An Implicit Parametric Morphable Dental Model

CONGYI ZHANG, The University of Hong Kong, Hong Kong and Max Planck Institute for Informatics, Germany
MOHAMED ELGHARIB, Max Planck Institute for Informatics, Germany
GEREON FOX, Max Planck Institute for Informatics, Germany
MIN GU, The University of Hong Kong, Hong Kong
CHRISTIAN THEOBALT, Max Planck Institute for Informatics, Germany
WENPING WANG, Texas A&M University, USA and The University of Hong Kong, Hong Kong

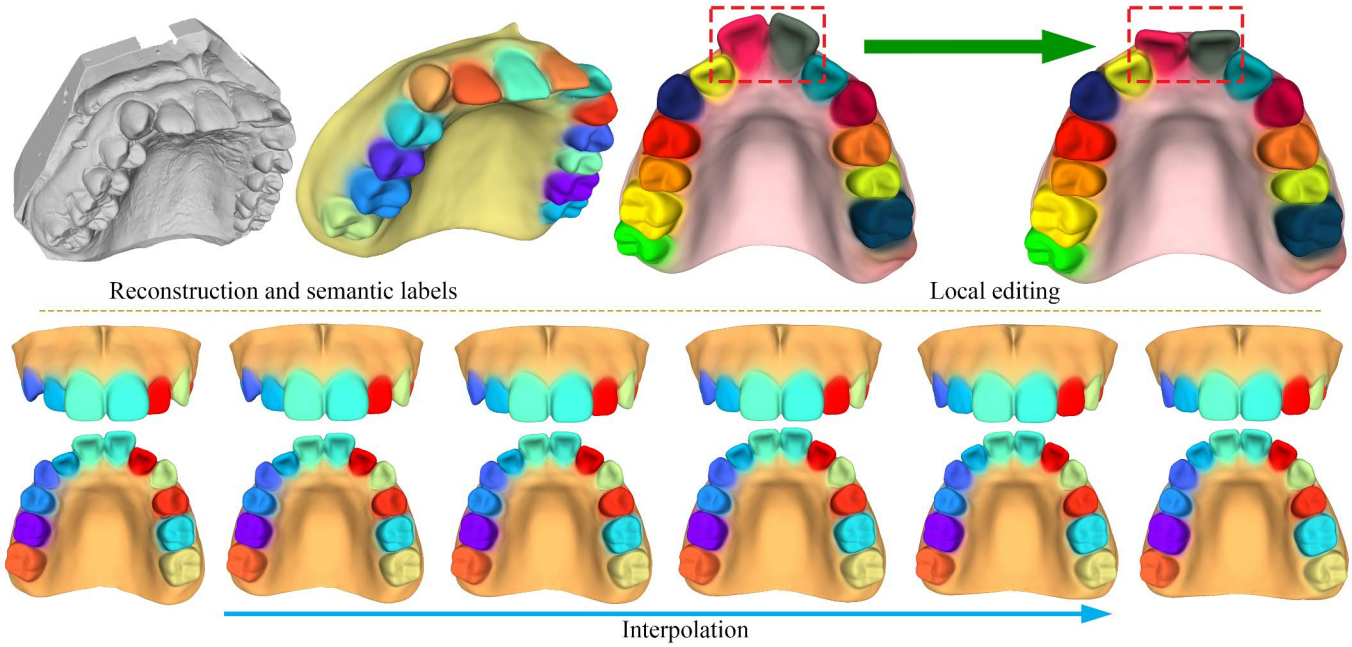


Fig. 1. We present an SDF-based morphable model for the human teeth and gums. Our model is *compositional*, i.e. the full geometry is a combination of a number of smaller components, that each model one semantically meaningful component: Each tooth and also the gums are controlled by separate latent codes. This allows our model to not just reconstruct geometry, but to also compute a semantic labelling in addition. Furthermore, it enables editing of specific components, e.g. individual teeth (see dashed red). Our model can also be used to smoothly interpolate between different teeth configurations, possibly serving as a visual aid in the communication between orthodontists and their patients.

3D Morphable models of the human body capture variations among subjects and are useful in reconstruction and editing applications. Current dental models use an explicit mesh scene representation and model only the teeth, ignoring the gum. In this work, we present the first parametric 3D morphable dental model for both teeth and gum. Our model uses an implicit scene representation and is learned from rigidly aligned scans. It is based on a component-wise representation for each tooth and the gum, together with a learnable latent code for each of such components. It also learns a template shape thus enabling several applications such as segmentation, interpolation and tooth replacement. Our reconstruction quality is on par with the most advanced global implicit representations while enabling novel applications. The code will be available at https://github.com/cong-yi/DMM

CCS Concepts: • **Computing methodologies** → *Shape modeling*; *Shape representations*.

Additional Key Words and Phrases: Implicit neural representation, Morphable model, Teeth

ACM Reference Format:
Congyi Zhang, Mohamed Elgharib, Gereon Fox, Min Gu, Christian Theobalt, and Wenping Wang. 2022. An Implicit Parametric Morphable Dental Model.

Authors' addresses: Congyi Zhang, The University of Hong Kong, Hong Kong and Max Planck Institute for Informatics, Saarbrücken, Germany, cyzhang@cs.hku.hk; Mohamed Elgharib, Max Planck Institute for Informatics, Saarbrücken, Germany, elgharib@mpi-inf.mpg.de; Gereon Fox, Max Planck Institute for Informatics, Saarbrücken, Germany, gfox@mpi-inf.mpg.de; Min Gu, The University of Hong Kong, Hong Kong, drgumin@hku.hk; Christian Theobalt, Max Planck Institute for Informatics, Saarbrücken, Germany, theobalt@mpi-inf.mpg.de; Wenping Wang, Texas A&M University, College Station, USA and The University of Hong Kong, Hong Kong, wenping@cs.hku.hk.

0730-0301/2022/12-ART217 $15.00
https://doi.org/10.1145/3550454.3555469

ACM Trans. Graph. 41, 6, Article 217 (December 2022), 13 pages. https://doi.org/10.1145/3550454.3555469

1 INTRODUCTION

The availability of morphable human face models has enabled various applications such as virtual face avatars for telecommunication [Lombardi et al. 2018; Wang et al. 2021], photorealistic animation in movies and media production [Flawless 2022; Kim et al. 2019; Synthesia 2022], single-photo editing [Tewari et al. 2020], and others [Hu et al. 2017; Thies et al. 2019; Zollhöfer et al. 2018b]. So far, however, most of such methods omit the modelling of the mouth interior, in particular teeth and gum. It is not only in the aforementioned applications that these parts of the human face are of importance, but also in medical research, for instance in orthodontic treatment. Capturing the geometry of the dental region (teeth and gum) is the basis for many interesting use cases, such as planning of the treatment or visualizing expected results for the patient. In this context, the availability of a morphable model with some control over original geometric components, e.g. over single tooth, could for example allow animating a transition from the status quo in a patient towards a desired treatment result.

Processing the human teeth poses multiple challenges: For one, human teeth can be shifted, rotated and generally misaligned in many ways, and some of them may even be missing altogether. Furthermore, the topological variety of teeth and their very uniform texture make it very hard to reliable detect, or even define any stable features on them. While there have been attempts to create morphable models for the human teeth only [Wu et al. 2016] (without gum), they are based on explicit representations such as meshes with a manually defined template shape. Learning such models requires accurate non-rigid registration of 3D scans and manual labeling of the teeth during reconstruction. Until now, there is no parametric morphable model for the human teeth that also includes the gums.

In this work, we present the first parametric morphable dental model for the geometry of the human teeth and gums. While the vast majority of human body models use explicit representations such as meshes [Egger et al. 2020; Loper et al. 2015; Romero et al. 2017], recently, there has been strong interest in using implicit representations [Alldieck et al. 2021; Corona et al. 2022; Palafox et al. 2021; Yenamandra et al. 2021; Zheng et al. 2022]. Motivated by this, we use, for the first time, an implicit representation for modeling teeth and gums: We adapt the implicit representation of DeepSDF [Park et al. 2019] where an object's geometry is represented by the decision boundary of a classifier that is supposed to tell whether points lie inside or outside of the examined object. To further identify dense correspondences, we learn a template shape and its deformations with the help of Hyper Networks [Sitzmann et al. 2020]. We use a compositional DeepSDF representation, i.e. each tooth or the gum is assigned a separate DeepSDF network with a learnable latent code. In addition to improving reconstruction accuracy over most of the original *global* implicit representations [Park et al. 2019; Zheng et al. 2021], the compositional representation allows various editing applications such as tooth replacement and interpolation (see Fig. 1). Our model can be learned from merely aligned 3D scans. The final overall model is a combination of the outputs of all the component models, weighted by segmentation indicators that are also predicted by the model.

In summary, we make the following contributions:

- We present the first implicit morphable model for the geometry of the human teeth and gum. We use a compositional implicit representation, with learnable latent codes for each region.
- Our model learns a template shape, which automatically establishes correspondences that help predict geometry segmentation during reconstruction.
- We introduce novel segmentation and teeth centroid losses that are crucial for training the model.
- Our technique produces accurate reconstructions that are on par with the state of the art, in addition to enabling novel applications such as tooth replacement, interpolation and segmentation.

As of yet, there is no publicly accessible model of the human teeth and gums, making ours the first such model when we release it.

