Image-Based 3D Reconstruction of Cleft Lip And Palate Using a Learned Shape Prior

Lasse Lingens[1], Baran Gözcü[1], Till Schnabel[1], Yoriko Lill[2], Benito K. Benitez[2,3,4], Prasad Nalabothu[2,3,4], Andreas A. Mueller[2,3,4], Markus Gross[1], and Barbara Solenthaler[1]

[1] Department of Computer Science, ETH Zurich, Switzerland
[2] Oral and Craniomaxillofacial Surgery, University Hospital Basel and University of Basel, Switzerland
[3] Department of Clinical Research, University of Basel, Switzerland
[4] Department of Biomedical Engineering, University of Basel, Switzerland
lasse.lingens@inf.ethz.ch

Abstract. We present a novel pipeline that takes smartphone videos of the intraoral region of newborn cleft patients as input and produces a 3D mesh. The mesh can be used to facilitate the plate treatment of the cleft and support surgery planning. A retrained LoFTR-based method creates an initial sparse point cloud. Next, we utilize our collection of existing scans of previous patients to train an implicit shape model. The shape model allows for refined denoising of the initial sparse point cloud and; therefore, enhances the camera pose estimation. Finally, we complete the model with a dense reconstruction based on multi-view stereo. With Moving Least Squares and Poisson reconstruction we convert the point cloud into a mesh. This method is low-cost in hardware acquisition and supports minimal training time for a user to utilize it.

Keywords: Image-based 3D reconstruction · data-driven modeling · shape prior · cleft lip and palate

1 Introduction

Cleft lip and palate is the most common craniofacial birth defect with an estimated prevalence of 1 in 700 [13]. Presurgical orthopedic (PSO) treatment is commonly used to narrow the cleft and to enable a single-surgical repair [12]. The treatment involves the fabrication of a patient-specific plate that is inserted into the mouth and on the palate of a patient. This prevents the tongue from reaching inside the palate cleft and supports a natural narrowing of the cleft. The plate additionally eases food consumption and helps early speech development [2]. The creation of such an orthopedic plate consists of two steps. First, the practitioner acquires a 3D model of the specific intraoral region, either using an intraoral scanner or through silicon impression and subsequent fabrication of a plaster cast. Second, the digital or physical 3D model is used to design a person-specific well-fitting plate.

By using an entirely digital process, and hence 3D digital models of the cleft lip and palate, the automatic computation of the plate is enabled [18]. While plaster casts can be digitized and serve as input to the digital plate computation, the mesh quality is typically lower, and more importantly, the impression is taken under airway-endangering conditions [6]. Therefore, a fully digital alternative via intraoral scanners is the preferred capture technology today. However, clinics in low- and middle-income countries (LMICs) very often do not have access to such scanning devices, due to their high costs and requirement of trained personnel.

In this work, we aim to provide an alternative solution to intraoral scanners targeted at LMICs, such that the previously developed digital plate computation [18] can be applied. Our method turns a smartphone into an intraoral scanner, which outputs a digital 3D model of the cleft lip and palate just from a set of captured photographs. We leverage state-of-the-art deep learning based methods from Computer Vision for the first step of our 3D reconstruction [21], and combine it with a cleft shape prior trained on a collected data set of cleft lip and palate scans. We show that the domain-specific prior serves as a denoiser, leading to higher-quality meshes than domain-agnostic approaches. We further present the entire digital processing pipeline - from the raw input video to the final fabricated plate - and discuss the design choices of each step. Our results highlight the enormous potential of smartphone scanners for LMICs, and our work can be seen as a first step towards achieving this goal. Our contributions can be summarized as:

- Introduction of the detector-free local feature matching using transformer networks (LoFTR) to the medical community.
- A learned shape prior for cleft lip and palate, which was trained on a dataset of patient scans and is based on deep signed distance function.
- A complete digital processing pipeline: from an RGB smartphone video as input to the final printed orthopedic plate.

2 Related work

Neural approaches have led to drastic improvements of image-based 3D reconstruction quality across disciplines. In the following, we focus our discussion on photogrammetry and data-driven shape models.

