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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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. ACM Trans. Graph., Vol. 41, No. 6, Article 217. Publication date: December 2022.
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 reliably 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; Faláfox 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 tooth processing with emphasis 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 point 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 DIF 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. Abdelrehim 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 current 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 submodule \( i \) predicts an SDF value \( s_i + \Delta s_i \) and an indicator \( \delta_i \). 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 \( \delta_i \) 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 shape, in which a template shape is embedded by a Ref-Net (see Fig. 3b). Deform-Net also predicts the SDF compensation \( \Delta s_i \) 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 digitizing 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 sub-spaces 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 \( \{ z_i \}_{i=1}^m \) of gum and teeth latent codes (where \( m = 15 \) is the number of geometric components in our implementation), and maps arbitrary spatial locations \( p \) to the signed distance \( s \) between \( p \) and the surface:

\[
f(p, z_1, \ldots, z_m) = s
\]

where the set \( \{ p \mid f(p, z_1, \ldots, z_m) = 0 \} \) 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 \( \{ f_i \}_{i=1}^m \) of sub-functions:

\[
f(p, z_1, \ldots, z_m) = \sum_{i=1}^m w_i \cdot f_i(p, z_i) = \sum_{i=1}^m w_i \cdot (s_i + \Delta s_i)
\]

where \( w_i \) are blending weights \( s_i \) are signed distance values for component \( i \) and \( \Delta s_i \) 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 \( f_i(p, z_i) = s_i + \Delta s_i \), 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 \( \mathcal{R}_i(p_i) = s_i \) predicts SDF values \( s_i \) for the \( i \)-th reference shape, independent of the latent code \( z_i \). It is queried at those points \( p_i \) that are produced by

ACM Trans. Graph., Vol. 41, No. 6, Article 217. Publication date: December 2022.
the deformation network $\Phi_i(p, z_i) = (p', \Delta s_i, \delta_i)$. The deformation network is supposed to warp the space in which the reference shape is embedded, such that querying $R_i$ at the points $p_i$ results in the shape encoded by the $z_i$.

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 $\Delta p$ of each query point $p$ so that $p' = p + \Delta p$. 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 $(r; t) \in \mathbb{R}^6$ and set $p' = e^r p + t$ by Rodrigues’ formula [Rodrigues 1816].

Note that the weights of $\Phi_i$ are themselves the output of a so-called Hyper-Net $\Psi_i(z_i)$, a mechanism introduced in recent works [Deng et al. 2021; Sitzmann et al. 2020]. We differ from previous uses, however, in that $\Psi_i$ not only predicts the deformed points $p'_i$ and SDF value correction deltas $\Delta s_i$, 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 $\bar{c}_i$ of tooth $i$ leads to a result $\bar{c}'_i$, that should coincide with the average centroid position $\bar{c}_i$ of the training data for that tooth, which we can precompute before training. We penalize the $l^2$ distance between the predicted and the expected centroid:

$$L_{\text{centroid}} = \| \bar{c}'_i - \bar{c}_i \|_2$$

(3)

**Segmentation loss.** Our deformation network $\Phi_i$ produces values $\delta_i$ that indicate a confidence with which a particular point belongs to geometric component $i$ (where components can be any of the teeth, or the gums). 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 $\delta_i$ 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 $\delta_i$ 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:

$$L^\text{BCE}_i = \sum_{p \in S_i} \text{BCE}(\delta_i(p), \ell_i(p) = i)$$

(4)

where $S_i$ is the tooth or gum surface and $\ell_i$ is the ground truth label for point $p$. 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 $R_T$ to be close to the mean shape of component $i$. We thus constrain the deformation field $\Phi_i$ to be smooth, as in [Deng et al. 2021]:

$$L^\text{smooth}_i = \sum_{p \in \Omega} \|\nabla (p_i'(p) - p)\|_2$$

(5)

where $\nabla (p_i'(p) - p)$ is the Jacobian of deformation offset with respect to the coordinates of $p$ and $\Omega$ 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:

$$L^\text{SDF}_i = \sum_{p \in S_i} |f_i(p)| + (1 - \langle \nabla f_i(p), \hat{n} \rangle) + \sum_{p \in \Omega \backslash S_i} |\nabla f_i(p)|_2 - 1 + \sum_{p \in \Omega \backslash S_i} \psi(f_i(p))$$