2 RELATED WORK

In this section we start by discussing implicit representations used for modeling the scene geometry. Here, we examine both global [Chen and Zhang 2019; Mescheder et al. 2019; Park et al. 2019] and local [Deng et al. 2021; Zheng et al. 2021] representations, and the advantages the latter brings to literature. We then discuss recent methods for modelling the human body using implicit representations. Finally, we discuss methods for teeth processing with emphasize on teeth reconstruction and modeling.

2.1 Scene geometry modeling

Modeling scene geometry is a fundamental task in both computer vision and computer graphics. Previous methods use explicit representations such as meshes [Thies et al. 2016], voxels [Nießner et al. 2013] or points clouds [Keller et al. 2013]. While such explicit representations have been successful in many applications [Zollhöfer et al. 2018a], they suffer from a number of limitations: Voxels are memory-intensive, meshes struggle to handle detailed structures and point clouds are sparse and lose a significant portion of the geometry. Thus, in the past few years several efforts were made to explore so-called *implicit* geometrical representation [Chen and Zhang 2019; Mescheder et al. 2019; Park et al. 2019]. Unlike the earlier approaches, implicit representations encode the geometry indirectly, for instance as the decision boundary of a classifier that decides whether a point lies inside or outside the examined object. Among the most popular implicit representations are DeepSDF [Park et al. 2019] and Occupancy Networks [Mescheder et al. 2019]. DeepSDF uses a signed distance function to measure how far a point is from the surface of the examined object, while Occupancy Networks predict the probability of a point lying inside the object. Both methods demonstrate interesting capabilities such as interpolating between latent codes learned either by an auto-decoder [Park et al. 2019] or an auto-encoder [Mescheder et al. 2019] architecture.

Follow-up works addressed limitations of implicit representations. One main limitation is the lack of correspondences, which limit editing and model learning capabilities. To this end, some

works proposed to learn a template that is shared by all the training samples, which may be similar to the mean shape for the training samples. For instance, "Deep Implicit Templates"[Zheng et al. 2021], or DIT for short, uses a network that learns the deformation to the template shape. "Deformed Implicit Field" [Deng et al. 2021] , or DIF, proposes a similar idea, but, inspired by [Sitzmann et al. 2019], they use so-called Hyper-Nets, that predict the weights of their deformation networks.

There are several implicit representations that decompose the geometry into a number of localized components [Chabra et al. 2020; Chen et al. 2021; Genova et al. 2020; Peng et al. 2020; Tretschk et al. 2020; Wu et al. 2020] as opposed to the single global representation used in the earlier works [Chen and Zhang 2019; Mescheder et al. 2019; Park et al. 2019]. The motivation here is that while implicit representations are powerful, their global formulation could limit their generalization capabilities and reconstruction accuracy. Genova *et al.* [Genova et al. 2020] use localized deep implicit representations and assign latent codes to each local region. The method estimates a template shape as well as geometrical details through a local shape encoder. However, it automatically decomposes the watertight shape into a pre-defined number of components, the definition of which cannot be controlled and would lead to problems and mis-shaped geometry when teeth are missing. Chabra *et al.* [Chabra et al. 2020] uses a localized SDF representation with local latent codes defined in a voxel grid. Both methods show finer geometric details than global formulations and better generalization capabilities. Another potentially useful application of localized representations is the ability to perform novel applications. For instance Yin *et al.* [Yin et al. 2020] proposed a method that combines different parts of the same object together. This is done by learning the connections/joints of the various parts. Joints are learned using the implicit representation of Chen *et al.* [Chen and Zhang 2019]. They are learned in a way to agree with the remaining components while being smooth and topologically valid. The solution is trained with segmented components extracted from ShapeNet. PQ-Net [Wu et al. 2020] represents and generates 3D shapes in a sequential part assembly manner. However, since this method is not able to automatically segment the input data, it requires segmentation annotations at test time for the reconstruction task. Their geometric components are rigidly assembled to compose a model, while we believe that dental models require smooth blending of components into one reconstruction.

2.2 Implicit-based Modeling

Recently there has been increasing interest in building models of the various parts of the human body using implicit representations. This includes models for the human head [Yenamandra et al. 2021; Zheng et al. 2022], hands [Corona et al. 2022] and body [Alldieck et al. 2021; Deng et al. 2020; Palafox et al. 2021]. The work of Yenamandra *et al.* [Yenamandra et al. 2021] was the first in this regard. They presented the first 3D morphable model of the human head, including hair. The model learns latent codes for identity, albedo, expression and hairstyle. The model, named i3DMM, is based on a SDF-based architecture learned from 3D scans of various subjects with different hairstyles and performing different expressions. The method learns a template shape and a deformation to this shape. The template shape establishes correspondences and hence, unlike early 3DMM explicit face models [Egger et al. 2020], it does not need complicated non-rigid alignment of the scans, but merely rigidly aligned ones. The method shows novel interpolation applications in the latent spaces of all components e.g. identity, expressions and hairstyle. ImFace [Zheng et al. 2022] is a concurrent work to the one we present here. The aim is to improve the reconstruction accuracy of i3DMM [Yenamandra et al. 2021] using a localized SDF representation. To this regards, the entire face is decomposed into 5 regions with separate networks for expression and identity learning. A meta-learning approach is used, where hyper-nets learn the weights of the expression and identity networks. Results show more accurate reconstructions over i3DMM [Yenamandra et al. 2021]. An important difference between ImFace and our work is that since we aim at providing semantic control over each tooth individually, we assign one dedicated latent code to each geometric component, i.e. to each tooth and to the gums, whereas ImFace's latent codes cannot be partitioned into distinct regions of the geometry. Also, our method is different from NASA [Deng et al. 2020]: For a fixed number of joints, NASA encodes geometry as pose-conditioned occupancy. This is ill-suited for dental geometry, as single teeth might be missing, and annotating individual teeth poses and skinning weights for the dental scans in the training set would be very difficult.

2.3 Human Teeth Processing

There are several methods for processing the human teeth, including methods for teeth reconstruction [Abdelrehim et al. 2014; Farag et al. 2013; Wirtz et al. 2021; Wu et al. 2016; Zheng et al. 2011], restoration and completion [Mostafa et al. 2014; Ping et al. 2021], orthodontic treatment [Yang et al. 2020], segmentation [Cui et al. 2021; Zhang et al. 2021], pose estimation [Beeler and Bradley 2014; Murugesan et al. 2018; Yang et al. 2019] and others [Velinov et al. 2018; Wei et al. 2020]. The closest to our work are methods for teeth reconstruction and restoration [Abdelrehim et al. 2014; Mostafa et al. 2014; Ping et al. 2021; Wirtz et al. 2021; Wu et al. 2016]. The vast majority of these methods use explicit representations, with the exception of Ping *et al.* [Ping et al. 2021]. This work, however, does not propose a morphable model, but rather takes teeth crowns as input and completes them with gum. Abderlrehim *et al.* [Abdelrehim et al. 2014] reconstructs individual teeth from a single image by using shape-from-shading with shape priors. Here, shape priors are learned using PCA on height maps. Mostafa *et al.* [Mostafa et al. 2014] proposes a method for tooth restoration based on a single captured image. Restoration is achieved by aligning to a model derived from ensemble of oral cavity shapes and textures. Wirtz *et al.* [Wirtz et al. 2021] reconstructs teeth only, without gums, from 5 images of different viewpoints. They propose a model-based method that deforms a mean teeth shape to fit the input images. The optimum fitting is estimated by minimizing a silhouette loss between the 2D projections of the fitted 3D model and the observed 2D images.

Wu *et al.* [Wu et al. 2016] utilizes a teeth model learned from high quality dental scans and fit it to multiple images shot from

either a multiview camera setup or from a handheld moving camera. The teeth model follows a mesh representation and does not include the gum. To build the model, the 3D scans are aligned with respect to a manually defined template. Alignment is done through user-intervention, followed by a combination of rigid and non-rigid registration. An average shape model is estimated and PCA is used to estimate local shape variations. The final model accounts for the global deformations of the entire tooth row and the variations of each tooth individually.

We present the first implicit-based model for the human teeth and gum. It is morphable by design, producing intermediate shapes by interpolating between latent codes. Furthermore, it learns a reference shape that allows the computation of correspondences. This produces segmentation masks of the teeth as a by-product, as well as allowing interesting editing applications such as teeth replacement. In comparison to Wu *et al.* [Wu et al. 2016], our model does not need any labeling of the teeth for reconstruction, except for a binary vector indicating the presence/absence of individual teeth. Our model is trained on scans that were aligned only rigidly, in contrast to Wu *et al.*, who require sophisticated non-rigid alignment coupled with user-intervention. It also jointly models the gum with the teeth, unlike Wu *et al.*, who only model teeth. We believe our model is a useful contribution to recent efforts of building implicit models of the full human body as discussed in Sec. 2.2. We will release our model for research purposes, thus making it the only publicly available morphable teeth model including gums.

3 OVERVIEW

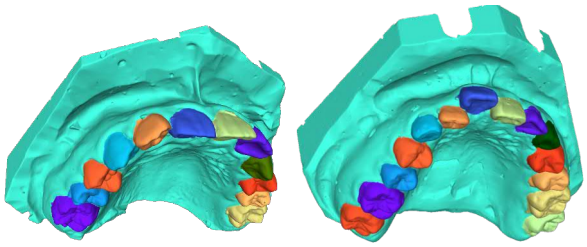


Fig. 2. Examples from our dataset of ground truth teeth geometries. Teeth identities have been annotated manually. We visualize them by different colours.

