructions from dental imagery is challenging due to complex textures, low-contrast regions, and occlusions caused by soft tissues and variable lighting conditions.

Traditional feature detection methods such as SIFT and ORB often fail to capture the fine details needed for robust 3D reconstruction. While deep learning–based feature matching approaches, including LoFTR and DISK + LightGlue, can provide notable improvements, they tend to be computationally expensive, which may limit their applicability in real-time clinical settings. In response to these issues, this paper proposes a 3D reconstruction pipeline that integrates KeyNetAffNetHardNet for feature detection and matching, coupled with Surface-Aligned Gaussian Splatting (SuGaR) for efficient and high-quality mesh reconstruction. By applying KeyNet for discerning salient keypoints, AffNet for robust affine normalization, and HardNet for discriminative descriptors, the pipeline can capture precise correspondences in scenes characterized by reflective surfaces and repetitive patterns. Meanwhile, SuGaR improves structural accuracy by aligning surface Gaussians with the underlying 3D geometry, resulting in more faithful dental meshes.

Although our experiments demonstrate strong results on data collected from a single participant, we recognize that working with a limited dataset may constrain the generalizability of our approach. To address broader clinical scenarios, future research will include multiple subjects, more diverse imaging conditions. Even in its current form, however, the proposed pipeline shows that accurate 3D models can be generated using standard photographic equipment, potentially reducing hardware costs and enhancing accessibility in dental practices.

2. RELATED WORK

Accurate 3D reconstruction remains a core challenge in computer vision, and dental imagery poses especially difficult conditions due to frequent occlusions, complex textures, and low-texture regions. Traditional feature detectors such as SIFT [1] and ORB [2] provide basic robustness to scale and rotation but often fail to detect enough reliable keypoints on smooth or repetitive surfaces typically found in intraoral images [3, 4]. Deep learning approaches, such as LoFTR [5] and DISK combined with LightGlue [6, 7], rely on dense matching or attention mechanisms to overcome low-texture problems but require considerable computational resources, potentially hindering live clinical use.

In contrast, KeyNetAffNetHardNet effectively unifies three specialized components. KeyNet focuses on detecting salient features even in dental imagery with reflective enamel or low contrast, AffNet normalizes patches to handle large viewpoint variations, and HardNet creates robust descriptors that mitigate mismatches in repetitive or glossy areas [8, 9]. Together, these steps produce consistent keypoint correspondences in scenarios where other methods struggle. To generate high-fidelity 3D meshes from these matched points, our pipeline employs SuGaR [10], which aligns surface Gaussians through a regularization step that is particularly helpful for reconstructing curved dental surfaces. Although earlier studies addressed photogrammetry-based dental reconstruction [11, 12] or specialized scanning devices [13], the combined use of advanced feature matching and Gaussian Splatting can provide a more accessible and cost-effective solution. By pairing learned descriptors with an efficient, geometry-aligned reconstruction strategy, this method captures both overall tooth contours and fine-grained surface details that are crucial for clinical evaluations.

3. METHODS

This study developed and evaluated a 3D reconstruction pipeline tailored for dental imagery, addressing the unique challenges posed by complex textures, repetitive patterns, and low-texture regions inherent in dental photographs. The methodology encompasses data acquisition, keypoint detection and matching, 3D reconstruction using Gaussian Splatting and SuGaR, and evaluation using quantitative metrics.

3.1. Experiment Design

To create a comprehensive dataset for the experiments, three high-definition videos of a single participant's dental panorama were recorded, designated as Experiment IDs 1, 2, and 3. Specialized dental retractors were used during filming to expose the maximum surface area of the teeth, minimising occlusions from lips and cheeks and providing unobstructed views crucial for accurate 3D reconstruction. The camera followed an ellipsoidal trajectory (Figure 1) around the oral cavity to effectively cover the main regions of the teeth.

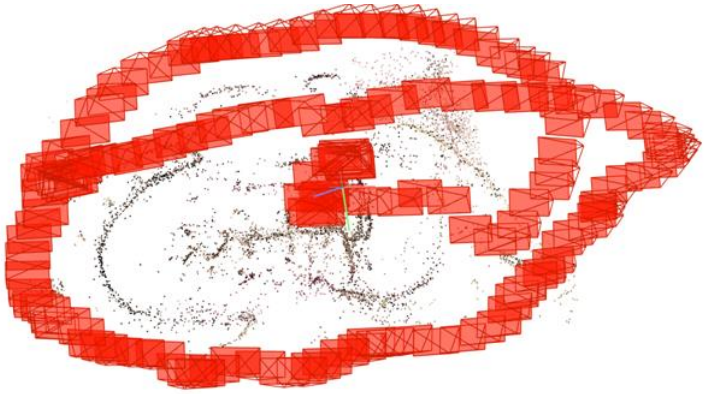


Figure 1. Camera trajectory in COLMAP visualization [14].

3.2. Dataset Description

Each video was captured at a resolution of 1920×1080 pixels (Full HD) with a frame rate of 30 frames per second (FPS) and a duration of approximately 25 to 30 seconds. To assess the robustness of the reconstruction pipeline under varying data densities, frames were extracted from each video at intervals corresponding to every 5, 10, 15, 20, 25, and 30 frames, resulting in different sets of images for each experiment. This approach generated multiple datasets with varying numbers of frames, simulating different sampling conditions and allowing evaluation of the pipeline's performance with respect to data sparsity.

For instance, extracting every 5th frame from a 25-second video at 30 FPS yields a dataset of approximately 150 images, while extracting every 30th frame results in a dataset of about 25 images. This systematic variation in frame extraction intervals enabled testing how temporal spacing between frames affects keypoint detection, matching, and ultimately, the quality of the 3D reconstruction.

