# An End-to-End System for 3D Reconstruction of Intraoral Cleft Anatomy in Children from Smartphone Video

Supplementary Material

## 7. Scanner Hardware

The 3D intraoral scans are acquired with Medit-i500 and utilize the corresponding medit software.

## 8. Data Sharing and Reproducibility

**Datasets** Given the ethical considerations in handling patient data, particularly the sensitivity of medical information from children and partial facial data, we cannot share our entire dataset. Both real and synthetic data are directly based on the intraoral region of patients, which can in theory be used to identify the patient. We aim to publish a small, anonymized subset of patient data that have explicitly consented to its use for scientific purposes, demonstrating our commitment to ethical data sharing.

**Code** We will publish our code together with instructions and parameters to preprocess and reconstruct a video.

## 9. Alternative Reconstruction Methods

Neural Radiance Fields [17] and Gaussian Splatting [12] are cutting-edge techniques in computer graphics and vision to synthesize novel views of complex 3D scenes from a set of input images. These methods excel in scenarios requiring photorealistic rendering, where traditional methods might struggle with complex scene structures and material properties. In addition, the surface of the object can be reconstructed. NeuS [29] builds on NeRFs to extract the 3D geometry of the scene. Trim3DGS [9] allows for enhanced geometry recovery of Gaussian Splatting. However, the original focus of these methods is mostly the appearance and not always geometry, especially in challenging settings. In addition, the filming of the intraoral region offers very limited viewing angles, while novel view synthesis methods usually rely on many different perspectives. Finally, these methods need camera positions as input. In a setting where the camera pose estimation is already a difficult task, they are bound to struggle to achieve high quality reconstructions.

An alternative can be depth based reconstruction methods that rely on projecting the depth map as points into 3D space from a single image. Hardware devices, such as LiDAR or depth cameras, can be used to capture an object. Truncated Signed Distance Function Fusion Integration [27] allows the fuse of multiple depth maps to create a surface object from the acquired depth maps. These scans are often used as reference and ground truth in public datasets [25]. For the purpose of intraoral reconstruction they have a few shortcomings. Mainly, most sensors do not work with high enough resolution and reliability in a very close range. In addition, the sensors used often struggle with reflective surfaces. Finally, the goal of smartphone based reconstruction is to lower cost and improve accessibility. Reliance on other hardware would compromise these goals. An alternative can therefore be the depth estimation based on RGB-images captured by a camera or smartphone. DepthAnything [31] is trained on a large dataset for monocular depth estimation. However, the cleft images are out of distribution, and therefore the predicted depth maps are not precise enough for clinical application.

## 10. Additional Pipeline Information

We achieved the best results with Mast3r-SfM. The images are previously masked, sub-sampled and cut to the region of interest. We reconstruct with 50 images. We used the default values given by the authors for the reconstruction. Next, we denoise the point cloud using CloudCompare. We first apply noise removal with n=6 nearest neighbors and a relative error multiplier of 1.5. Next, we apply statistical outlier removal with n=20 nearest neighbors and a sigma multiplier of 2 for the threshold. Finally, we mesh the result with the default parameters of NKSR and choose a level of detail of 0.95.

## 11. ArUco Dataset Creation

To obtain a pose estimate from intraoral video, we developed the following workflow. First, we laser cut a small MDF wood square with a thickness of 3mm. The length of the side varies between 2 cm and 3 cm to offer different sizes depending on the current size of the mouth. We sand the edges to ensure that there are no safety hazards. Next, we print out the ArUco code and leave a small white edge to allow for detection. We attach the code to the wooden plate. Then we use dental paste to press the wooden plate with the sticker on the back of the front teeth. This allows for a good viewing