Our aim is to build a morphable model for the geometry of the human teeth and gum, with the ability to control each component individually. To this end, we present a compositional SDF representation where separate models are used to represent each tooth and the gum. For a standard dental scan (maxilla or mandible), we typically have one gum and up to 14 teeth (excluding 2 wisdom teeth), and thus we build our model assuming $m = 15$ components in total (Fig. 2). We learn 1 dedicated latent code for each of these 15 components. This allows editing applications such as tooth replacement and morphing.

Our proposed network consists of m sub-modules that represent the different components of the dental scan (Fig. 3a). Conditioned on a latent code and given a point in 3D space as input, each sub-module i predicts an SDF value $s_i + \Delta s_i$ and an indicator δ_i. The latter provides an estimate of the probability that the input point belongs to that part of the geometry that the sub-module is responsible for. Based on the δ_i values for all components, we compute a set of blending weights w_i to linearly combine the accompanying SDF values to one final value.

For each sub-module, we use a variant of DIF [Deng et al. 2021] (as shown in Fig. 3b): The input spatial point will first be warped by a Deform-Net. This network is generated by a Hyper-Net that is conditioned on a latent code. It maps each input point to a learned canonical reference space, in which a template shape is embedded by a Ref-Net (see Fig. 3b). Deform-Net also predicts the SDF compensation Δs_i to refine geometric details. Hence the component-wise SDF is computed as $s_i + \Delta s_i$, where s_i is the SDF value predicted by the Ref-Net.

Our method is trained on a dataset of dental geometries, manually annotated with semantic labels that segment the surface into the individual tooth types and the gums (see Fig. 2). Even though these models have been acquired by different methods (e.g. by digizing the traditional teeth impression of a patient, or more directly by an intra-oral scanning method) we will refer to them mostly as "dental scans", to avoid confusion between the various meanings of the word "model". Each scan is available as a high-resolution mesh, in particular allowing us to obtain normal vectors for supervision. More details are given in Sec. 5.1

4 METHOD

To enable control over each geometric component individually, we decompose the overall latent space of the entire model into subspaces that correspond to teeth and gums respectively. Inspired by DeepSDF [Park et al. 2019], we model the entirety of the teeth scan geometry as a function f that is conditioned on a set $\{\mathbf{z}_i\}_{i=1,\dots,m}$ of gum and teeth latent codes (where $m = 15$ is the number of geometric components in our implementation), and maps arbitrary spatial locations $\mathbf{p}$ to the signed distance s between $\mathbf{p}$ and the surface:

$$f(\mathbf{p}, \mathbf{z}_1, \dots, \mathbf{z}_m) = s \tag{1}$$

where the set $\{\mathbf{p} \mid f(\mathbf{p}, \mathbf{z}_1, \dots, \mathbf{z}_m) = 0\}$ constitutes the surface of the model. To learn sub-spaces of the geometry of individual components, we decompose the overall function f into a set $\{f_i\}_{i=1,\dots,m}$ of sub-functions:

$$f(\mathbf{p}, \mathbf{z}_1, \dots, \mathbf{z}_m) = \sum_{i=1}^{m} w_i \cdot f_i(\mathbf{p}, \mathbf{z}_i) = \sum_{i=1}^{m} w_i \cdot (s_i + \Delta s_i) \tag{2}$$

where w_i are blending weights s_i are signed distance values for component i and Δs_i are small-scale correction offsets for these values. In the following sections, we discuss how to compute the blending weights w_i and the function values $f_i(\mathbf{p}, \mathbf{z}_i) = s_i + \Delta s_i$, and we will define the objective function that we use for supervision.

4.1 Network structure

Each sub-function f_i is implemented as a pair $\Phi_i, \mathcal{R}_i$ of neural networks (see Fig. 3b): The *reference shape network* $\mathcal{R}_i(\mathbf{p}'_i) = s_i$ predicts SDF values s_i for the i-th reference shape, independent of the latent code $\mathbf{z}_i$. It is queried at those points $\mathbf{p}'_i$ that are produced by

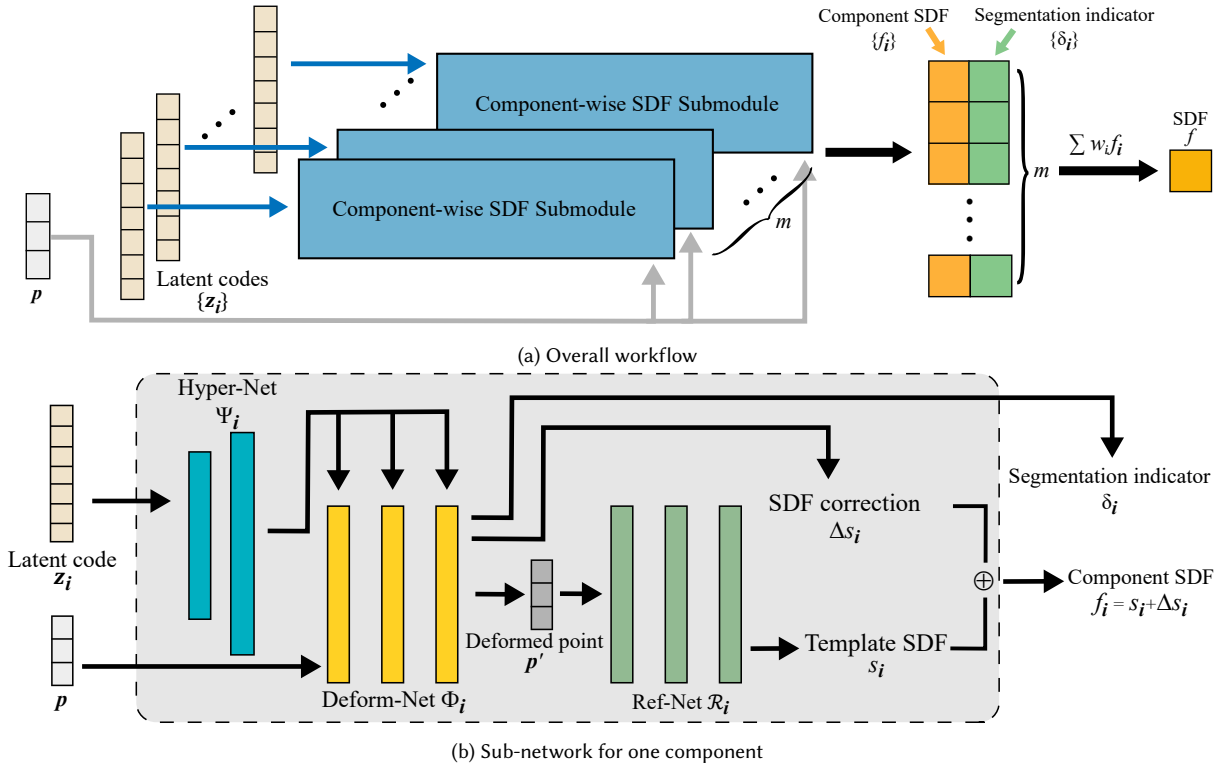


Fig. 3. The pipeline of our proposed method. We use a component-wise SDF representation where each tooth and the gum is represented by a separate "Component Shape Model". These models learn a reference shape for each component (in the Ref-Net), that is queried at those points to which the input points are warped by Deform-Net. Based on the component-wise SDF values and the segmentation indicators δ predicted for each component, we compute the full geometry as a weighted sum (see top right).

the *deformation network* $\Phi_i(\mathbf{p}, \mathbf{z}_i) = (\mathbf{p}'_i, \Delta s_i, \delta_i)$. The *deformation network* is supposed to warp the space in which the reference shape is embedded, such that querying $\mathcal{R}_i$ at the points $\mathbf{p}'_i$ results in the shape encoded by the $\mathbf{z}_i$.

In terms of output, there is a slight difference between the deformation networks for the gums and those for the teeth: For the gums, deformation is represented directly by the offset $\Delta\mathbf{p}$ of each query point $\mathbf{p}$ so that $\mathbf{p}' = \mathbf{p}+\Delta\mathbf{p}$. For the teeth however, the points close to them are always subject to almost the same rotation and translation, because teeth have characteristic rigid shapes. Therefore, inspired by [Park et al. 2021], we represent the transform for each point by a screw axis $(\mathbf{r}; \mathbf{t}) \in \mathbb{R}^6$ and set $\mathbf{p}' = e^{\mathbf{r}}\mathbf{p} + \mathbf{t}$ by Rodrigues' formula [Rodrigues 1816].

Note that the weights of Φ_i are themselves the output of a so-called Hyper-Net $\Psi_i(\mathbf{z}_i)$, a mechanism introduced in recent works [Deng et al. 2021; Sitzmann et al. 2020]. We differ from previous uses, however, in that Φ_i not only predicts the deformed points $\mathbf{p}'_i$ and SDF value correction deltas Δs_i, but also an additional value $\delta_i \in [0, 1]$ (see Fig. 3b), which serves as an indicator for teeth segmentation. This indicator is used to estimate the blending weights from Eq. (2):

$$w_i := \frac{\delta_i}{\sum_{j=1}^{m} \delta_j}$$

4.2 Objective Function

To train our component-wise implicit neural representations, we not only build on loss terms from previous work [Deng et al. 2021; Sitzmann et al. 2020], but also contribute two new loss terms that help decompose the geometry into semantically meaningful components, namely our *centroid loss* and our *segmentation loss*.