(6)

where $\psi(x) = \exp(-\alpha \cdot |x|)$ with $\alpha \gg 1$. The first sum in this term encourages points on the surface $S_i$ 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 $\hat{n}$. 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:

$$L^\text{SDF}_g = \sum_{p \in S} |f(p)| + (1 - \langle \nabla f(p), \hat{n} \rangle) + \sum_{p \in \Omega \backslash S} |\nabla f(p)|_2 - 1 + \sum_{p \in \Omega \backslash S} \psi(f(p))$$

(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 $R$. 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:

$$L^\text{corres}_i = \sum_{p \in S_i} (1 - \langle \nabla R_i(p'), \hat{n} \rangle)$$

(8)

where $\nabla R_i$ is the spatial gradient of the reference net, and $\hat{n}$ is the ground truth normal of the query point $p$ on the original surface $S_i$.

**SDF correction regularization loss.** To make sure that SDF correction values $\Delta \lambda_i$ 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]

$$L^\text{correction}_i = \sum_{p \in \Omega} |\Delta \lambda_i(p)|$$

(9)

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

$$L^\text{latent}_i = \|z_i\|^2$$

(10)

In summary, we optimize the following objective:

$$L = \frac{1}{|I|} \sum_{i \in I} (\lambda_1 L^\text{centroid}_i + \lambda_2 L^\text{SDF}_i + \lambda_3 L^\text{corres}_i + \lambda_4 L^\text{smooth}_i + \lambda_5 L^\text{correction}_i + \lambda_6 L^\text{SDF}_g + \lambda_7 L^\text{SDF}_g + \lambda_8 L^\text{SDF}_g)$$

(11)

Here, $I$ 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. $\lambda_1$ to $\lambda_7$) 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 $\lambda_1 = 1$, $\lambda_2 = 10^2$, $\lambda_3 = 10^2$, $\lambda_4 = 50$, $\lambda_5 = 10^3$, $\lambda_6 = 10^3$, $\lambda_7 = 1$, $\lambda_8 = 0.1$. 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:

$$\arg\min_{\{z_i\}} \left( L_{SDF} + \lambda_{\text{latent}} \sum_{i \in I} L_{i}^{\text{latent}} \right)$$

where $I$ 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 the gums, thus enabling interesting editing applications (see Sec. 5.3).
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 to 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.

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     | 0.005322           | 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...
an implicit parametric morphable dental model

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 geometry (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 overcome 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 (17201419 and 17212120) from the RGC of Hong Kong and by Seed Fund for Basic Research (20211159232) 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-Megly. 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, Björn Menze, Georg Langs, Albert Montillo, Michael Kelm, Henning Müller, and Zhounuo Tu (Eds.). Springer International Publishing, Cham, 44–52.

Thiemo Alldreck, Hongyi Xu, and Cristian Sminchisescu. 2021. imgHUM: Generative Models of 3D Human Shape and Articulated Pose. In 2021 IEEE 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 144 (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 Tosset, and Edoardo Ladoitro. 2022. Teeth3DS: a benchmark for tooth segmentation and labeling from intra-oral 3D scans. arXiv preprint (2022).

Rohan Chhabra, Jan E. Lensen, Eddy Igl, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. 2020. Deep Local Shapes: Learning Local SDF Prions 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/ACM International Conference on Computer Vision (ICCV). 10967–10976. https://doi.org/10.1109/ICCV48922.2021.01284

Enric Corona, Tomas Hodan, Minik Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingui 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, Changjiang Li, Nenglun Chen, Guodong Wei, Runman Chen, Yuancheng Zhou, and Wengang Wang. 2021. TsegNet: An efficient and accurate tooth segmentation network on 3D dental model. Medical Image Analysis 69 (2021), 101945. 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/CVPR44457.2021.01015

Bernhard Egger, William A. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Achraf Ben-Hamadou, Oussama Smaoui, Houda Chaabouni-Chouayakh, Ahmed Rekik, Sergi Pujades, Edmond Boyer, Julien Strippoli, Aurélien Thollot, Hugo Setbon, Cyril Tosset, and Edoardo Ladoitro. 2022. Teeth3DS: a benchmark for tooth segmentation and labeling from intra-oral 3D scans. arXiv preprint (2022).

Aly Farag, Shireen Elhabian, Aly Abdelrehim, Wael Abdoelmaaty, 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, Björn H. Menze, Georg Langs, Le Lu, Albert Montillo, Zhounuo Tu, and Antonio Criminisi (Eds.) 263–272.

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