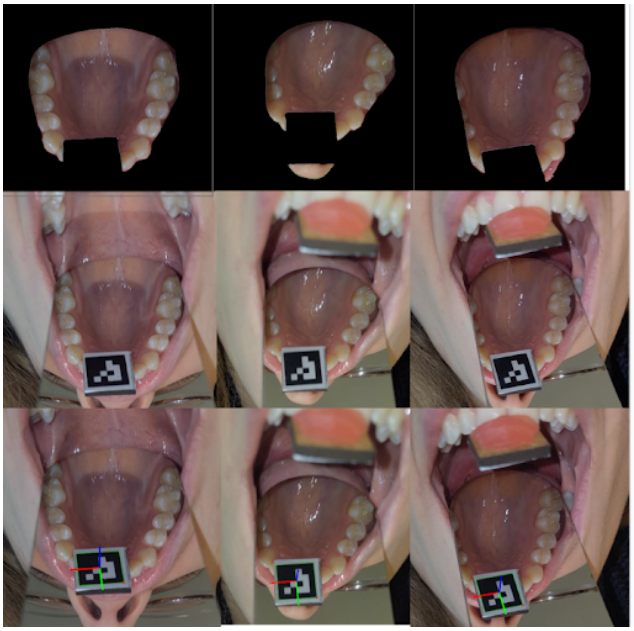

Figure 6. Example images from an ArUco video, masked and non masked. One can see the dental paste with the MDF plate and the printed ArUco code on top. The final row show the estimated camera poses.

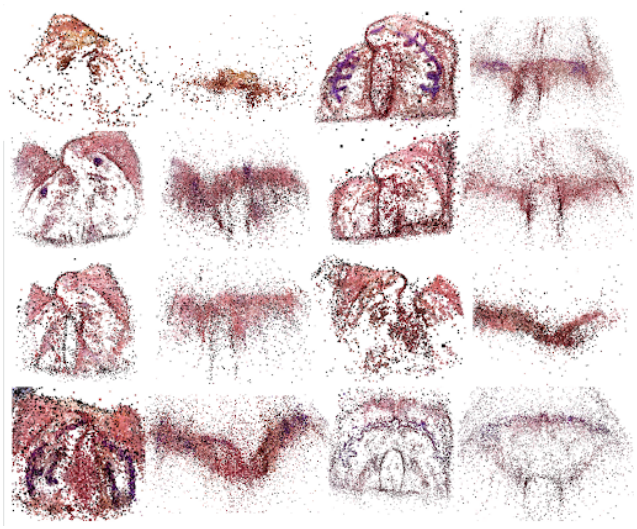

Figure 7. We show the eight largest point clouds, by number of points, of SIFT-based Structure from Motion. We show two angles for each: a top-down view and a side view. We note that even the most successful reconstructions have large noise variance and significantly lack point density in relevant regions. This would result in challenges for the meshing. We therefore decided to exclude these results from the quantitative evaluation, as they qualitatively failed.

angle when filming through the mirror. For each video, we fix the focal length of the camera and film a calibration checkerboard pattern within the same video. The pose is then estimated with OpenCV [4], utilizing the prior calibrated intrinsics.
We mask the images similar to the intraoral video captured in infants, but ensure that the ArUco code is completely masked to not offer any easy-to-match texture during reconstruction.

## 12. Additional Results and Material

In this section, we present additional qualitative results of the different reconstruction methods and medical evaluation.

### 12.1. COLMAP

We utilize the automatic reconstruction with COLMAP and complete a sparse reconstruction. We choose the high quality flag. Since our input images are masked, we use them both for the image input and as mask. We show the results in Figure 7.

### 12.2. Cleft-LoFTr

We use the method proposed by Lingens et al. [14] and evaluate the results. We thank the authors for sharing their methods. However, due to data privacy concerns, we will not be able to provide the complete reconstruction method. We will publish a LoFTR-based reconstruction using hloc [21], which does not rely on a shape model for denoising. We show the results in Figure 8.

### 12.3. VGGT

We refer to the author's github page for reference. We used the code provided. Due to GPU memory limitations, we could only reconstruct each case with six images. We experimented once with a setup of 15 images and observed similar layering issues as described in the main paper. We show the results in Figure 10.

### 12.4. Denoise Results

We show the quantitative and qualitative results for the denoising steps. The quantitative results are similar and do not show an improvement over the original point cloud. PointCleanNet outperforms the statistical methods. However, qualitative comparison shows that statistical methods are the better choice to generate final meshes. We show the results in Table 3 and Figure 13.

### 12.5. Medical Evaluation

We show some additional results and provide the detailed feedback received from the two medical experts. We show some plate examples in Figure 14, additional landmark information in Figure 17, and the feedback

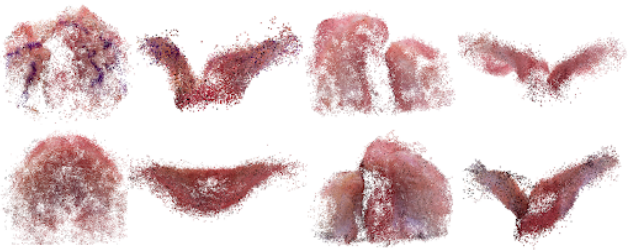

Figure 8. We show four representative successful Cleft-LoFTR reconstructions in the bottom two rows.