Photogrammetry: Photogrammetry was dominated for a long time by detector-based local feature matchers. Two successful and prominent techniques are Scale-Invariant Feature Transform (SIFT) [11] and ORB [16]. These methods are hand-crafted and have been adopted in most computer vision-based tasks until recently. With the success of learning-based methods in many fields, photogrammetry progressed as well. NeRF-based methods such as NeuS [22] build a full implicit representation of the shape from the input images and camera pose estimations. Other recent notable methods include SuperPoint [8] as a feature extractor and SuperGlue [17] as a feature matcher that works in tandem with SuperPoint. The recently proposed detector-free method Detector-Free Local

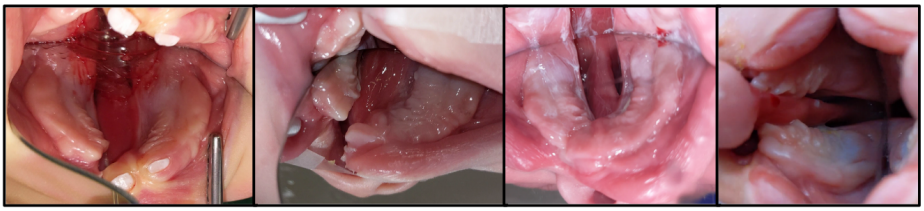

Fig. 1. Selected input images highlighting the challenges of the uncontrolled capture.

Feature Matching with Transformers (LoFTR) added the transformer network structure to correlate points spatially and build semi-dense correspondences between two images, offering more robust reconstructions for low feature surfaces. These properties are of great benefit for the reconstruction of the cleft region and is; therefore, featured as a central part of our proposed solution.

Data-driven shape models: Data-Driven Shape Models find their origin in the concept of PCA-based models. They have been explored in a variety of different fields, though the main area of research focuses on faces [3, 4]. The main focus of a morphable model is to learn the shape of an object class and compress that information into a compact latent. They are often used as a prior to fit observational data to and create a result within expectation of possible observations. In recent years, the statistical approach was replaced with learning-based methods. One prominent method is DeepSDF [14] and its variants [5, 9]. We leverage the representative power of DeepSDF and train it on cleft data to create a domain-specific shape prior, which is particularly useful in our setting where we have noisy and incomplete point data.

3 Methods

The goal of our work is to compute a digital 3D model of the cleft lip and palate based on an intraoral smartphone video, which is precise enough to compute and 3D print an orthopedic plate for the pre-surgical treatment. The smartphone video is captured in an uncontrolled environment, specifically, by doctors in a clinical setting at hospitals. This comes with multiple challenges for an image-based 3D reconstruction technique, including data that is captured through a mirror, with unsteady hand motion, movement of the infant during the capturing process, varying light conditions, occlusions due to the operators' hands, small capturing angles and limited mouth opening. Moreover, the intraoral surface has low quality features, no clear edges or corners, the surface is very reflective and the object of interest might undergo movement of even non-rigid nature. Since not all mobile phones are equipped with depth sensors, our reconstruction method is solely using RGB input data. Figure 1 shows example images that serve as input to our method.

Our reconstruction pipeline consists of multiple steps. We first pre-process the video to mask out the relevant region and sub-sample the frames based on a quality score and a given interval (Section 3.1). Next, we create a semi-dense reconstruction with LoFTR [21] (Section 3.2) and refine the reconstruction with DeepSDF [14] (Section 3.3). These two steps represent the core of our method. The refined semi-dense reconstruction is then completed to a dense reconstruction with Multi-View Stereo (MVS) [20]. Next, we fit our shape prior to our dense reconstruction and remove points with a distance greater than 0.5mm. We use Moving Least Squares (MLS) [1] (*radius = 2mm, order = 3*) to smooth out the resulting point cloud. This step is manually verified and the parameters adapted, if necessary. Finally, we use Poisson Reconstruction [10] (*tree depth = 8*) to create a mesh. This resulting mesh then serves as the input to the orthopedic plate computation [18]. In Figure 2 we show our pipeline to reconstruct the palatal area.