3.3. Pairwise Matching

Before proceeding with the 3D reconstruction, an initial evaluation of keypoint detection and matching performance for the different methods was conducted. This step aimed to assess the effectiveness of each method in establishing reliable correspondences under varying conditions.

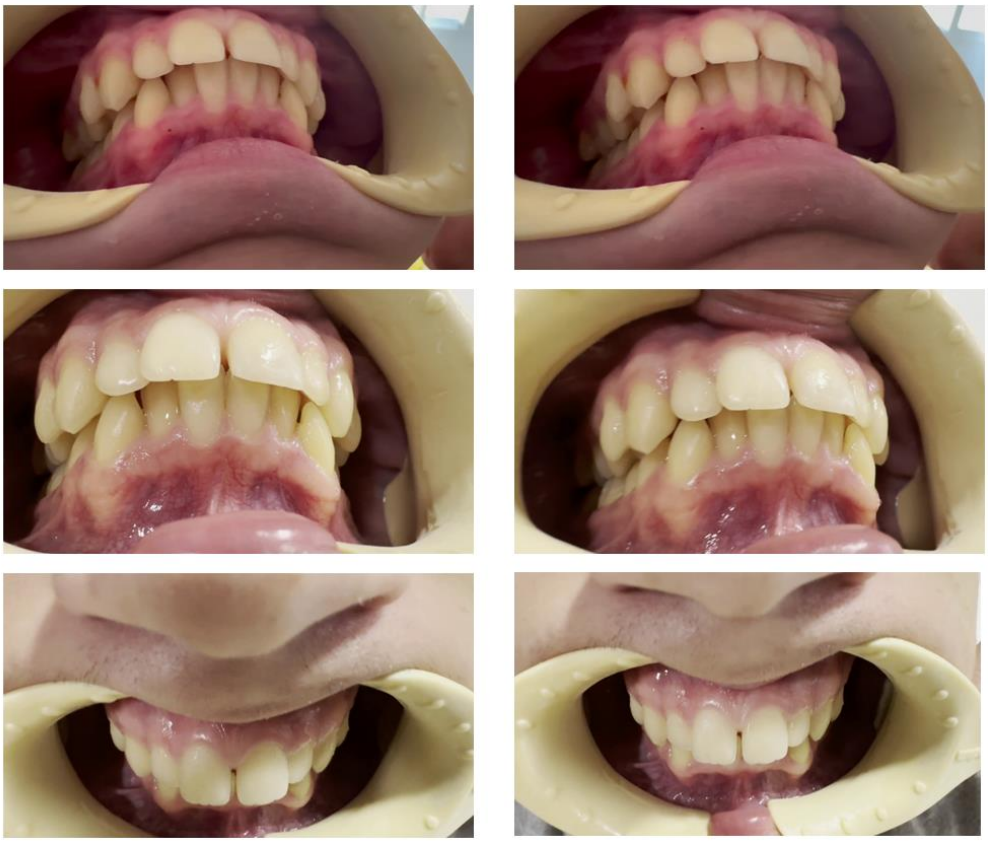


Figure 2. Pairs samples from each experiment video of the dental panorama.

For each image set extracted at different frame intervals, initially, two neighboring frames were selected (e.g., from video 2 with a frame gap of 10 frames) and applied keypoint detection and matching using the following methods:

- **SIFT (Scale-Invariant Feature Transform)** is a classical feature detection and description algorithm renowned for its robustness to changes in scale, rotation, and illumination. SIFT operates by detecting keypoints in scale space using the Difference-of-Gaussian Method [1]. It identifies extrema that are invariant to scale and orientation by locating peaks in the DoG pyramid, which is constructed by subtracting successive Gaussian-blurred images at different scales [15].
- **ORB (Oriented FAST and Rotated BRIEF)** is a feature detection and description algorithm that is an efficient alternative to more computationally intensive methods like SIFT. ORB combines the FAST (Features from Accelerated Segment Test) keypoint detector with the BRIEF (Binary Robust Independent Elementary Features) descriptor, enhancing them to provide orientation invariance and robustness to rotation [2].
- **LoFTR (Local Feature TRansformer)** is a deep learning-based method that performs dense matching without relying on explicit keypoint detection and descriptor computation. In order to model long-range dependencies and contextual information

across images, it uses a transformer architecture, thereby also being capable of finding pixel-wise correspondences in scenarios that are difficult as low texture or repetitive pattern regions [16]. LoFTR directly outputs dense matches in pixel coordinates by being given pairs of images. It combines two modules, a CNN backbone for feature extraction, and a transformer module for feature correlation, then a matching head. Self-attention mechanisms used by the transformer module make it possible to capture the global context which proves especially useful to capture correspondences that are not locally distinctive [5].

- **KeyNetAffNetHardNet** is an integrated framework that combines three deep learning-based models to enhance keypoint detection, affine shape estimation, and descriptor generation, respectively. This combination is particularly effective in challenging imaging conditions, as it prioritizes discriminability and robustness in keypoint matching.

  KeyNet is a convolutional neural network designed to detect salient keypoints that are both repeatable and discriminative [17]. It learns to focus on regions with rich structural information by training on datasets where keypoint locations are annotated. KeyNet employs a combination of handcrafted and learned filters to balance computational efficiency and detection performance [18].

  AffNet estimates the local affine transformation around each keypoint, effectively normalising the keypoints to a canonical form [19, 20]. Robustness to viewpoint changes, scale variations, and imaging distortions — common in the dataset due to dental retractors and non-frontal camera angles — is also improved by this affine adaptation. Given the output of KeyNet the patches around each detected keypoint are extracted and affine parameters that describe the local shape are predicted using AffNet. The estimated affine transformation is used to warp the patches to a canonical shape so that a consistent descriptor can be computed [21].