Centroid loss. To guide and regularize the deformation field, most face-related works use landmarks as their spatial constraints for deformation [Yenamandra et al. 2021; Zheng et al. 2022]. However, it is not easy to detect or even define landmarks on the teeth geometry. We solve this problem by enforcing that deforming the centroid point $\mathbf{c}_i$ of tooth i leads to a result $\mathbf{c}'_i$, that should coincide with the average centroid position $\bar{\mathbf{c}}_i$ of the training data for that tooth, which we can precompute before training. We penalize the ℓ^1 distance between the predicted and the expected centroid:

$$\mathcal{L}_i^{\text{centroid}} = \|\mathbf{c}'_i - \bar{\mathbf{c}}_i\|_1 \tag{3}$$

Segmentation loss. Our deformation network Φ_i produces values δ_i that indicate a confidence with which a particular point belongs to geometric component i (where components can be any of the teeth, or the gum). These values are very important to ensure the locality of each deformation network: Only if component i is sufficiently close to a given point should δ_i take a significant value and thus

be able to contribute to the overall SDF value for that point. Any SDF contributions coming from components that are not closest to the point should be suppressed by multiplication with the blending weights. Since the ground truth for segmentation is annotated on the surface, we will only actively supervise those sample points that reside on the surface with their ground truth labels. Since points that lie on the surface of one component are usually *off* the surfaces of the other components, this strategy, in combination with the SDF loss on the global level (see below) is sufficient for making the δ_i values behave in the required way (see Tab. 3 for an ablative evaluation). We force each Deform-Net to solve a binary classification problem that can be formulated by point-wise binary cross-entropy (BCE) loss:

$$\mathcal{L}_i^{\text{seg}} = \sum_{\mathbf{p}\in\mathcal{S}_i} \text{BCE}(\delta_i(\mathbf{p}), \ell_i(\mathbf{p}) == i) \tag{4}$$

where $\mathcal{S}_i$ is the tooth or gum surface and ℓ_i is the ground truth label for point $\mathbf{p}$. The Deform-Net basically learns to classify surface locations as either belonging to component i, or *not* belonging to component i. Note that, in contrast to Mu *et al.* [Mu et al. 2021], who use one shared network to represent all shapes and multi-class segmentation labels, each of our submodules learns the deformation field and label of only one corresponding geometric component.

Deformation smoothness loss. The deformation fields for all components should be rather smooth, because we expect the template shape learned by $\mathcal{R}_i$ to be close to the mean shape of component i. We thus constrain the deformation field Φ_i to be smooth, as in [Deng et al. 2021]:

$$\mathcal{L}_i^{\text{smooth}} = \sum_{\mathbf{p}\in\Omega} \left\| \nabla \left(\mathbf{p}_i'(\mathbf{p}) - \mathbf{p} \right) \right\|_2 \tag{5}$$

where $\nabla(\mathbf{p}_i'(\mathbf{p}) - \mathbf{p})$ is the Jacobian of deformation offset with respect to the coordinates of $\mathbf{p}$ and Ω is the 3D spatial domain.

SDF loss. We do not use ground-truth SDF values for supervision Instead we use a loss term from previous work [Sitzmann et al. 2020] that merely ensures that SDF values predicted by our model behave in a way that is consistent with the surface normals of the geometry and that satisfies the conditions that are generally expected from a signed distance field. We use this loss term twice, namely on the component level and also on the global level:

Component-level SDF loss. For each component we minimize:

$$\begin{aligned}\mathcal{L}_i^{\text{SDF}} = &\sum_{\mathbf{p}\in\mathcal{S}_i} |f_i(\mathbf{p})| + (1 - \langle \nabla f_i(\mathbf{p}), \bar{n} \rangle) \\ &+ \sum_{p\in\Omega} \left| \|\nabla f_i(\mathbf{p})\|_2 - 1 \right| + \sum_{\mathbf{p}\in\Omega\setminus\mathcal{S}_i} \psi\left(f_i(\mathbf{p})\right)\end{aligned} \tag{6}$$

where $\psi(x) = \exp(-\alpha \cdot |x|)$ with $\alpha \gg 1$. The first sum in this term encourages points on the surface $\mathcal{S}_i$ to be mapped to SDF values close to 0, while the spatial gradient of the SDF value at those points should be directed parallel to the ground truth normal vectors $\bar{n}$. The last sum in the loss term forces *non*-surface points to be mapped to SDF values that are *different* from zero. The sum term in the middle forces the gradient of the signed distance field to be 1 almost everywhere.

Global-level SDF+Normal loss. We use the same loss term also on the global level, to make sure that the result of blending all the geometric components together has the same desirable properties as each component individually:

$$\begin{aligned}\mathcal{L}^{\text{SDF}} = &\sum_{\mathbf{p}\in\mathcal{S}} |f(\mathbf{p})| + (1 - \langle \nabla f(\mathbf{p}), \bar{n} \rangle) \\ &+ \sum_{p\in\Omega} \left| \|\nabla f(\mathbf{p})\|_2 - 1 \right| + \sum_{\mathbf{p}\in\Omega\setminus\mathcal{S}} \psi\left(f(\mathbf{p})\right)\end{aligned} \tag{7}$$

Normal consistency loss. The SDF loss (Eqs. (6) and (7)) makes sure that our signed distance fields are consistent with the ground truth normal vectors for each instance. However, it is possible to satisfy the SDF loss while deforming points from different instances to different parts of the template shape. We need to discourage this, in order to establish consistent correspondences between the surface points of a particular instance and the surface points of the template shape learned by $\mathcal{R}$. As pointed out before [Deng et al. 2021], this can be achieved by forcing the normal directions at the points of the template surface to coincide with the normals at the corresponding points in each instance:

$$\mathcal{L}_i^{\text{corres}} = \sum_{\mathbf{p}\in\mathcal{S}_i} \left(1 - \left\langle \nabla\mathcal{R}_i\left(\mathbf{p}'\right), \bar{n} \right\rangle\right) \tag{8}$$

where $\nabla\mathcal{R}_i$ is the spatial gradient of the reference net, and $\bar{n}$ is the ground truth normal of the query point $\mathbf{p}$ on the original surface $\mathcal{S}_i$.

SDF correction regularization loss. To make sure that SDF correction values Δs_i remain reasonably small and serve only to encode fine geometric details (as opposed to encoding the entire shape in general), we add the regularizer [Deng et al. 2021]

$$\mathcal{L}_i^{\text{correction}} = \sum_{\mathbf{p}\in\Omega} |\Delta s_i(\mathbf{p})| \tag{9}$$

Latent code regularization loss. The latent codes are regularized by minimizing the term [Park et al. 2019]:

$$\mathcal{L}_i^{\text{latent}} = \|\mathbf{z}_i\|_2^2 \tag{10}$$

In summary, we optimize the following objective:

$$\begin{aligned}\mathcal{L} = \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} (&\lambda_1\mathcal{L}_i^{\text{centroid}} + \lambda_2\mathcal{L}_i^{\text{seg}} + \lambda_3\mathcal{L}_i^{\text{corres}} + \lambda_4\mathcal{L}_i^{\text{smooth}} \\ &+ \lambda_5\mathcal{L}_i^{\text{latent}} + \lambda_6\mathcal{L}_i^{\text{correction}} + \lambda_7\mathcal{L}_i^{\text{SDF}}) + \lambda_8\mathcal{L}^{\text{SDF}}\end{aligned} \tag{11}$$

Here, $\mathcal{I}$ is the set of all component identities (i.e. all teeth and gum components) and the $\{\lambda_k\}$ are hyperparameters that control the weight of each term. Note that because dental patients may be missing one or more teeth, we dynamically set the component-level weights (i.e. λ_1 to λ_7) to zero in such cases, i.e. we do not supervise the submodules for components that are not present in a particular training scan.

4.3 Data Preprocessing and Sampling

In order to provide the training data for our morphable model, we first collect a set of dental scans (see Fig. 2). Each of the scanned models was manually annotated, to segment them into gums and individual teeth, with each tooth being labelled by its identification

number in the FDI World Dental Federation notation system (ISO-3950) [ISO-3950 2016].

Since the scans have been acquired using a variety of different devices, they are not aligned in a common coordinate system. We thus select one of the scans and normalize it to occupy the volume $[-1, 1]^3$. Then we align all scans with this template, by aligning the centroids of corresponding teeth in a generalized Procrustes analysis (GPA) [Gower 1975] that estimates the optimal scaling, rotation and translation for each instance, minimizing the sum of square distances. Note that the centroids of teeth are approximately on a plane which makes the GPA unstable. We handle this problem by checking the determinant of the rotation matrix and making sure that reflections are converted to rotations when necessary [Arun et al. 1987]. Based on the aligned scans, we compute the average centroid for each tooth position.