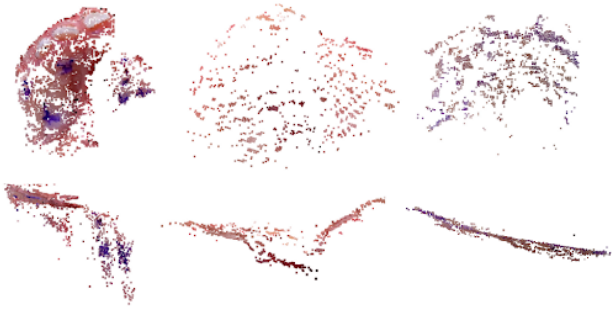

Figure 9. We show some failure cases of Cleft-LoFTR in the first two rows. Other failure cases had zero or close to zero points. These cases can not be visualized.

| Method | Mast3r | PCN | SOR/NR |
|---|---|---|---|
| F1@1.5 | 72.8 | 70.9 | 65.6 |
| Acc@1.5 | 76.5 | 80.9 | 83.6 |
| Com@1.5 | 73.8 | 67.3 | 58.5 |
| F1@0.5 | 46.8 | 47.5 | 45.5 |
| Acc@0.5 | 44.6 | 50.2 | 55.1 |
| Com@0.5 | 54.5 | 49.2 | 43.3 |

Table 3. The F1, accuracy and completeness score at 1.5mm and 0.5mm. Mast3r describes the original point cloud obtained with Mast3r-SfM. PCN describes the denoised point cloud with PointCleanNet and SOR/NR the statistical approach to denoising..

of the medical examiners in Table 4

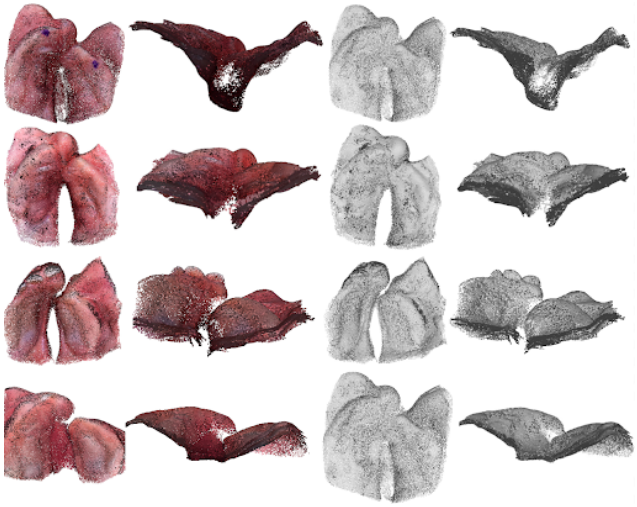

Figure 10. We present four example reconstructions of VGGT. We choose to visualize the results with normals to give a better understanding of depth. We further visualize the results without color, to enhance the geometry perception. Especially in the side view of the reconstructions, the layering problem can be observed.

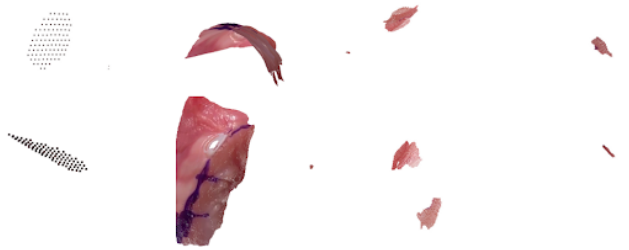

Figure 11. We show some failure cases of VGGT. Other failure cases had zero or close to zero points. These cases can not be visualized.