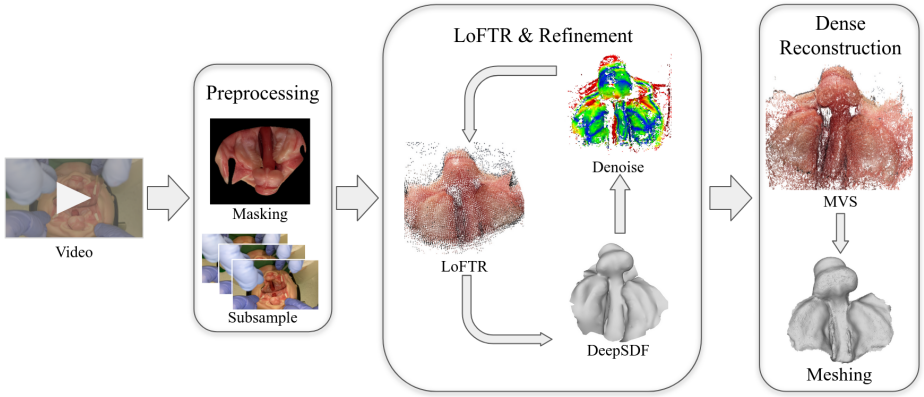

Fig. 2. Our pipeline uses a smartphone video as input that was captured in an uncontrolled clinical setting. After pre-processing, we compute a semi-dense reconstruction and use our DeepSDF shape prior as a denoiser, before computing the final mesh.

3.1 Data pre-processing

We semi-automatically mask all the frames of the video input to only include the palate region using MiVOS [7] as masking tool. This prevents ill-posed equation systems for the camera and point positions which are caused by the mirror in the image. We then automatically sub-sample the frames based on a heuristic approach to improve the average quality of the images and further reduce the processing time of the pipeline. We sub-sample at fixed intervals while considering within a range of each interval the quality of the image depending on their blurriness. To calculate the quality score, we apply a Laplacian kernel pixel-wise to each image.

3.2 Semi-dense reconstruction

For the semi-dense reconstruction, we use the state-of-the-art approach LoFTR [21], which outperforms classical feature extraction and matching methods. We verified the performance of LoFTR against a classic Structure from Motion and MVS approach with COLMAP [19, 20], SuperGlue+SuperPoint [17, 8] and NeuS [22]. LoFTR was the most robust approach over all cases, while it occasionally was outperformed for a single reconstruction. LoFTR takes a number of image pairs to be matched against each other as input. We use the two matching methods NETVLAD and sequential matching for pair finding. NETVLAD calculates the best n pairs based on global feature descriptions in all images. Sequential matching takes advantage that the input format is a video and matches each frame to the next m sub-sampled frames. LoFTR finds matches between the resulting image pairs and we use COLMAP to extract a semi-dense reconstruction out of the resulting meshes.

3.3 Data-driven shape prior

We use Deep Signed Distance Function (DeepSDF) [14], which has the advantage that it can be fit to point clouds, even if they are a noisy representation of a shape. The quality of the meshes produced by the model when fitted to a noisy point cloud is not sufficient for our targeted medical purpose. However, it provides a rough shape estimation for the current reconstruction. We utilize this shape estimation to denoise the point cloud that we get with the semi-dense reconstruction with LoFTR. In DeepSDF, two networks are trained in parallel. The first network is an encoder that receives sampled points of an observation and their signed distance to the mesh. The output is a latent code. The second network takes the latent code together with a single point and estimates the signed distance of that point to the mesh. We first align our point cloud with the canonical frame of reference of DeepSDF. We then sample the point cloud and create a shape estimation using our DeepSDF model. We apply the inverse transform of the initial alignment to the shape model estimation, such that it is re-projected into the reconstruction space. We calculate the distance from the point cloud to the closest point on the mesh and reject points over a threshold d. We track the removed points to their corresponding features and matches, and discard those as well. This in turn leads to refined camera poses, as the noisy points are no longer part of the equations system. Finally, we recompute the triangulation - now with reduced noise - to refine the result.

3.4 Data collection

We collected a data set of cleft lip and palate shapes, which consists of 188 intraoral scans and 553 plaster casts of the intraoral region of 489 cleft patients. The patients at scan time have an age of mostly 1-14 months. 178 of the patients are classified with a unilateral cleft and 86 with a bilateral, while the remaining are either not clearly classifiable or classified as a different cleft type. This data

set was used to train the DeepSDF model. For video acquisition we used a Google Pixel 4 and chose the 4k camera with 25 fps. We instructed the doctors to fulfill a steady slow ellipsoid movement with either a camera or a mirror to capture as many different viewing angles as possible. The duration of a video is usually between 15 and 30 seconds. Additionally, we tried to minimize occlusions, such as tubes, and non-rigid movement in the area of interest. As the object in focus are infants, sometimes awake, the videos have high variance in quality.