In order to supervise our model, we need to sample the training data, which, for computational efficiency reasons, we do once before the start of training. We follow the sampling strategy proposed in DIF-Net [Deng et al. 2021] with slight modifications to ensure that the distribution of the samples is adequate for capturing geometric details. Specifically, we require most sample points to be on the surface of the geometry and we make sure that each tooth is sampled equally often, even though teeth differ in surface area. For each dental scan, we sample 25,000 points on each tooth and 100,000 on the gum. All the surface sample points come with their labels and surface normals. We also sample 500,000 free space points uniformly from $[-1, 1]^3$.

5 RESULTS

In this section, we evaluate our approach with regards to its reconstruction quality, its suitability for editing applications and the importance of its design choices. We also compare it to a number of methods for implicit scene representations [Deng et al. 2021; Park et al. 2019; Zheng et al. 2021]. While each of these methods shares some of the features of our method, none of them decomposes the geometry into semantically delineated components in the way we do, and thus cannot provide separate control over the semantic components of the geometry (e.g. over individual teeth). We refer the reader to the supplemental material for video results.

5.1 Implementation details

As described in Sec. 3, our model is trained on a dataset of dental scans. This dataset contains 1077 maxilla geometries, about half of which are malaligned. We split them randomly into 1027 for training, and 50 for testing. By flipping left and right and interchanging labels, we augmented the training set to 2054 geometries. Fig. 2 shows some examples from the dataset.

For each geometric component (i.e. for each tooth type and for the gum), we use a latent code of length 10. We weigh our loss terms by $\lambda_1 = 1$, $\lambda_2 = 10^2$, $\lambda_3 = 10^2$, $\lambda_4 = 50$, $\lambda_5 = 10^6$, $\lambda_6 = 10^3$, $\lambda_7 = 1$, $\lambda_8 = 0.1$. Our model is trained on 2 NVIDIA Quadro RTX 8000 GPUs for 120 epochs, which takes about 36 hours. In each iteration we sample 16384 points from 8 randomly selected ground truth scans. For all modules, the learning rate is $1 \cdot 10^{-4}$, which we halve every 30 epochs.

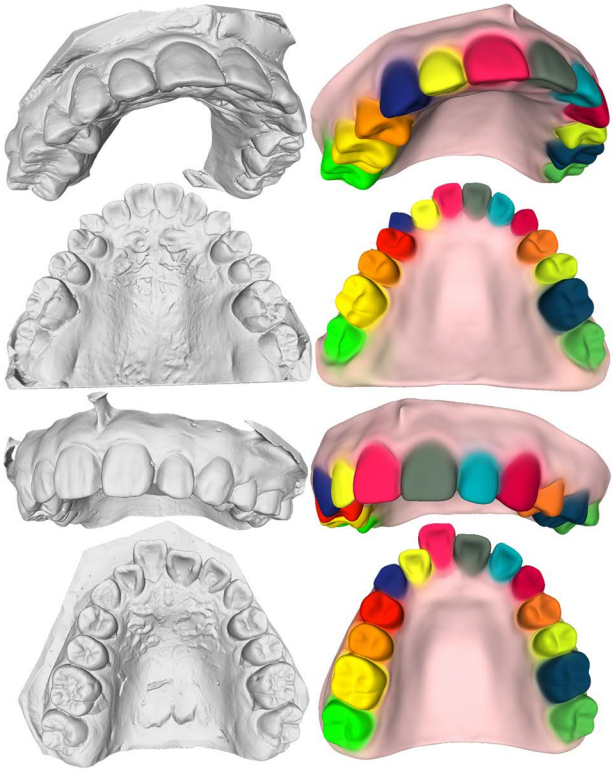


Fig. 4. Reconstruction with teeth labelling. Left column: raw dental scan data; Right column: reconstruction results with teeth labelled by our method.

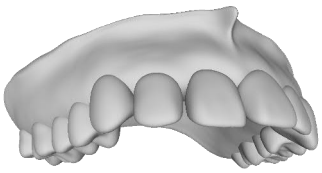


Fig. 5. The template shapes learned by our submodules, combined into one overall geometry.

5.2 Reconstruction & semantic labelling

To reconstruct a given dental scan, we keep the weights of our trained model fixed and solve, similar to previous work [Park et al. 2019], the following optimization problem:

$$\arg\min_{\{\mathbf{z}_i\}} \left(\mathcal{L}^{\text{SDF}} + \lambda^{\text{latent}} \sum_{i \in \mathcal{I}} \mathcal{L}_i^{\text{latent}} \right) \tag{12}$$

where $\mathcal{I}$ is the set of teeth indices that are present in the scan. Note that, for a user of our method, merely providing a boolean vector that indicates the presence or absence of individual teeth is significantly easier than manually segmenting the individual teeth in the raw scan data.

In the first and second column of Fig. 6 we compare some ground truth dental scans to our reconstruction results. We observe that in

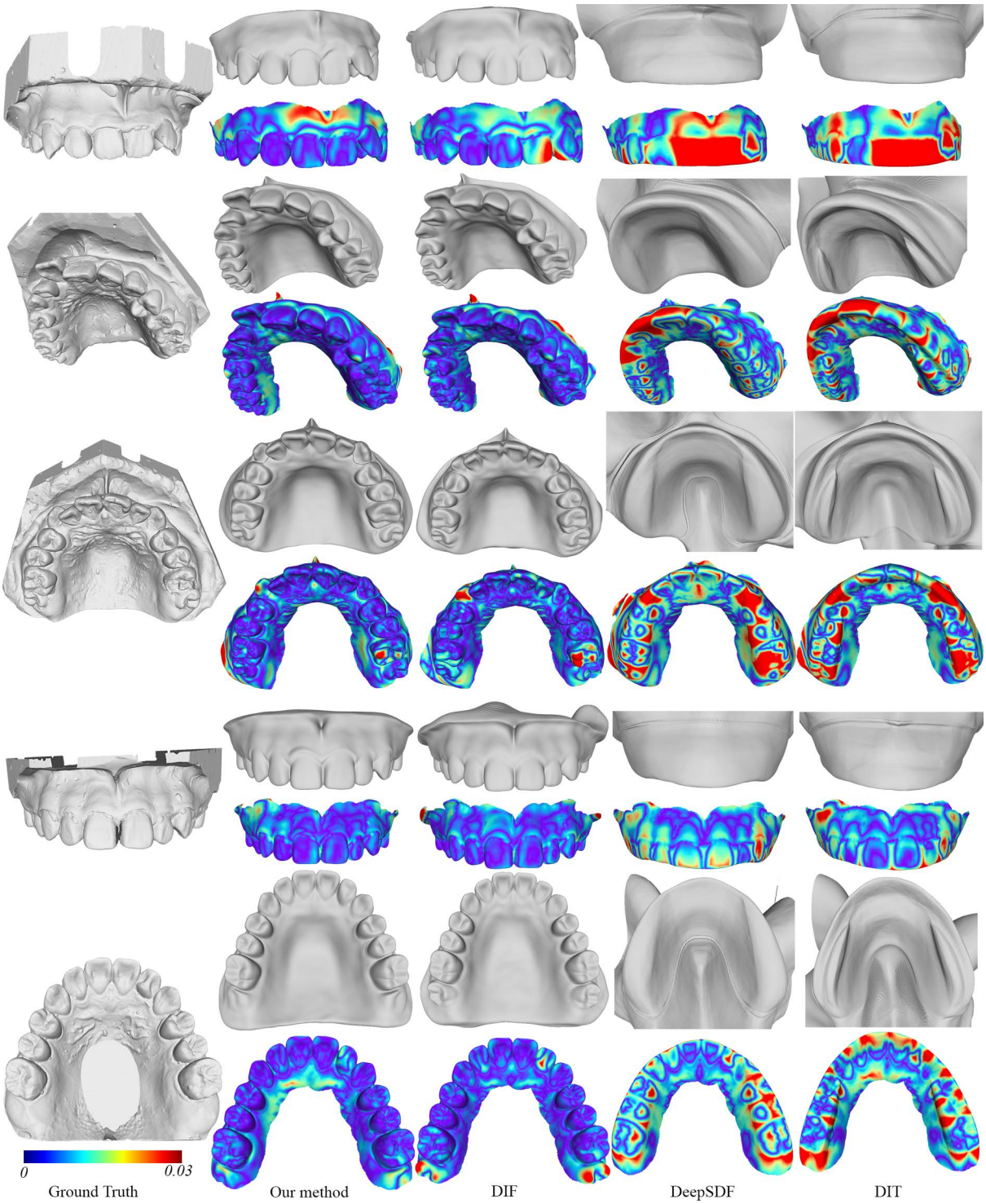


Fig. 6. Comparison of reconstruction results with DIF [Deng et al. 2021], DeepSDF [Park et al. 2019], and DIT [Zheng et al. 2021]. Our method clearly outperforms DIT and DeepSDF. Furthermore, the error heat map shows that our method reconstructs the teeth region more accurately than DIF. Note that our method is the only one that offers independent control over each tooth and and the gums, thus enabling interesting editing applications (see Sec. 5.3).

general our reconstructions quite faithful, even in cases where teeth are severely misaligned, such as in the second row. In particular, our reconstructions exhibit clearly visible gum lines.

Fig. 5 shows that our method also learns a meaningful template shape. It is via this template shape, that our method is not only able to *reconstruct* a given dental scan, but also to *label* it semantically, identifying each tooth in it (Fig. 4). The construction of our dataset still required this labelling to be done manually, but our method now provides a means of automating this task.

