Cross-Species 3D Face Morphing via Alignment-Aware Controller

Xirui Yan, Zhenbo Yu, Bingbing Ni*, Hang Wang

Shanghai Jiao Tong University, Shanghai 200240, China
{xiruiYan, yuzhenbo, nibingbing, Wang–Hang} @sjtu.edu.cn

Abstract

We address cross-species 3D face morphing (*i.e.*, 3D face morphing from human to animal), a novel problem with promising applications in social media and movie industry. It remains challenging how to preserve target structural information and source fine-grained facial details simultaneously. To this end, we propose an *Alignment-aware 3D Face Morphing (AFM)* framework, which builds semantic-adaptive correspondence between source and target faces across species, via an alignment-aware controller mesh (*Explicit Controller, EC*) with explicit source/target mesh binding. Based on EC, we introduce *Controller-Based Mapping (CBM)*, which builds semantic consistency between source and target faces according to the semantic importance of different face regions. Additionally, an inference-stage coarse-to-fine strategy is exploited to produce fine-grained meshes with rich facial details from rough meshes. Extensive experimental results in multiple people and animals demonstrate that our method produces high-quality deformation results.

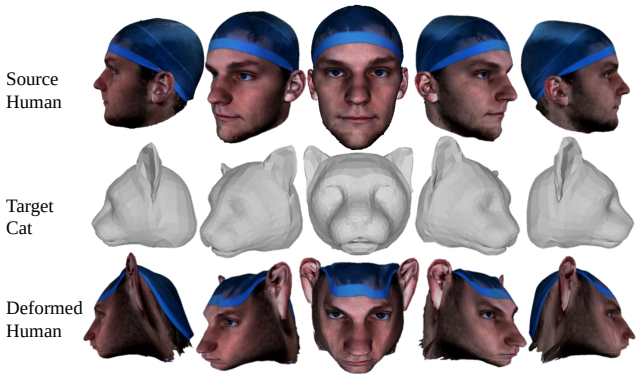

Figure 1: Visualization of the proposed alignment-aware 3D face morphing (AFM) framework. Morphing a textured human face of rich details into a cat.

Introduction

3D face morphing with artificial design has received considerable attention in computer vision community for a long time. Previous methods mainly fall into two categories: 1) Deforming a 3D face towards different expressions and shapes (Smith et al. 2020; Li et al. 2020; Geng, Cao, and Tulyakov 2019; Jiang et al. 2019; Ranjan et al. 2018; Chandran et al. 2020), and 2) Transferring deformations of source faces to control a new target avatar (*e.g.*, face retargeting or reenactment) (Thies et al. 2016; Chaudhuri, Vesdapunt, and Wang 2019; Ouzounis, Kilias, and Mousas 2017; Ribera et al. 2017; Gao et al. 2020; Yao et al. 2020; Le and Deng 2017). These works do not consider directly morphing 3D human faces to different structures such as human-to-animal morphing where the structures and feature details of source human faces and target animal faces are quite different.

In this study, we are interested in morphing a 3D human face mesh into a specific 3D animal face mesh while preserving human features, *i.e.*, *cross-species 3D face morphing*. As shown in Fig. 1, the human face mesh is morphed into a cat while the human details are preserved, including human identity features, face expression, wrinkles and so on. This task has a huge number of application potentials, *e.g.*, 3D animated characters (Ho, Sun, and Tsai 2019), virtual reality (Clay, König, and König 2019) and game simulation (Wang et al. 2019a). To our knowledge, it is the first attempt at this fancy but challenging problem.

Specifically, challenges of cross-species 3D face morphing are two folds: **1) Shape alignment and human feature preservation**. Due to the large gap between the geometric structures of 3D human faces and animal faces, it is difficult to learn the deformation that satisfies two competing objectives simultaneously, that is animal shape alignment and human feature preservation. Neural Cages (NC) (Wang et al. 2020) introduces cage-based deformation to preserve surface details regardless of the shape's intricacy and performs well on deforming chairs, tables and cars. However, we empirically show that NC is only capable of performing rough deformations on human faces when applied to cross-species 3D face morphing, as illustrated in Fig. 2(a). The reason is that the cage used in NC is rough as a result of being implicitly trained in network without any cage-specific constraints. This adds much regularization to the deformation process and limits the degrees of freedom to deform the source objects, which makes NC lack the ability of fine deformation. **2) Semantic consistency.** Semantic mismatch

*Corresponding author: Bingbing Ni.