4 Results

In the following, we evaluate the quality of the 3D reconstructions (Section 4.1) and show the resulting pre-surgical plates (Section 4.2), demonstrating a proof-of-concept for the clinical use of a smartphone-based cleft and palate scanner. We further evaluate the effectiveness of the learned shape prior in Section 4.3.

4.1 3D reconstruction

We show the reconstruction quality for two unilateral cleft and two bilateral cleft cases in Figure 3, and compare the reconstructed shapes (second row) to the ground truth intraoral scans (top row). We display the color-coded error maps for the entire shape (third row), with blue and red corresponding to 0mm and 1.5mm, respectively. As expected, higher errors can be observed near the boundary, while smaller errors can be found in the relevant region near the ridges. The latter reflects the area that is relevant for the plate, as the final, fabricated plate needs to fit tightly to these ridges. We therefore evaluate the error for the particular area of interest as visualized in the last row. For the selected patient cases, we achieve a mean error of $[0.11, 0.39, 0.37, 0.28]$mm in the relevant area.

4.2 Plate evaluation

In order to evaluate if the accuracy of our smartphone based reconstruction is high enough for clinical settings, we used the digital plate computation algorithm of Schnabel et al. [18] and quantitatively assess the difference of the resulting digital plates when using an intraoral scan as input (second column) versus using our 3D reconstructed shape (third column) in Figure 4. For the two selected patient cases, we achieve a mean error of 0.11mm and 0.24mm, respectively. Note that the error is again only relevant along the ridges, and hence larger errors in the area that bridges the ridge areas are acceptable. Since it is difficult to conclude from these numbers if the resulting physical plate will fit well on a patient's palate, we 3D printed four selected plates using the previously reported clinical procedure [18], and collected feedback from three healthcare professionals who assessed the fitting quality. Out of four plates, they assessed two, three and three plates, respectively, to be applicable after no or only minor subtractive adjustments.

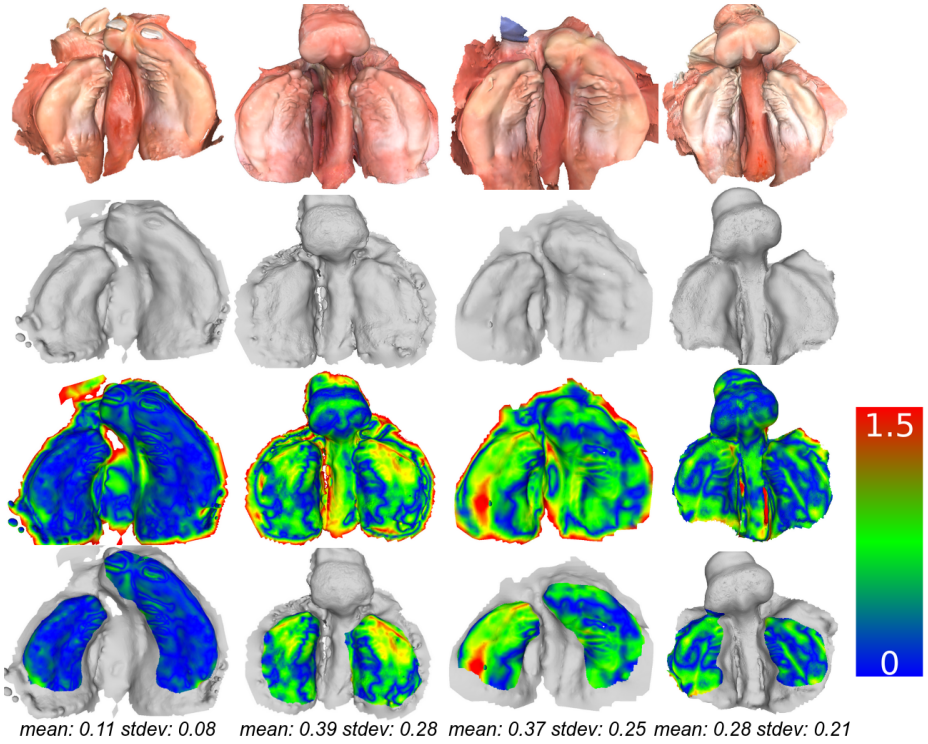

Fig. 3. From left to right we show four selected unilateral and bilateral patient cases. From top to bottom we show the ground truth meshes acquired with an intraoral scanner, our reconstructed meshes, and the error maps for the entire shape and partial area relevant for the pre-surgical plate.