5.3 Editing applications

The major strength of our method is the fact that it decomposes teeth geometry into a number of semantically meaningful components that can be controlled individually. This allows us, for example, to edit a particular reconstruction result by replacing teeth that are mis-shaped or posed unaesthetically, by some more desirable counterparts (for example from a catalogue of aesthetically more pleasing teeth). Fig. 8 give an impression of this kind of editing application. Note that we only edited the incisor teeth, while all the other teeth remain unchanged. Fig. 1 gives another example (see dashed red).

Especially in an orthodontic context, for example as part of a telehealth application, such editing could be used to visualize different treatment outcomes to a patient. Since orthodontic treatments such as correcting misaligned teeth can be a lengthy and continuous process, there may also be some merit in visualizing them as one continuous animation, which we illustrate in Fig. 7 and in our supplemental video. Note that in those results we merely linearly interpolate latent codes from one teeth configuration to another. In an actual orthodontic use case, the animation would probably need to contain additional "keyframes", based on orthodontic expert knowledge.

5.4 Comparison to related works

We compare our model to a number of implicit-based reconstruction methods, including the original DeepSDF [Park et al. 2019] and two more recent evolutions of it that both learn a template shape, DIT [Zheng et al. 2021] and DIF [Deng et al. 2021]. The latter is closest to our method, because it also uses Hyper-Nets to predict the weights of the deformation network. None of these methods, however use a component-wise representation and thus unlike our method they are unable to provide component-wise editing as we have shown in Sec. 5.3.

In Fig. 6 we compare our reconstruction results to those of the other methods. We observe that our reconstruction quality is clearly superior to that of DeepSDF or DIT. The comparison to DIF requires closer inspection: While at first sight DIF and our method seem to be on par, we consider the clearer gum line in our results as an advantage, given that this is also a clearly visible feature in the ground truth geometry. In the error maps we observe that while both methods struggle in similar areas of the gums, our method shows fewer and smaller erroneous regions on the teeth (see rows 1 and 3), especially when teeth are misaligned.

In Tab. 1 we compare the methods numerically: We extract meshes from our reconstructions by marching cubes. After sampling a point cloud from such a mesh and from the ground truth mesh, we can compute symmetric Chamfer distance and an F-score based on that, as was done in previous work [Yenamandra et al. 2021] (after applying a threshold of 0.01 to the Chamfer distances). In all of our results, the width of the mean bounding box around the teeth geometry is approximately 2.

Tab. 1 confirms that our method and DIF are very close in reconstruction quality and superior to the other methods. This is a very satisfying finding, given that our method is fulfilling the additional requirements of providing semantic labelling and allowing independent control over individual geometric components. DIF does not have these capabilities, but our method achieves them without compromising reconstruction quality at all.

Table 1. Quantitative comparison with related works. The reconstruction accuracy is evaluated by the symmetric Chamfer distance (lower is better) and F-score (higher is better). Our overall reconstruction accuracy is on par with DIF. However, DIF does not enable the novel applications our method is capable of (Sec. 5.3).

Metrics	Chamfer distance ↓	F-score ↑
DeepSDF	0.01497	42.132
DIT	0.01353	47.668
DIF	0.0058	**88.125**
Ours	**0.00552**	88.029

We also provide a comparison based on a publicly available 3D dental model dataset [Ben-Hamadou et al. 2022], which includes 595 dental models with ground truth labelling (after filtering out 2 duplicates and 3 cases with wisdom teeth). We split them into 545 for training (1090 after flipping data augmentation) and 50 for testing. The numerical results are shown in Tab. 2, confirming that our method achieves similar performance as DIF for the reconstruction task. All the methods yield better numerical results in Tab. 2 than Tab. 1 because the teeth in this dataset are more regular in shapes.

Table 2. Quantitative comparison to related works on a publicly available dataset [Ben-Hamadou et al. 2022]. The accuracy and F-scores are evaluated in the same way as Tab. 1. Similarly, our overall reconstruction accuracy is on par with DIF.

Metrics	Chamfer distance ↓	F-score ↑
DeepSDF	0.01196	53.761
DIT	0.01263	50.317
DIF	0.00514	**92.622**
Ours	**0.00463**	92.182

5.5 Ablation Study

To evaluate our design choices, we have conducted an ablation study, comparing our full method to variants that lack our centroid loss or our segmentation loss (see Sec. 4.2), as well as variants that omit the Deform-Net or train its weights directly, without obtaining

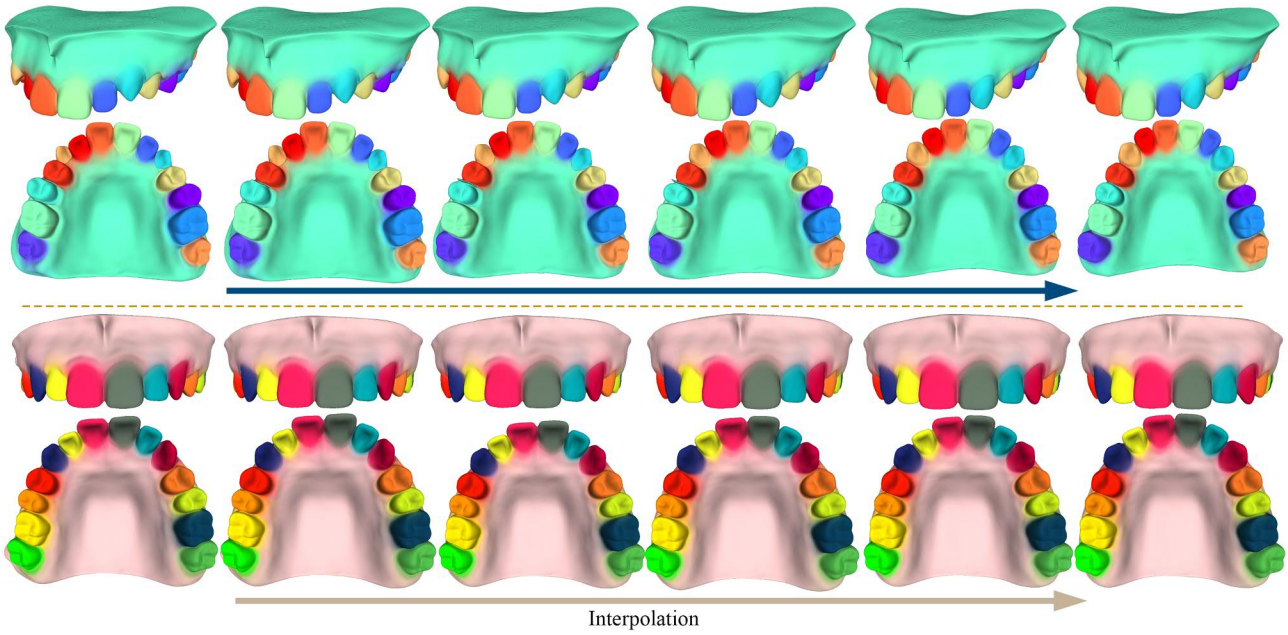


Fig. 7. In each row we interpolate between the reconstruction of a pre-treatment scan (first column) and the reconstruction of a post-treatment scan (last column). The arrows show the direction of interpolation. We can render plausible visualizations of orthodontic treatment plans in this way, which is best illustrated by our supplemental video results.

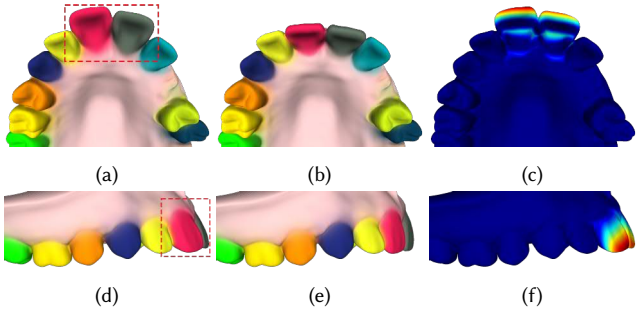


Fig. 8. Teeth replacement demonstration. (a) and (d) show two malaligned incisors from bottom and side view respectively (see dashed red). (b) and (e) show the result of replacing these two incisor teeth by some counterparts that are aligned better, while keeping all the other teeth unchanged. (c) and (f) encode the difference between before and after the edit. Note that the original model has no canine. Thus, we selected to process this example to show that we can reconstruct a model with originally missing teeth.

them via a Hyper-Net. Tab. 3 shows that removing any of these design choices strongly deteriorates reconstruction accuracy. Our centroid and segmentation losses, which are absolutely vital for segmenting the geometry into separate semantic components also seem to improve reconstruction quality, as numbers consistently get worse when any of them is omitted. The importance of these two losses, that we introduce in this work, is further highlighted by Figs. 9 and 10: Fig. 10 clearly shows that semantic labelling becomes very inaccurate when our segmentation loss is omitted. Note that the semantic labelling is not only a visualization but also an indicator of the locality of each individual component. When the semantic labelling is not accurate, the SDF contribution from each component is also not accurate. Fig. 9 is consistent with Tab. 3: While all the design choices including the deformation field and the use of Hyper-Net have clear impact on the final result, omitting either centroid or segmentation loss can still significantly increase the error, for example as observed in the molar on the right in the second row (see dashed red).