  HardNet is a CNN-based descriptor that generates compact and highly discriminative feature descriptors [17]. It is trained with a triplet margin loss to maximize the distance between descriptors of different keypoints while minimising the distance between descriptors of the same keypoint in different images.

- **DISK + LightGlue** leverages deep learning for both keypoint detection and feature matching, aiming to improve performance in challenging conditions.

  The deep learning-based method DISK jointly learns keypoint detection and descriptor generation in an end-to-end fashion. However, it tries to produce repeatable and highly informative keypoints for matching. DISK is trained with reinforcement learning to choose optimal keypoint locations and descriptors from a model trained for a downstream task [6].

  LightGlue is an advanced feature matcher benefiting from attention mechanisms to improve feature-matching speed and robustness. It uses facets of DISK descriptors and matches on a transformer-like architecture that incorporates local and global context. Given two images, LightGlue's descriptors are processed to compute a matching matrix that contains the similarity of each descriptor. The matching scores are refined by using self-attention and cross-attention layers that allow it to handle difficult matching scenarios like repetitive patterns and large viewpoint changes [7, 22].

The number of matched inliers for each method was measured by applying geometric verification using MAGSAC [23] to eliminate outliers. This evaluation provided insights into each method's ability to handle typical frame-to-frame variations in dental videos. To increase the difficulty of the keypoint matching algorithms, three frame pairs were selected from each video, ensuring they contained the largest differences in camera position and viewpoint. Due to significant viewpoint changes, occlusions, and differences in illumination, these pairs were the most difficult to match features in. For these pairs, the same keypoint detection and matching methods were used, and the number of matched inliers was counted after geometric verification. This step was enabled to determine how robust each method was under adverse conditions.

3.4. Image Processing Pipeline

The entire computational process was executed on a workstation equipped with an NVIDIA RTX A2000 12GB.

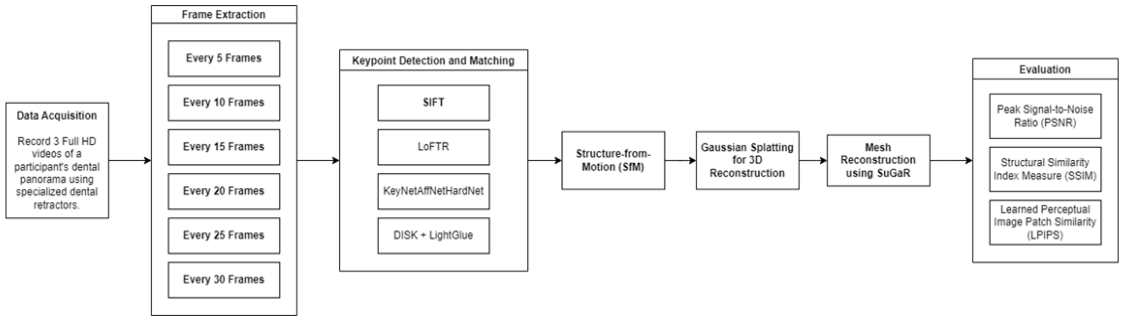


Figure 3. Block diagram of full 3D reconstruction investigation.

Figure 3 illustrates our complete reconstruction pipeline. In the first stage, we detect and match keypoints using KeyNetAffNetHardNet, SIFT, ORB, LoFTR, or DISK + LightGlue. Next, a Structure-from-Motion (SfM) [24, 25] step estimates camera poses and generates a sparse point cloud, which we refine using Gaussian Splatting. Finally, SuGaR enhances mesh quality by aligning surface Gaussians with the underlying geometry, resulting in more accurate and visually coherent dental models.

Firstly, various feature detection and matching algorithms were applied to the extracted frames. The methods evaluated were: SIFT, ORB, LoFTR, KeyNetAffNetHardNet and DISK + LightGlue. Using the matched keypoints from each method, SfM was performed to estimate camera poses and reconstruct initial sparse 3D point clouds. This involved triangulating matched keypoints to generate 3D points and refining the camera parameters and point positions through sparse bundle adjustment, minimising the overall reprojection error across all images. Then Gaussian Splatting was applied to the sparse point clouds, running 7.000 iterations to optimize the positions, orientations, and appearances of the 3D Gaussians representing the scene geometry.

For mesh reconstruction, SuGaR was employed, extending Gaussian Splatting by introducing a regularization term that aligns the Gaussians with the surfaces of the dental structures. This alignment is achieved by minimising the difference between the actual and ideal Depth-Normal consistency regularizer (dn_consistency) [26, 10] of the scene, under the assumption that the Gaussians are flat and distributed across the surface. A total of 12.000 iterations were executed using the "dn_consistency" method for mesh reconstruction, followed by an additional 4.000 iterations for refinement. This process resulted in high-quality meshes that accurately captured

the intricate geometries of the dental anatomy. Subsequently, a Taubin filter [27] was applied to smooth the resulting mesh, using parameters $\lambda = 0.5$, $\mu = -0.53$, and 10 iterations.

3.5. Evaluation Metrics

To quantitatively assess the quality of the reconstructed 3D models, the following standard metrics [28] were employed:

- **Peak Signal-to-Noise Ratio (PSNR)**: Measures the fidelity of reconstructed images compared to the input multi-view images, with higher values, indicating better reconstruction fidelity [29].
- **Structural Similarity Index Measure (SSIM)**: Evaluates the structural similarity between images, considering luminance, contrast, and structure. SSIM values range from 0 to 1, with higher values indicating greater similarity [30].
- **Learned Perceptual Image Patch Similarity (LPIPS)**: Uses deep neural networks to evaluate perceptual differences between images, with lower values indicating more perceptually similar images [31].