4.3 Learned shape prior

The learned shape prior is a crucial part of our pipeline, and we therefore evaluate the expressiveness of our DeepSDF model for two selected cleft shapes in Figure 5. For the reconstruction of these two introral scans, the model achieves an average error of 0.14mm and 0.16mm in the area of interest, respectively. While the overall shape is approached quite accurately, it is also visible that very fine structural details are smoothed, which is a common problem of DeepSDF.

In our algorithm we use the DeepSDF shape prior as a denoiser. We have compared our method with the common denoisers Statistical Outlier Removal (SOR) and PointCleanNet (PCN) [15], and evaluated the methods based on correct identification of noise and of points that should be retained. Our data-driven shape prior noise removal outperforms the other methods (in percentages) for 1) correctly retained correct points (PCN: 45, SOR: 60, ours: 66), 2) incorrectly retained noisy points (PCN: 55, SOR: 40, ours: 34), 3) correctly removed noisy points (PCN: 81, SOR: 81, ours: 86), 4) incorrectly removed correct points

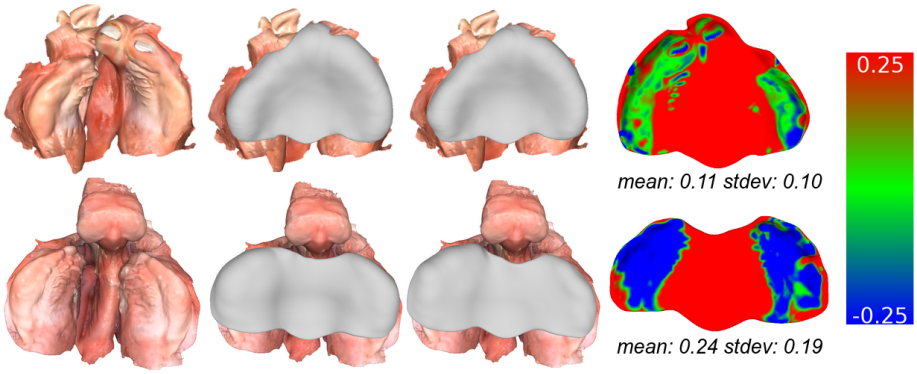

Fig. 4. For two selected cases we compare the resulting digital plates, once computed with an intraoral scan (second column) and our 3D reconstructed shape (third column). We visualize the plates on the original scan (left). The color-coded errors (right) are absolute distances of the region of interest around the ridges.

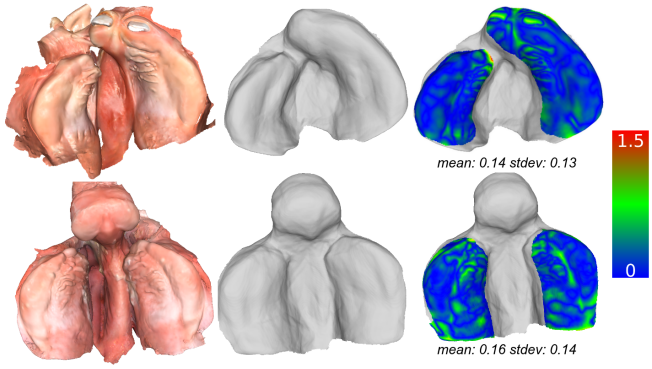

Fig. 5. Reconstruction of an intraoral scan (left) through our DeepSDF model (middle) and color-coded errors in the area of interest (right).

(PCN: 19, SOR: 19, ours: 14). Note that we hand-tuned the parameters for the alternative methods and applied multiple iterations to optimize their results.

5 Conclusion

We have presented a first smartphone-based scanning solution for the 3D reconstruction of the cleft and palatal region. All steps in our pipeline are data-driven and outperform conventional approaches when applied to input captured in uncontrolled clinical settings. We have demonstrated a proof-of-concept by computing and fabricating plates based on our reconstruction, which can then be used for the pre-surgical treatment of cleft lip and palate. The evaluation of the clinicians was overall positive, indicating great promise for using smartphone-based

scanners in clinical settings. However, a larger evaluation and clinical study is needed to draw a resulting conclusion.

Limitations Our current approach is heavily dependent on the quality of the input video. In some cases the video was too short, the infant moved too much or the camera was too shaky. This led to a failure of reconstruction. In addition, in some cases it proved to be difficult to capture the outside regions of the ridges as they were often occluded by the lips. Some error margin can be explained due to the global shape model and the necessary smoothing in the post-processing step, both of which reduce high frequency details in the reconstructions. Finally, there is a potential to further increase the automation level of our pipeline and eliminate the remaining manual steps such as initial mask segmentation or alignment of shape model and reconstruction.