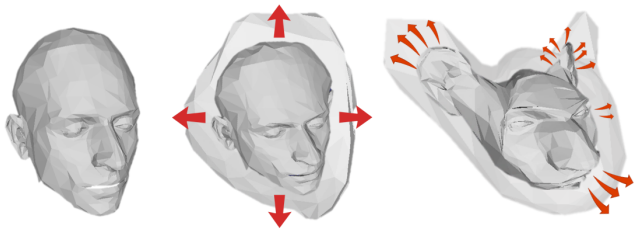

Figure 2: The baseline (a) using Neural Cages (Wang et al. 2020) only performs rough morphing on the human face due to the limit of the rough cage. Instead, the proposed AFM (b) achieves semantic-adaptive fine deformation of the human face via an alignment-aware controller.

across species is a challenging problem to be solved. For example, the positions of the ears, eyes, and chin between the human face and animal face should be consistent after morphing. The change of semantic information in the morphing process is uncontrollable since there is no ground truth of the deformed human face.

To address the issues above, we propose an *Alignment-aware 3D Face Morphing* (AFM) framework to build the uniform semantic-adaptive correspondence across species. To achieve fine control of face morphing while preserving human feature details, we propose *Explicit Controller* (EC), a simplified mesh that encloses the object and transfers the deformation to the object by weight binding. The weight binding scheme interpolates the translations of the controller vertices and is carefully designed to preserve the source features. EC is alignment-aware under being trained with explicit constraints and thus yields a structured deformation operation set, which is sufficient for fine deformation. To maintain the semantic consistency across species, we propose *Controller-Based Mapping* (CBM), which takes advantage of ECs to map the semantic information between human and animal faces. The proposed CBM builds correspondences between source ECs and target ECs in a semantic importance manner, *i.e.*, face regions with more dedicated local structures such as eyes and ears are constrained with denser correspondences while regions with plain structures receive less attention. In this way, the subsequent shape deformation will be conducted in a semantic-adaptive manner. As demonstrated in Fig. 2(b), the proposed AFM achieves semantic-adaptive fine deformation of human faces, which significantly outperforms the prior art (Wang et al. 2020). Notably, our method is end-to-end trainable as a whole framework, where the output (mesh) of each step is explainable and intuitive.

Furthermore, in recent learning-based deformation works related to our study (Wang et al. 2020; Pan et al. 2019; Tretschk et al. 2020), a well-trained model could only work on meshes with a similar number of vertices during training and inference stage, usually no more than 5,000 vertices which is sparse. In real applications, a human face mesh with dense vertices cannot be processed by the trained model directly. We propose an inference-stage sparse-to-dense strategy to solve this problem. Extensive experiments validate the effectiveness of the proposed method on deforming human faces to animal faces of multiple species.

Related Work

Human Face Morphing. Recently, 3D human face morphing has increasingly attracted the interest of researchers and could be divided into two aspects: deforming directly on 3D human faces and deformation transfer. 1) For deforming directly on 3D human faces, Blanz and Vetter (Blanz and Vetter 1999) derived a morphable face model known as 3DMM. Following this framework, many researchers (Smith et al. 2020; Li et al. 2020; Geng, Cao, and Tulyakov 2019; Jiang et al. 2019; Ranjan et al. 2018; Chandran et al. 2020; Chen and Kim 2021; Bailey et al. 2020) directly deformed the original 3D human face meshes to have different expressions. In these works, only facial expressions changed after the deformation while the overall shapes were still similar. Some recent works of 3D face reconstruction from 2D images devoted to learning the deformation of a template mesh to generate human face meshes (Zhou et al. 2019; Lee and Lee 2020a; Sidhu et al. 2020; Chaudhuri et al. 2020; Lee and Lee 2020b; R et al. 2021) but still did not pay attention to the change of semantic information during morphing. 2) Researchers also made significant progress in deformation transfer, *i.e.*, face reenactment or face retargeting. Some works (Thies et al. 2016; Chaudhuri, Vesdapunt, and Wang 2019; Ouzounis, Kilias, and Mousas 2017; Ribera et al. 2017; Gao et al. 2020; Yao et al. 2020; Le and Deng 2017; Zhang, Chen, and Zheng 2020; Bai et al. 2021; Sumner and Popović 2004) retargeted the facial animation in a new character. These works controlled an avatar by human faces rather than morphing human face meshes themselves. Different from the above works, we make the first attempt at cross-species 3D human face morphing.