4. RESULTS

4.1. Pairwise Matching

The initial evaluation aimed to identify the most effective way to establish reliable correspondences in dental imagery, especially given reflective surfaces and partial occlusions. Five techniques — ORB, SIFT, LoFTR, KeyNetAffNetHardNet, and DISK + LightGlue — were compared with attention to matching performance, computational efficiency, and robustness (Table 1).

ORB applies a FAST corner detector and BRIEF descriptor for high speed but struggles on smooth or repetitive dental surfaces. SIFT finds scale-invariant features using a Difference-of-Gaussians pyramid, yet can still miss keypoints in low-contrast intraoral views. LoFTR bypasses explicit keypoint detection with a transformer-based, dense matching approach, while DISK + LightGlue combines learned keypoint detection and attention-guided descriptor matching. These deep-learning methods handle low-texture regions but can be computationally demanding.

By contrast, KeyNetAffNetHardNet unifies learned keypoint detection (KeyNet), affine normalization (AffNet), and robust descriptor generation (HardNet). This integrated design is well-suited to reflective enamel and subtle texture variations, ensuring repeatable keypoints and fewer false matches even under large viewpoint changes. Our tests across multiple frame extraction intervals show that KeyNetAffNetHardNet consistently achieves a good balance of inlier matches, memory usage, and overall stability, making it the strongest candidate for subsequent 3D reconstruction in a dental context.

Table 1. Comparison of ORB, SIFT, LoFTR, KeyNetAffNetHardNet, and DISK + LightGlue for pairwise matching under different conditions (**bold** numbers are best).

		SIFT	ORB	LoFTR (0.65 size)	KeyNetAffNetHardNet	DISK + LightGlue
		1.76 GB	1.71 GB	6.76 GB	3.50 GB	9.42 GB
Exp ID	Frames	Number of Matched Inliers				
1	5	58	36	**2147**	283	349
	10	30	16	**326**	91	138
	15	20	12	**208**	154	144
	20	18	8	**147**	102	102
	25	13	11	84	**103**	121
	30	10	0	**40**	36	30
2	5	99	75	**2029**	751	669
	10	82	64	**1953**	835	642
	15	90	48	**1380**	404	253
	20	21	13	71	**120**	85
	25	36	25	**152**	56	45
	30	14	10	**132**	71	49
3	5	92	137	**1919**	556	307
	10	29	25	**317**	195	93
	15	85	38	**811**	322	222
	20	50	37	**1051**	452	163
	25	8	0	17	**53**	16
	30	13	13	18	**40**	20
Extreme	1	0	11	7	**19**	19
	2	8	0	8	**72**	29
	3	10	0	10	**28**	17

When applying pairwise matching of neighboring frames extracted from the videos, ORB exhibited the lowest matching rate among all methods. Although ORB required minimal memory usage, approximately 1.71 Gb of RAM (Random Access Memory), it failed to provide a sufficient number of inlier matches after geometric verification. SIFT performed slightly better than ORB, with a marginally higher matching rate and a similar memory footprint of about 1.76 Gb of RAM. However, both traditional methods were outperformed by their deep learning-based counterparts in terms of matching accuracy and robustness.

LoFTR achieved the highest matching rate in the neighboring frames scenario, indicating excellent performance in establishing correspondences. However, this advantage came with significant computational costs. LoFTR required approximately 6.76 Gb of RAM and operated exclusively on grayscale images. To accommodate memory constraints, the size of the input images was reduced to 65% of their original resolution, which may have compromised the level of detail necessary for accurate 3D reconstruction.

KeyNetAffNetHardNet emerged as a strong contender, providing very robust matching results while maintaining a reasonable memory usage of around 3.50 Gb of RAM. This method demonstrated a favorable balance between matching performance and computational efficiency, making it suitable for practical applications where resources are limited. DISK + LightGlue ranked third in matching performance, offering good results but at the cost of high memory consumption, approximately 9.42 Gb of RAM, which could be prohibitive in resource-constrained environments.

To further challenge the robustness of these methods, tests were conducted on frame pairs with extreme differences in camera positions and viewpoints. In these challenging conditions, only KeyNetAffNetHardNet maintained robust matching performance, consistently providing a sufficient number of inlier matches for reliable camera pose estimation. The other methods, including LoFTR and DISK + LightGlue, struggled significantly, failing to produce enough matches due to the substantial changes in perspective and occlusions present in the dental images.

The evaluation demonstrated that ORB consistently yielded an order of magnitude fewer matched inliers compared to other methods, both in neighboring frames and in frames with extreme viewpoint differences. ORB's binary descriptors are computationally efficient but lack discriminability in the complex and repeated textures of dental imagery. The small number of inliers suggests that there are not enough reliable correspondences for accurate camera pose estimation and 3D reconstruction.

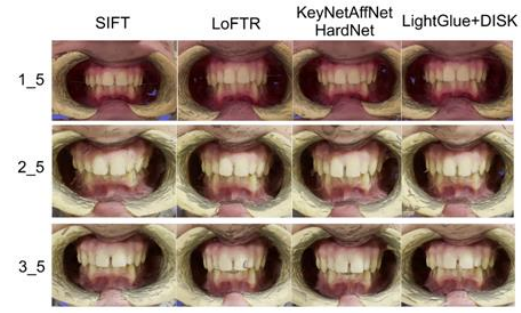
Due to ORB's poor performance in keypoint matching, it was disregarded in the subsequent 3D reconstruction process using Gaussian Splatting. As there are not many reliable matches, including ORB, it would most likely lead to false camera poses and poor reconstruction quality. The focus was placed on the methods that demonstrated satisfactory performance in the initial evaluation.