Table 3. Ablating the various design choices of our method. Our full method achieves the best results.

Metrics	Chamfer distance ↓	F-score ↑
Our full method	**0.005522**	**88.029**
w/o centroid loss	0.006644	83.306
w/o Segmentation loss	0.005985	86.706
w/o Deform-Net	0.01377	51.818
w/o Hyper-Net	0.01841	35.225

6 LIMITATIONS & FUTURE WORK

Even though the availability of a morphable teeth model with some control over individual teeth is a very useful contribution to the state of the art in this area, our model still has some limitations. For example, at reconstruction time, we assume that the user provides a binary vector indicating the presence/absence of teeth in their respective positions. Without this information, missing teeth might

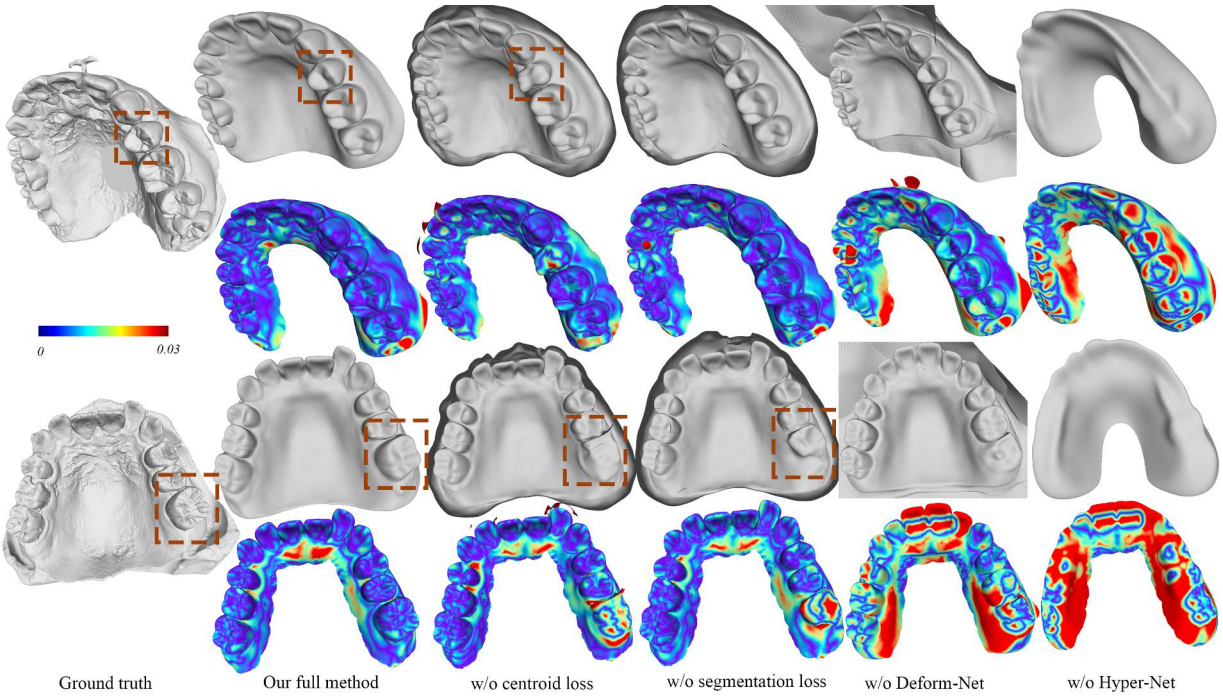
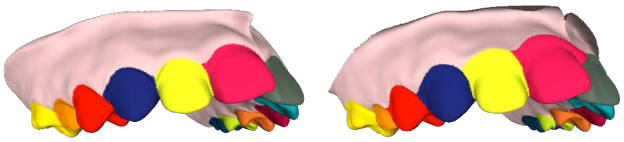
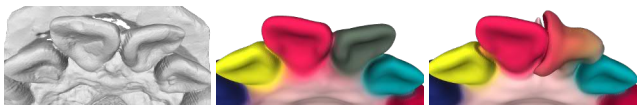


Fig. 9. Similar to Fig. 6, we visualize reconstruction results of our method and its ablated variants. As is confirmed by Tab. 3, our full method achieves the best results. Especially omitting our centroid loss or our segmentation loss can lead to severe artifacts (see dashed red).



(a) Our full method (b) w/o Segmentation loss

Fig. 10. Disabling our segmentation loss (b) clearly throws off the semantic labelling that our method (a) provides.



(a) (b) (c)

Fig. 11. If a boolean vector indicating the presence/absence of individual teeth is provided, the ground truth geomtry (a) can be faithfully reconstructed (b) by our method. If one gives an incorrect boolean vector, e.g. specifying that the right incisor (green in (b)) is missing, the resulting reconstruction (c) can have strong artifacts.

lead to flawed results, as shown in Fig. 11 While providing the binary vector is not too difficult for a user (such as a doctor or even a patient), it would be more satisfying to be able to infer it automatically. To make the model usable in a broader range of use cases, it would have to include texture, which we have not addressed in this work. In addition, we completely omit the modelling of the tongue, which would be one major step on the way from a dental model to a full-fledged intra-oral model. We believe the modelling of the tongue to be a very difficult problem, because it is not even clear what data can be acquired for tongues. A limitation much more easily covercome would be the extension of our model to the lower jaw, which does not require any changes to the method itself, but just training it on an additional dataset.

7 CONCLUSION

We have presented the first compositional, implicit neural representation for the modelling of teeth+gum geometry. Not only does our representation achieve state of the art quality in reconstruction (Sec. 5.2), but also it decomposes the geometry into a number of semantically meaningful components, i.e. into individual teeth and the gum. The benefits of this decomposition are two-fold: First, it makes our reconstruction approach provide a semantic labelling for teeth geometry as an additional by-product (Fig. 4). Second, it shows interesting editing applications, such as the replacement of individual teeth by more aesthetically desirable alternatives (Sec. 5.3). Together with the fact that our model can be used for rendering smooth interpolations between different teeth states, it could be a valuable tool for the communication between orthodontists and their patients.

Beyond the concrete domain of teeth geometry, the contributions of this work lie in the way we achieve the local-based decomposition of complex geometry into smaller semantically meaningful components, which we hope can be beneficial in other domains as well. We will release the pre-trained model and the inference code,

which will make it the first publicly available teeth model of its kind.

ACKNOWLEDGMENTS

This work is partially supported by the ERC Consolidator Grant 4DRepLy (770784), GRF grants (17210419 and 17212120) from the RGC of Hong Kong and by Seed Fund for Basic Research (202111159232) from Hong Kong. The authors would like to thank Yanhong Lin, Zhiming Cui, and Ayush Tewari for their instructive suggestion and generous help on the project.

REFERENCES

Aly S. Abdelrehim, Aly A. Farag, Ahmed M. Shalaby, and Moumen T. El-Melegy. 2014. 2D-PCA Shape Models: Application to 3D Reconstruction of the Human Teeth from a Single Image. In *Medical Computer Vision. Large Data in Medical Imaging*, Bjoern Menze, Georg Langs, Albert Montillo, Michael Kelm, Henning Müller, and Zhuowen Tu (Eds.). Springer International Publishing, Cham, 44–52.

Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. 2021. imGHUM: Implicit Generative Models of 3D Human Shape and Articulated Pose. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 5441–5450. https://doi.org/10.1109/ICCV48922.2021.00541

K. S. Arun, T. S. Huang, and S. D. Blostein. 1987. Least-Squares Fitting of Two 3-D Point Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-9, 5 (1987), 698–700. https://doi.org/10.1109/TPAMI.1987.4767965

Thabo Beeler and Derek Bradley. 2014. Rigid Stabilization of Facial Expressions. *ACM Trans. Graph.* 33, 4, Article 44 (July 2014), 9 pages.

Achraf Ben-Hamadou, Oussama Smaoui, Houda Chaabouni-Chouayakh, Ahmed Rekik, Sergi Pujades, Edmond Boyer, Julien Strippoli, Aurélien Thollot, Hugo Setbon, Cyril Trosset, and Edouard Ladroit. 2022. Teeth3DS: a benchmark for teeth segmentation and labeling from intra-oral 3D scans. *arXiv preprint* (2022).

Rohan Chabra, Jan E. Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. 2020. Deep Local Shapes: Learning Local SDF Priors for Detailed 3D Reconstruction. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 608–625.

Zhiqin Chen and Hao Zhang. 2019. Learning Implicit Fields for Generative Shape Modeling. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5932–5941. https://doi.org/10.1109/CVPR.2019.00609

Zhang Chen, Yinda Zhang, Kyle Genova, Sean Fanello, Sofien Bouaziz, Christian Häne, Ruofei Du, Cem Keskin, Thomas Funkhouser, and Danhang Tang. 2021. Multiresolution Deep Implicit Functions for 3D Shape Representation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 13067–13076. https://doi.org/10.1109/ICCV48922.2021.01284

Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. 2022. LISA: Learning Implicit Shape and Appearance of Hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20533–20543.