Cage-Based Deformation. Sederberg (Sederberg and Parry 1986) firstly proposed cage, a simple mesh enclosing the object, to deform any solid geometric models in a free-form manner. All the surface points on the object were binding to the cage vertices by linear combinations with weights. Many works (Hoppe 1997; Xia and Varshney 1996; Sander et al. 2000; Calderon and Boubekeur 2017; Sacht, Vouga, and Jacobson 2015; Joshi et al. 2007) focused on how to compute the weights. Ju (Ju, Schaefer, and Warren 2005) generalized Mean Value Coordinates from closed 2D polygons to closed triangular meshes. The works (Huang et al. 2006; Zhang et al. 2020; Ju et al. 2008; Chen et al. 2010; Sung et al. 2020; Ramachandran et al. 2018) used manually calibrated cages or optimization-based cages to deform objects, which suffered from heavy computation overheads. Neural Cages (Wang et al. 2020) introduced neural networks to learn the deformation on cages. However, the cages in (Wang et al. 2020) conveyed no semantic information and could not handle fine deformation since it was not constrained explicitly. (Jakab et al. 2021) also introduced cages, but it only performs rough deformation on objects from the same object category. Differently, we propose an explicit controller based on cages and controller-based mapping to achieve the semantic-adaptive fine deformation on

human faces.

Method

In this paper, a novel Alignment-aware 3D Face Morphing Framework called AFM is proposed to address cross-species 3D face morphing. Explicit Controller (EC) and Controller-Based Mapping (CBM) are proposed to build semantic-adaptive correspondences between human and animal faces. The encoder-decoder structure is used to predict controllers and learn deformations. In this section, we start from revisiting the cage, as the weight binding and key ideas of our method are inspired from this technique.

Preliminary: Revisiting Cage

Cage is a closed mesh surrounding the object, which is also called bounding proxy(Calderon and Boubekeur 2017). The control of the cage to the object is realized by binding the cage to the object with the weight computed by a designed interpolation function. Given a mesh $\mathcal{M} = (\mathcal{V}_\mathcal{M}, \mathcal{F}_\mathcal{M})$ and a cage $\mathcal{C} = (\mathcal{V}_\mathcal{C}, \mathcal{F}_\mathcal{C})$, where $\mathcal{V}_\mathcal{M}$ and $\mathcal{V}_\mathcal{C}$ are sets of vertices, $\mathcal{F}_\mathcal{M}$ and $\mathcal{F}_\mathcal{C}$ are sets of triangles, the i-th vertex $\mathbf{v}_i^m \in \mathcal{V}_\mathcal{M}$ can be expressed as an interpolation result by the designed weight w:

$$\mathbf{v}_i^m = \sum_{0 \le j \le |\mathcal{V}_\mathcal{C}|} w_{i,j} \mathbf{v}_j^c, \tag{1}$$

where $\mathbf{v}_j^c \in \mathcal{V}_c$ and $w_{i,j}$ is the weight of the j-th vertex of $\mathcal{C}$ to the i-th vertex of $\mathcal{M}$. Mean value coordinates (MVC) (Ju, Schaefer, and Warren 2005) are used to compute the weight w with good feature preservation properties and differentiability as:

$$w = \mathrm{MVC}(\mathcal{V}, \mathcal{V}_\mathcal{C}, \mathcal{F}_\mathcal{C}). \tag{2}$$

The computed weight w could be used for deforming a cage into a new one, where the corresponding deformed mesh of $\mathcal{M}$ is computed by w following Eq. 1.