4.2. Full 3D Reconstruction of Dental Structures

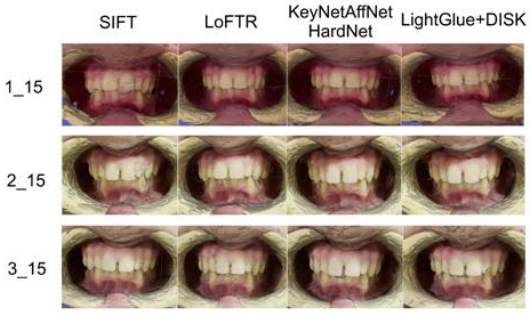
After evaluating the keypoint detection and matching methods, the main reconstruction pipeline was applied, utilising Gaussian Splatting and SuGaR for 3D reconstruction. The goal was to assess how the different keypoint detection methods impacted the quality of the final 3D models of dental structures. Three key metrics were used to evaluate the reconstructed models: SSIM, PSNR, and LPIPS. After generating the 3D reconstructions from the point clouds obtained through each keypoint detection method, images were rendered from the reconstructed models (Figure 4) and compared with the original input frames in Table 2.

Table 2. Comparison of 3D mesh reconstruction by SuGaR with SIFT, LoFTR, KeyNetAffNetHardNet, and DISK + LightGlue keypoint detectors (**bold** numbers are best).

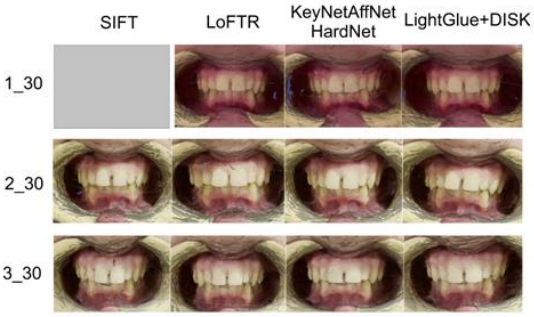
		SIFT			LoFTR (0.65 of initial size)			KeyNetAffNetHardNet			DISK + LightGlue		
Exp ID	Frames	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
1	5	0.9517	28.73	0.2590	0.9472	28.11	0.2557	**0.9541**	**28.92**	**0.2435**	0.9522	28.83	0.2473
	10	0.9440	27.09	0.2639	0.9381	26.58	0.2621	**0.9477**	**27.74**	**0.2452**	0.9437	27.10	0.2559
	15	0.8892	21.57	0.3270	0.9229	24.06	0.2730	**0.9356**	**25.54**	**0.2584**	0.9273	24.53	0.2698
	20	0.8381	18.81	0.3756	0.9075	22.95	0.2926	**0.9225**	**24.66**	**0.2779**	0.9162	23.61	0.2836
	25	–	–	–	0.9056	22.57	0.3006	**0.9232**	**23.91**	**0.2735**	0.9099	22.90	**0.2908**
	30	–	–	–	0.8998	22.54	0.2981	**0.9017**	**23.06**	0.3141	0.8946	21.52	0.2939
2	5	0.9416	27.39	0.2659	0.9430	27.40	0.2474	**0.9511**	**28.77**	**0.2317**	0.9493	28.49	0.2344
	10	0.9339	26.57	0.2724	0.9355	26.73	0.2541	**0.9447**	**27.73**	0.2376	0.9429	27.63	0.2373
	15	0.9277	25.69	0.2766	0.9305	25.69	0.2551	**0.9396**	**26.77**	**0.2367**	0.9364	26.44	0.2413
	20	0.9128	23.29	0.2884	0.9157	24.04	0.2671	0.9207	**24.14**	0.2583	**0.9218**	23.96	0.2544
	25	0.9064	22.71	0.3044	0.9030	23.09	0.2832	**0.9199**	**24.35**	**0.2585**	0.9175	24.16	0.2622
	30	0.8513	21.04	0.3291	0.8831	21.34	0.3011	0.9057	**23.32**	**0.2724**	**0.9061**	23.20	0.2728
3	5	0.9498	27.36	0.2487	0.9450	26.79	0.2383	**0.9526**	**27.55**	0.2265	0.9514	27.35	**0.2253**
	10	0.9405	26.26	0.2532	0.9372	25.60	0.2389	**0.9458**	**26.61**	0.2277	0.9450	26.29	**0.2229**
	15	0.9280	24.60	0.2612	0.9197	23.88	0.2619	**0.9306**	**24.69**	0.2437	0.9290	24.45	**0.2423**
	20	0.9261	24.25	0.2629	0.9180	23.51	0.2581	**0.9292**	**24.47**	**0.2390**	0.9289	23.59	0.2403
	25	0.9151	23.76	0.2740	0.9035	22.00	0.2813	**0.9156**	**23.24**	**0.2598**	0.9123	22.86	0.2635
	30	0.8772	20.88	0.3119	0.8787	20.77	0.3026	0.8872	**20.99**	0.2859	**0.8873**	20.44	**0.2843**



(a) Dental 3D reconstruction with every 5 frames extraction.



(b) Dental 3D reconstruction with every 15 frames extraction.



(c) Dental 3D reconstruction with every 30 frames extraction.

Figure 4. Final 3D mesh reconstruction by SuGaR with SIFT, LoFTR, KeyNetAffNetHardNet, and DISK + LightGlue keypoint detectors.

KeyNetAffNetHardNet consistently resulted in the highest quality reconstructions. Models reconstructed using this method achieved the highest SSIM and PSNR values across all datasets and frame extraction intervals. This indicates that the reconstructed images were structurally similar to the original images and maintained high fidelity, effectively capturing fine details and textures of the dental anatomy.