Zhiming Cui, Changjian Li, Nenglun Chen, Guodong Wei, Runnan Chen, Yuanfeng Zhou, and Wenping Wang. 2021. TSegNet: An efficient and accurate tooth segmentation network on 3D dental model. *Medical Image Analysis* 69 (2021), 101949. https://doi.org/10.1016/j.media.2020.101949

Boyang Deng, J. P. Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. 2020. NASA Neural Articulated Shape Approximation. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 612–628.

Yu Deng, Jiaolong Yang, and Xin Tong. 2021. Deformed Implicit Field: Modeling 3D Shapes with Learned Dense Correspondence. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10281–10291. https://doi.org/10.1109/CVPR46437.2021.01015

Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 2020. 3D Morphable Face Models—Past, Present, and Future. *ACM Trans. Graph.* 39, 5, Article 157 (jun 2020), 38 pages. https://doi.org/10.1145/3395208

Aly Farag, Shireen Elhabian, Aly Abdelrehim, Wael Aboelmaaty, Allan Farman, and David Tasman. 2013. Model-Based Human Teeth Shape Recovery from a Single Optical Image with Unknown Illumination. In *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, Bjoern H. Menze, Georg Langs, Le Lu, Albert Montillo, Zhuowen Tu, and Antonio Criminisi (Eds.). 263–272.

Flawless 2022. Flawless AI. https://www.flawlessai.com/.

Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. 2020. Local Deep Implicit Functions for 3D Shape. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4856–4865. https://doi.org/10.1109/CVPR42600.2020.00491

John C Gower. 1975. Generalized procrustes analysis. *Psychometrika* 40, 1 (1975), 33–51.

Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017. Avatar Digitization from a Single Image for Real-Time Rendering. *ACM Trans. Graph.* 36, 6, Article 195 (nov 2017), 14 pages.

ISO-3950 2016. Dentistry — Designation system for teeth and areas of the oral cavity. https://www.iso.org/standard/68292.html.

Maik Keller, Damien Lefloch, Martin Lambers, Shahram Izadi, Tim Weyrich, and Andreas Kolb. 2013. Real-Time 3D Reconstruction in Dynamic Scenes Using Point-Based Fusion. In *2013 International Conference on 3D Vision - 3DV 2013*. 1–8. https://doi.org/10.1109/3DV.2013.9

Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. 2019. Neural Style-Preserving Visual Dubbing. *ACM Trans. Graph.* 38, 6, Article 178 (nov 2019), 13 pages. https://doi.org/10.1145/3355089.3356500

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.* 37, 4, Article 68 (July 2018), 13 pages.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4455–4465. https://doi.org/10.1109/CVPR.2019.00459

Eslam Mostafa, Shireen Elhabian, Aly Abdelrahim, Salwa Elshazly, and Aly Farag. 2014. Statistical morphable model for human teeth restoration. In *2014 IEEE International Conference on Image Processing (ICIP)*. 4285–4288. https://doi.org/10.1109/ICIP.2014.7025870

Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, and Xiaolong Wang. 2021. A-SDF: Learning Disentangled Signed Distance Functions for Articulated Shape Representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13001–13011.

Yahini Prabha Murugesan, Abeer Alsadoon, Paul Manoranjan, and P. W.C. Prasad. 2018. A novel rotational matrix and translation vector algorithm: Geometric accuracy for augmented reality in oral and maxillofacial surgeries. *International Journal of Medical Robotics and Computer Assisted Surgery* 14, 3 (June 2018), 1–14.

Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. 2013. Real-Time 3D Reconstruction at Scale Using Voxel Hashing. *ACM Trans. Graph.* 32, 6, Article 169 (nov 2013), 11 pages. https://doi.org/10.1145/2508363.2508374

Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. 2021. NPMs: Neural Parametric Models for 3D Deformable Shapes. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 12675–12685. https://doi.org/10.1109/ICCV48922.2021.01246

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 165–174. https://doi.org/10.1109/CVPR.2019.00025

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5865–5874.

Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. 2020. Convolutional Occupancy Networks. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 523–540.

Yuhan Ping, Guodong Wei, Lei Yang, Zhiming Cui, and Wenping Wang. 2021. Self-attention implicit function networks for 3D dental data completion. *Computer Aided Geometric Design* 90 (2021), 102026. https://doi.org/10.1016/j.cagd.2021.102026

Olinde Rodrigues. 1816. De l'attraction des sphéroïdes. In *Correspondence Sur l'École Impériale Polytechnique*. 361–385.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Trans. Graph.* 36, 6, Article 245 (nov 2017), 17 pages. https://doi.org/10.1145/3130800.3130883

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. 2020. Implicit Neural Representations with Periodic Activation Functions. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 7462–7473. https://proceedings.neurips.cc/paper/2020/file/53c04118df112c13a8c34b38343b9c10-Paper.pdf

Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. 2019. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/b5dc4e5d9b495d0196f61d45b26ef33e-Paper.pdf

Synthesia 2022. Synthesia. https://www.synthesia.io/.

Ayush Tewari, Mohamed Elgharib, Mallikarjun B R, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2020. PIE: Portrait Image Embedding for Semantic Control. *ACM Trans. Graph.* 39, 6, Article 223 (nov 2020), 14 pages. https://doi.org/10.1145/3414685.3417803

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis Using Neural Textures. *ACM Trans. Graph.* 38, 4, Article 66 (jul 2019), 12 pages. https://doi.org/10.1145/3306346.3323035

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2387–2395. https://doi.org/10.1109/CVPR.2016.262

Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Carsten Stoll, and Christian Theobalt. 2020. PatchNets: Patch-Based Generalizable Deep Implicit 3D Shape Representations. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 293–309.

Zdravko Velinov, Marios Papas, Derek Bradley, Paulo Gotardo, Parsa Mirdehghan, Steve Marschner, Jan Novák, and Thabo Beeler. 2018. Appearance Capture and Modeling of Human Teeth. *ACM Trans. Graph.* 37, 6, Article 207 (2018), 13 pages.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10034–10044. https://doi.org/10.1109/CVPR46437.2021.00991

Guodong Wei, Zhiming Cui, Yumeng Liu, Nenglun Chen, Runnan Chen, Guiqing Li, and Wenping Wang. 2020. TANet: Towards Fully Automatic Tooth Arrangement. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). 481–497.

Andreas Wirtz, Florian Jung, Matthias Noll, Anqi Wang, and Stefan Wesarg. 2021. Automatic model-based 3-D reconstruction of the teeth from five photographs with predefined viewing directions. In *Medical Imaging 2021: Image Processing*, Ivana Išgum and Bennett A. Landman (Eds.), Vol. 11596. International Society for Optics and Photonics, SPIE, 198 – 212. https://doi.org/10.1117/12.2582253

Chenglei Wu, Derek Bradley, Pablo Garrido, Michael Zollhöfer, Christian Theobalt, Markus Gross, and Thabo Beeler. 2016. Model-Based Teeth Reconstruction. *ACM Trans. Graph.* 35, 6, Article 220 (nov 2016), 13 pages. https://doi.org/10.1145/2980179.2980233

Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. 2020. PQ-NET: A Generative Part Seq2Seq Network for 3D Shapes. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 826–835. https://doi.org/10.1109/CVPR42600.2020.00091

Lingchen Yang, Zefeng Shi, Yiqian Wu, Xiang Li, Kun Zhou, Hongbo Fu, and Youyi Zheng. 2020. iOrthoPredictor: Model-Guided Deep Prediction of Teeth Alignment. *ACM Trans. Graph.* 39, 6, Article 216 (nov 2020), 15 pages. https://doi.org/10.1145/3414685.3417771

Wenwu Yang, Nathan Marshak, Daniel Sỳkora, Srikumar Ramalingam, and Ladislav Kavan. 2019. Building anatomically realistic jaw kinematics model from data. *The Visual Computer* 35, 6-8 (2019), 1105–1118.

Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2021. i3DMM: Deep Implicit 3D Morphable Model of Human Heads. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12798–12808. https://doi.org/10.1109/CVPR46437.2021.01261

K. Yin, Z. Chen, S. Chaudhuri, M. Fisher, V. G. Kim, and H. Zhang. 2020. COALESCE: Component Assembly by Learning to Synthesize Connections. In *2020 International Conference on 3D Vision (3DV)*. IEEE Computer Society, Los Alamitos, CA, USA, 61–70. https://doi.org/10.1109/3DV50981.2020.00016

Lingming Zhang, Yue Zhao, Deyu Meng, Zhiming Cui, Chenqiang Gao, Xinbo Gao, Chunfeng Lian, and Dinggang Shen. 2021. TSGCNet: Discriminative Geometric Feature Learning With Two-Stream Graph Convolutional Network for 3D Dental Model Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6699–6708.

Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. 2022. ImFace: A Nonlinear 3D Morphable Face Model With Implicit Neural Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20343–20352.

Shu-Xian Zheng, Jia Li, and Qing-Feng Sun. 2011. A novel 3D morphing approach for tooth occlusal surface reconstruction. *Computer-Aided Design* 43, 3 (2011), 293–302.

Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. 2021. Deep Implicit Templates for 3D Shape Representation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1429–1439. https://doi.org/10.1109/CVPR46437.2021.00148

Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. 2018a. State of the Art on 3D Reconstruction with RGB-D Cameras. *Computer Graphics Forum* 37, 2 (2018), 625–652. https://doi.org/10.1111/cgf.13386 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13386

M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. 2018b. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum* 37, 2 (2018), 523–550. https://doi.org/10.1111/cgf.13382 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13382