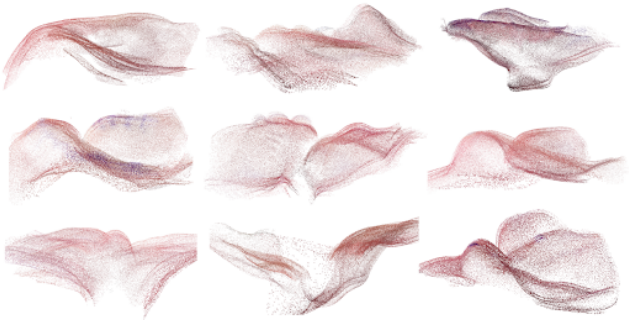

Figure 12. We reduce the rendered point size to highlight the layering challenges of VGGT. These layers are observed in every successful reconstruction.

| Plate Number | Plate Treatment | | | | Valuable for research measurements | Valuable for additional patient documentation |
|---|---|---|---|---|---|---|
| | without any adjustment | after minor subtractive adjustments | after major subtractive adjustments | not usable | | |
| 1 | | | O | X | OX | O |
| 2 | | | OX | | OX | O |
| 3 | | | OX | | OX | |
| 4 | | | O | X | OX | O |
| 5 | | | O | X | OX | O |
| 6 | O | | | X | OX | O |
| 7 | | O | X | | OX | O |
| 8 | O | | | X | OX | |
| 9 | | | | OX | | |
| 10 | | O | | X | O | |
| 11 | O | | | X | OX | O |
| 12 | O | | X | | O | OX |
| 13 | O | | | X | O | OX |
| 14 | O | X | | | O | OX |
| 15 | O | | | X | O | OX |

Table 4. The feedback received by the two healthcare practitioners. The O and X represent each a different healthcare professional.

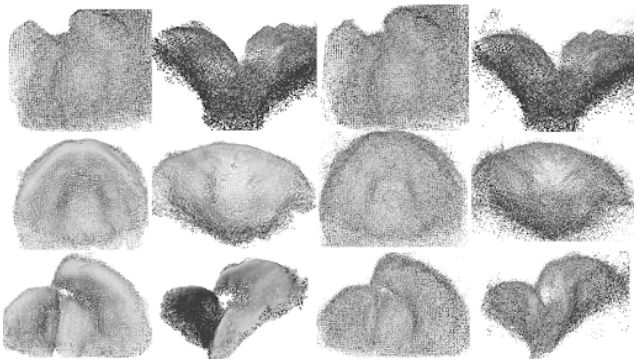

Figure 13. We visualize different denoise results. Each row represents another input. The first and second column are denoised with statistical methods, the third and fourth with PointCleanNet.

Figure 14. Five different plate reconstructions based on the digital pipeline by Schnabel et al. [22] and our pipeline reconstructions.

Figure 15. We show examples of a plate on a plaster-cast. The image is from Schnabel et al. [22]

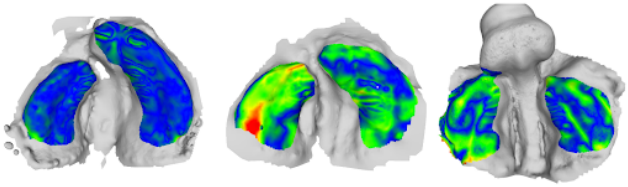

Figure 16. We highlight the plate contact region on three meshes. It is the region, which will have direct contact with the plate and is required to be the most accurate. The image is a modified version from Lingens et al. [14]

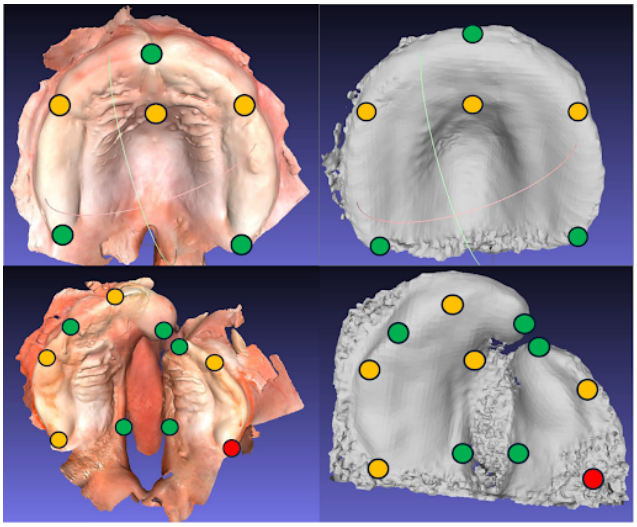

Figure 17. A doctor provided additional feedback on points relevant for common medical measurements. They marked the points on the intraoral scan and the reconstructions. Green points can be reliably placed, orange ones can sometimes be placed correctly and red points are never seen in any of the reconstructions. The categorization of these points happened on the set of 15 reconstructions send to the doctors.