The LPIPS metric sometimes produced slightly better scores for the models reconstructed using DISK + LightGlue, but overall, KeyNetAffNetHardNet was better across all three metrics collectively. However, the marginal advantage of DISK + LightGlue in LPIPS did not compensate for its additional computational cost and less reliable performance in both SSIM and PSNR.

SIFT and LoFTR-based reconstructions did not perform as well. The SIFT-based method was shown to be less accurate in capturing structural details and textures, based on lower SSIM and PSNR values. Even though LoFTR had strong matching performance for neighboring frames, the final 3D models from LoFTR-based reconstructions were of reduced quality.

4.3. Timing Comparison of Keypoint Methods

Total processing times were measured across multiple experiments and frame extraction intervals. KeyNetAffNetHardNet typically required 900 s to 1,200 s (about 15–20 minutes) per experiment, showing only moderate variation between different frame spacings. DISK + LightGlue exhibited similar or slightly shorter run times, ranging from around 1,100 s to 1,700 s (18–27 minutes). In contrast, LoFTR demonstrated the widest variability, with durations spanning 1,100 s to 1,900 s (roughly 18–31 minutes), especially when processing denser frame extractions. Meanwhile, SIFT generally finished within 700 s to 1,200 s (12–20 minutes) but occasionally produced anomalously short measurements, possibly due to reduced subsets of frames in certain trials. Overall, these results indicate that LoFTR, although robust in low-texture regions, may incur notably higher computational costs, whereas KeyNetAffNetHardNet and DISK + LightGlue strike a more favorable balance of speed and consistent keypoint matching performance.

5. DISCUSSION

This study presented a novel pipeline for 3D reconstruction of dental structures, addressing the inherent challenges of dental imagery, including complex textures, repetitive patterns, and low-texture regions. By integrating advanced keypoint detection with KeyNetAffNetHardNet, the goal was to enhance the discriminability and robustness of keypoint matching in dental images. The incorporation of Gaussian Splatting for efficient scene representation and SuGaR for precise mesh reconstruction further facilitated efficient mesh reconstruction, leading to high-quality 3D models.

The experimental results demonstrated that the KeyNetAffnetHardNet significantly outperformed traditional methods like SIFT and ORB, as well as advanced techniques such as LoFTR and DISK + LightGlue, in terms of keypoint matching accuracy and computational efficiency. Specifically, KeyNetAffnetHardNet consistently achieved the highest number of inlier matches and maintained robust performance even under challenging conditions with extreme viewpoint differences. This robustness is crucial for dental imaging, where variations in perspective and occlusions are common due to the confined space of the oral cavity.

The experimental results demonstrate significant enhancements achieved by the proposed pipeline in 3D dental reconstruction. With dense frame extraction intervals (every 5 frames), the reconstructed models attained high SSIM values of approximately 0.95 and PSNR values up to 29, indicating excellent structural similarity and fidelity to the original images. The LPIPS metric was as low as 0.24, confirming high perceptual quality. Compared to traditional methods like SIFT and ORB, which yielded SSIM values around 0.9 and PSNR below 25, the proposed pipeline improved SSIM and PSNR by up to 10% and 15%, respectively. This improvement is attributed to the advanced keypoint detection and matching capabilities of KeyNetAffNetHardNet, effectively handling complex textures and repetitive patterns in dental imagery. Even with increased frame intervals (every 30 frames) leading to sparse datasets, the method maintained robust performance, with SSIM values between 0.89 and 0.91 and PSNR ranging from 21 to 23. The LPIPS metric remained low at approximately 0.28, indicating preserved perceptual quality despite reduced data density. When compared to state-of-the-art

methods like LoFTR and DISK + LightGlue — which achieved SSIM values around 0.88 and PSNR up to 22 on sparse datasets — the proposed pipeline consistently outperformed them across all metrics. This demonstrates superior reconstruction quality and robustness to data sparsity. The integration of SuGaR further enhanced mesh accuracy and rendering quality. The pipeline effectively captured fine dental structures, resulting in smoother surfaces and more detailed 3D models.

These results highlight the significant contributions of the new pipeline. By effectively addressing challenges such as complex textures, repetitive patterns, and low-texture regions, the method advances 3D dental reconstruction. It provides a robust and efficient solution for generating high-fidelity 3D models, which is crucial for precise dental treatment planning and diagnostics.

The experiments indicate that the superior performance of KeyNetAffNetHardNet is partly attributed to several inherent design characteristics. In particular, KeyNet is geared towards detecting keypoints in complex textures, such as those found in dental images, where the peppered texture of teeth and gums makes picking out corners of details difficult. The capacity to focus on the salient features helps to offer more reliable keypoint detection on repetitive patterns or low texture regions, where conventional methods such as SIFT and ORB are vulnerable. In addition to KeyNet, AffNet estimates the affine shape of each keypoint and is robust to viewpoint changes and imaging distortions. Such an affine adaptation is greatly desired in dental imaging, where capturing different angles of the teeth requires major perspective variation [32]. AffNet normalizes the key points to a canonical form such that there is consistent feature representation across images from different viewpoints [21]. Furthermore, HardNet gives rise to highly discriminative descriptors for the detected keypoints. These are optimized to be able to distinguish between similar features, where those familiar to us are likely to be given lower priority so as to discount mismatches in places that are very repetitive [33]. If this level of discriminability is not achieved, high inlier match counts and accurate keypoint correspondences across images may not be obtained.