Prospect of application Our image-based 3D reconstruction approach enables the use of the PSO in low- and middle- income countries, where intraoral scanners are often not available. Our method relies solely on RGB images, which reduces requirements related to hardware. It further supports remote check ups, as the equipment is affordable and available and the image capture process is innocuous.

References

1. Alexa, M., Behr, J., Cohen-Or, D., Fleishman, S., Levin, D., Silva, C.: Computing and rendering point set surfaces. IEEE Transactions on Visualization and Computer Graphics **9**(1), 3–15 (2003). https://doi.org/10.1109/TVCG.2003.1175093
2. Alzain, I., Batwa, W., Cash, A., Murshid, Z.A.: Presurgical cleft lip and palate orthopedics: an overview. Clin Cosmet Investig Dent **9**, 53–59 (May 2017)
3. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. p. 187–194. SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co., USA (1999). https://doi.org/10.1145/311535.311556, https://doi.org/10.1145/311535.311556
4. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5543–5552 (2016). https://doi.org/10.1109/CVPR.2016.598
5. Chabra, R., Lenssen, J.E., Ilg, E., Schmidt, T., Straub, J., Lovegrove, S., Newcombe, R.A.: Deep local shapes: Learning local SDF priors for detailed 3d reconstruction. CoRR **abs/2003.10983** (2020), https://arxiv.org/abs/2003.10983
6. Chate, R.A.C.: A report on the hazards encountered when taking neonatal cleft palate impressions (1983–1992). British Journal of Orthodontics **22**(4), 299–307 (1995). https://doi.org/10.1179/bjo.22.4.299, https://doi.org/10.1179/bjo.22.4.299, pMID: 8580095
7. Cheng, H.K., Tai, Y.W., Tang, C.K.: Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In: CVPR (2021)

8. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. CoRR **abs/1712.07629** (2017), http://arxiv.org/abs/1712.07629
9. Genova, K., Cole, F., Sud, A., Sarna, A., Funkhouser, T.A.: Deep structured implicit functions. CoRR **abs/1912.06126** (2019), http://arxiv.org/abs/1912.06126
10. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the Fourth Eurographics Symposium on Geometry Processing. p. 61–70. SGP '06, Eurographics Association, Goslar, DEU (2006)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (nov 2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94, https://doi.org/10.1023/B:VISI.0000029664.99615.94
12. Mishra, B., Singh, A.K., Zaidi, J., Singh, G.K., Agrawal, R., Kumar, V.: Presurgical nasoalveolar molding for correction of cleft lip nasal deformity: experience from northern india. Eplasty **10** (Jul 2010)
13. Mossey, P., Modell, B.: Epidemiology of Oral Clefts 2012: An International Perspective. In: Cleft Lip and Palate: Epidemiology, Aetiology and Treatment. S.Karger AG (06 2012). https://doi.org/10.1159/000337464, https://doi.org/10.1159/000337464
14. Park, J.J., Florence, P.R., Straub, J., Newcombe, R.A., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. CoRR **abs/1901.05103** (2019), http://arxiv.org/abs/1901.05103
15. Rakotosaona, M., Barbera, V.L., Guerrero, P., Mitra, N.J., Ovsjanikov, M.: POINTCLEANNET: learning to denoise and remove outliers from dense point clouds. CoRR **abs/1901.01060** (2019), http://arxiv.org/abs/1901.01060
16. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International Conference on Computer Vision. pp. 2564–2571 (2011). https://doi.org/10.1109/ICCV.2011.6126544
17. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020)
18. Schnabel, T.N., Gözcü, B., Gotardo, P., Lingens, L., Dorda, D., Vetterli, F., Emhemmed, A., Nalabothu, P., Lill, Y., Benitez, B.K., Mueller, A.A., Gross, M., Solenthaler, B.: Automated and data-driven plate computation for presurgical cleft lip and palate treatment. International Journal of Computer Assisted Radiology and Surgery (Apr 2023). https://doi.org/10.1007/s11548-023-02858-6, https://doi.org/10.1007/s11548-023-02858-6
19. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
20. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
21. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. CVPR (2021)
22. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. CoRR **abs/2106.10689** (2021), https://arxiv.org/abs/2106.10689