In comparison, advanced methods like LoFTR and DISK + LightGlue, while powerful, may not be as effective in this specific context. LoFTR relies on dense matching without explicit keypoint detection, which can be computationally intensive and may struggle with the repetitive patterns and low-texture regions common in dental images [5, 16]. DISK + LightGlue, although incorporating attention mechanisms to improve matching, may be more prone to confusion in areas with low texture or repetitive features [7, 34]. Additionally, these methods often require more computational resources and may not handle extreme viewpoint changes as robustly as KeyNetAffNetHardNet. While the proposed pipeline showed considerable improvements, it is important to acknowledge that the quality of the reconstructed teeth surfaces is still not perfect. Despite achieving higher SSIM and PSNR values compared to other methods, some fine details and surface textures of the dental structures were not captured with complete accuracy. This limitation can be attributed to several factors, including the reflective nature of dental enamel, the presence of specular highlights, and the challenges in capturing subtle variations in tooth morphology. The reconstructed models serve as effective representations of the significant geometric features of dental structures, which can meet the needs of further clinical applications (for instance, orthodontic assessments and implant planning). Although the identified imperfections do not severely obstruct the practical utility of the models, further refinement in these areas is needed.

In future work, other keypoint matching and detection algorithms will be investigated to enable higher accuracy in reconstructing fine-grained details. The reconstructions could be improved by incorporating methods for tackling reflective surfaces and specular highlights. Moreover, by

refining the mesh reconstruction process by means of more advanced regularization methods and larger iteration counts, it may be possible to extract smaller details and smoother surfaces. The dataset is also to be expanded to encompass a broader selection of dental anatomies, and also different imaging conditions. Limitations to the generalizability of the results include use of videos from one single participant in current study. An additional evaluation would involve including datasets that involve diverse datasets regarding dental structures, lighting environments, and heterogeneous patient demographics.

6. CONCLUSIONS

This study presents a comprehensive evaluation of keypoint detection and matching techniques for dental 3D reconstruction, demonstrating that the integration of KeyNetAffNetHardNet with Gaussian Splatting and SuGaR yields superior results compared to traditional and advanced methods. The proposed pipeline effectively addresses the unique challenges of dental imagery, providing high-fidelity 3D models that are crucial for further clinical applications in dentistry. Although the quality of the reconstructed teeth surfaces is not yet perfect, the results are promising and indicate that the pipeline functions effectively in capturing the essential geometries of dental structures. The imperfections observed offer valuable insights into areas for future improvement, guiding subsequent research efforts toward enhancing the accuracy and utility of the models. The ability to generate accurate and efficient 3D reconstructions has significant implications for dental diagnostics, treatment planning, and patient outcomes. By improving upon existing methods and introducing a practical solution that balances performance and computational efficiency, this work contributes to advancing the application of computer vision techniques in dentistry.

Future research will aim to refine the pipeline further, focusing on enhancing the quality of the reconstructed models and expanding the dataset for greater generalizability. Exploring additional techniques to handle the reflective properties of dental surfaces and incorporating advanced algorithms for capturing fine details will be essential steps forward.

7. LIMITATIONS AND FUTURE WORK

While our pipeline demonstrates promising results, one limitation is the relatively small dataset, that is currently drawn from a single participant. As a result, there may be variations in patient anatomy, lighting conditions, and enamel reflections that are not fully captured. In future work, we aim to expand our dataset to include multiple subjects, which would enable broader validation and potentially improve the generalizability of our methods. Additionally, integrating real-time processing or hardware-accelerated techniques could further support clinical adoption.

ACKNOWLEDGEMENTS

We thank the reviewers for their comments and suggestions.

REFERENCES

[1] T. Lindeberg, Scale invariant feature transform (2012).
[2] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, Orb: An efficient alternative to sift or surf, in: 2011 International conference on computer vision, Ieee, 2011, pp. 2564–2571.
[3] D. Pojda, A. A. Tomaka, L. Luchowski, M. Tarnawski, Integration and application of multimodal measurement techniques: relevance of photogrammetry to orthodontics, Sensors 21 (23) (2021) 8026.

[4] F. I. Ali, Z. T. Al-dahan, Teeth model reconstruction based on multiple view image capture, in: IOP Conference Series: Materials Science and Engineering, Vol. 978, IOP Publishing, 2020, p. 012009.
[5] J. Sun, Z. Shen, Y. Wang, H. Bao, X. Zhou, Loftr: Detector-free local feature matching with transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8922–8931.
[6] M. Tyszkiewicz, P. Fua, E. Trulls, Disk: Learning local features with policy gradient, Advances in Neural Information Processing Systems 33 (2020) 14254–14265.
[7] P. Lindenberger, P.-E. Sarlin, M. Pollefeys, Lightglue: Local feature matching at light speed, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 17627–17638.
[8] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, G. Bradski, Kornia: an open source differentiable computer vision library for pytorch, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 3674–3683.
[9] F. Remondino, L. Morelli, E. Stathopoulou, M. Elhashash, R. Qin, Aerial triangulation with learning-based tie points, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 43 (2022) 77–84.
[10] A. Gu´edon, V. Lepetit, Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5354–5363.
[11] D. Claus, J. Radeke, M. Zint, A. B. Vogel, Y. Satravaha, F. Kilic, R. Hibst, B. G. Lapatki, Generation of 3d digital models of the dental arches using optical scanning techniques, in: Seminars in Orthodontics, Vol. 24, Elsevier, 2018, pp. 416–429.
[12] S. O'Toole, C. Osnes, D. Bartlett, A. Keeling, Investigation into the accuracy and measurement methods of sequential 3d dental scan alignment, Dental Materials 35 (3) (2019) 495–500.
[13] M. G´omez-Polo, A. B. Barmak, R. Ortega, V. Rutkunas, J. C. Kois, M. Revilla-Le´on, Accuracy, scanning time, and patient satisfaction of stereophotogrammetry systems for acquiring 3d dental implant positions: A systematic review, Journal of Prosthodontics 32 (S2) (2023) 208–224.
[14] L. Morelli, F. Ioli, R. Beber, F. Menna, F. Remondino, A. Vitti, Colmap-slam: A framework for visual odometry, The International Archives of 19 the Photogrammetry, Remote Sensing and Spatial Information Sciences 48 (2023) 317–324.
[15] Y. Wang, Z. Li, L. Wang, M. Wang, et al., A scale invariant feature transform based method., J. Inf. Hiding Multim. Signal Process. 4 (2) (2013) 73–89.
[16] Y. Wang, X. He, S. Peng, D. Tan, X. Zhou, Efficient loftr: Semi-dense local feature matching with sparse-like speed, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 21666–21675.
[17] A. Mishchuk, D. Mishkin, F. Radenovic, J. Matas, Working hard to know your neighbor's margins: Local descriptor learning loss, Advances in neural information processing systems 30 (2017).
[18] N. M. Jebreel, J. Domingo-Ferrer, D. S´anchez, A. Blanco-Justicia, Keynet: An asymmetric key-style framework for watermarking deep learning models, Applied Sciences 11 (3) (2021) 999.
[19] D. Mishkin, F. Radenovic, J. Matas, Repeatability is not enough: Learning affine regions via discriminability, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 284–300.
[20] X. Chen, C. Fu, M. Tie, C.-W. Sham, H. Ma, Affnet: An attention-based feature-fused network for surface defect segmentation, Applied Sciences 13 (11) (2023) 6428.
[21] Y. Jin, D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi, E. Trulls, Image matching across wide baselines: From paper to practice, International Journal of Computer Vision 129 (2) (2021) 517–547.
[22] Z. Zhao, C. Wu, X. Kong, Z. Lv, X. Du, Q. Li, Light-slam: A robust deep-learning visual slam system based on lightglue under challenging lighting conditions, arXiv preprint arXiv:2407.02382 (2024).
[23] D. Barath, J. Matas, J. Noskova, Magsac: marginalizing sample consensus, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 10197–10205.
[24] J. L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4104–4113.
[25] J. Iglhaut, C. Cabo, S. Puliti, L. Piermattei, J. O'Connor, J. Rosette, Structure from motion photogrammetry in forestry: A review, Current Forestry Reports 5 (2019) 155–168.

[26] H. Chen, F. Wei, C. Li, T. Huang, Y. Wang, G. H. Lee, Vcr-gaus: View consistent depth-normal regularizer for gaussian surface reconstruction, arXiv preprint arXiv:2406.05774 (2024).
[27] G. Taubin, T. Zhang, G. Golub, Optimal surface smoothing as filter design, in: Computer Vision—ECCV'96: 4th European Conference on Computer Vision Cambridge, UK, April 15–18, 1996 Proceedings, Volume I 4, Springer, 1996, pp. 283–292.
[28] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.
[29] J. Korhonen, J. You, Peak signal-to-noise ratio revisited: Is simple beautiful?, in: 2012 Fourth international workshop on quality of multimedia experience, IEEE, 2012, pp. 37–38.
[30] D. Brunet, E. R. Vrscay, Z. Wang, On the mathematical properties of the structural similarity index, IEEE Transactions on Image Processing 21 (4) (2011) 1488–1499.
[31] S. Ghazanfari, S. Garg, P. Krishnamurthy, F. Khorrami, A. Araujo, R-lpips: An adversarially robust perceptual similarity metric, arXiv preprint arXiv:2307.15157 (2023).
[32] A. Barroso-Laguna, E. Riba, D. Ponsa, K. Mikolajczyk, Key. net: Key-point detection by handcrafted and learned cnn filters, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 5836–5844.
[33] T. Li, J. Liu, W. Zhang, L. Duan, Hard-net: Hardness-aware discrimination network for 3d early activity prediction, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, Springer, 2020, pp. 420–436.
[34] N. Uzay, O. Aydemir, A. C. ¨Ozgen, M. Kayao˘glu, S. Calap, L. C¸ akıl, Image stitching evaluation by object detection and geometric transformation in retail, in: 2024 32nd Signal Processing and Communications Applications Conference (SIU), IEEE, 2024, pp. 1–4.

AUTHORS

Bohdan Vodianyk is a Ph.D. candidate in 3D Computer Vision at the University of Málaga. He holds a Master's degree (2023) in Micro and Nano Electronics from the Faculty of Electronics, Igor Sikorsky Kyiv Polytechnic Institute. His research focuses on multi-view 3D reconstruction using light field 3D cameras and advanced computer vision techniques.

Enrique Nava Baro is a Professor of Signal Theory and Communications at the University of Málaga since 1994. He holds a Ph.D. in Telecommunications Engineering from the Polytechnic University of Madrid. He has led numerous multidisciplinary research projects and collaborated with prestigious international institutions, including the University of Chicago and RWTH Aachen. His research interests include signal processing and communications.

Alfonso Ariza Quintana is a Professor of Electronic Technology at the University of Málaga. He holds a Ph.D. in Telecommunications Engineering from the University of Málaga. His research focuses on wired and wireless networks, and he has extensive experience in communication systems simulation. Before academia, he worked at ELIOP S.A. on real-time control applications.

Anton Popov is an Associate Professor at the Igor Sikorsky Kyiv Polytechnic Institute. He holds a Ph.D. in Biomedical Devices and Systems and has significant expertise in biomedical electronics and brain activity monitoring. He has served in various academic leadership roles and actively participates in IEEE conferences.