Explicit Controller

Neural Cages (Wang et al. 2020) is generally well-performing in feature preservation when deforming objects, while it has limited performance on fine deformation. To this end, we propose the Explicit Controller (EC), a controller mesh with explicit constraints to perform fine deformations on the human face. EC has a desirable shape aligned with the source human and the target animal respectively. Therefore, the deformation of the enclosed object can be restricted by explicitly restricting the deformation of the controller. EC provides the basis for subsequently building multi-density correspondences between human and animals based on semantic information, thereby realizing semantic-adaptive cross-species morphing.

Network Architecture. As illustrated in Fig. 3, the whole network architecture consists of 3 sub-networks: 1) a shape encoder Φ_E to embed the shape into a latent space, 2) a controller generator Φ_G to generate ECs for source and target respectively, *i.e.*, source controller and target controller, and 3) a deformation decoder Φ_D to predict the deformed controller from the source controller. The three networks work on mesh vertices as those in Neural Cages (Wang et al. 2020). In addition, to strengthen the shape alignment, we further introduce differentiable rendering layers.

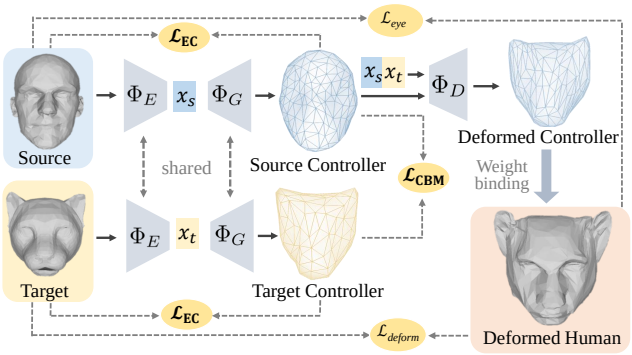

Figure 3: An overview of the proposed Alignment-aware 3D Face Morphing (AFM) framework. We first compute the Explicit Controllers (*i.e.*, source and target controllers) of source and target meshes, respectively. Semantic transformation is applied on the source controller to obtain a deformed controller via Controller-Based Mapping, from which the deformed human mesh is recovered via weight binding. Φ_E, Φ_G and Φ_D are three neural networks to process mesh vertices.

The inputs of AFM are a source human face mesh $\mathcal{M}_s = (\mathcal{V}_s, \mathcal{F}_s)$, a target animal face mesh $\mathcal{M}_t = (\mathcal{V}_t, \mathcal{F}_t)$ and several predefined landmarks. First of all, $\mathcal{M}_s$ and $\mathcal{M}_t$ are sent to the encoder Φ_E respectively to get embedding vectors x_s and x_t. Then Φ_G uses the embedding vector x_s / x_t as input to predict EC meshes $\mathcal{C}_s$ / $\mathcal{C}_t$ from a controller template (a 322-vertex sphere). The decoder Φ_D uses x_s, x_t and $\mathcal{C}_s$ as input, and outputs a deformed controller $\mathcal{C}_{s\to t}$ which is an offset from $\mathcal{C}_s$. Finally, according to the weights precomputed by Eq. 2 for $\mathcal{M}_s$ and $\mathcal{C}_s$, an output deformed mesh $\mathcal{M}_{s\to t}$ is attained by Eq. 1.

Shape Alignment Loss. To measure the shape difference between two meshes $\mathcal{M}_1$ and $\mathcal{M}_2$, we compute the Chamfer distance $\mathcal{L}_{\mathrm{CD}}$ to directly measure distances over 3D surfaces, and 2D rendering loss $\mathcal{L}_{\mathrm{R}}$ to measure distances over 2D silhouettes. We adopt the open-sourced soft rasterizer (Liu et al. 2019) for differentiable rendering, where gradients of the 2D image loss could be back-propagated to 3D coordinates of meshes. To strengthen the similarity of 3D meshes silhouettes, multi-view 2D rendering loss is designed. Assuming $\mathbf{A}$ represents the view point angles, the rendering image $\mathbf{I}_\mathcal{M}(\mathbf{A})$ of the mesh $\mathcal{M}$ can be denoted as:

$$\mathbf{I}_\mathcal{M}(\mathbf{A}) = \mathcal{R}(\mathcal{M}, \mathbf{A}), \tag{3}$$

where $\mathcal{R}$ represents the rendering layer. $\mathcal{L}_{\mathrm{CD}}$ and $\mathcal{L}_{\mathrm{R}}$ are computed as:

$$\mathcal{L}_{\mathrm{CD}}(\mathcal{M}_1, \mathcal{M}_2) = \sum_{\mathbf{x}\in\mathcal{V}_1} \min_{\mathbf{y}\in\mathcal{V}_2} \|\mathbf{x}-\mathbf{y}\|_2^2 + \sum_{\mathbf{y}\in\mathcal{V}_2} \min_{\mathbf{x}\in\mathcal{V}_1} \|\mathbf{x}-\mathbf{y}\|_2^2, \tag{4}$$

$$\mathcal{L}_{\mathrm{R}}(\mathcal{M}_1, \mathcal{M}_2, \mathbf{A}) = \sum_{a_i \in \mathbf{A}} \|\mathbf{I}_{\mathcal{M}_1}(a_i) - \mathbf{I}_{\mathcal{M}_2}(a_i)\|_2^2. \tag{5}$$

Explicit Constraints. Explicit constraints $\mathcal{L}_{\text{EC}}$ on the shape of controllers are proposed to predict ECs to align well with source and target respectively. It works on the shape encoder Φ_E and the controller generator Φ_G. The loss for EC is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{EC}} = \mathcal{L}_{\text{CD}}(\mathcal{C}_s, \mathcal{M}_s) + \mathcal{L}_{\text{CD}}(\mathcal{C}_t, \mathcal{M}_t) \\ + \mathcal{L}_{\text{R}}(\mathcal{C}_s, \mathcal{M}_s, \mathbf{A}) + \mathcal{L}_{\text{R}}(\mathcal{C}_t, \mathcal{M}_t, \mathbf{A}),\end{aligned} \tag{6}$$

where three angles are taken during rendering from the front, top and right.

To drive the deformation towards the target, $\mathcal{L}_{deform}$ is designed to compute the distance between the deformed human face and target animal face. The loss for deformation is computed by:

$$\begin{aligned}\mathcal{L}_{deform} = \mathcal{L}_{\text{CD}}(\mathcal{M}_{s\to t}, \mathcal{M}_t) + \mathcal{L}_{\text{R}}(\mathcal{M}_{s\to t}, \mathcal{M}_t, \mathbf{A}) \\ + \mathcal{L}_{\text{SYM}}(\mathcal{C}_{s\to t}),\end{aligned} \tag{7}$$

where $\mathcal{L}_{\text{SYM}}$ is a symmetry loss added on the controllers $\mathcal{C}_{s\to t}$ to maintain the symmetry of the source human face structure.

To avoid geometric aliasing that makes the deformation unreliable like self-intersections, flattening loss(Liu et al. 2019) and point-to-face (p2f) distance(Wang et al. 2020) are used as our smoothing loss $\mathcal{L}_{smooth}$. Unlike (Wang et al. 2020) to compute the difference of p2f values between $\mathcal{M}_S$ and $\mathcal{M}_{s\to t}$, we just penalize positive p2f values of $\mathcal{M}_{s\to t}$. $\mathcal{L}_{smooth}$ is computed as:

$$\mathcal{L}_{smooth} = \mathcal{L}_{fl}(\mathcal{C}_{s\to t}) + \mathcal{L}_{p2f}(\mathcal{M}_{s\to t}). \tag{8}$$

Moreover, to preserve the original characteristics of people as much as possible, eye-keeping loss is designed to retain the features of original human eyes, which is formulated as:

$$\mathcal{L}_{eye} = \mathcal{L}_{\text{CD}}(\mathbf{v}'_s, \mathbf{v}'_{s\to t}), \quad \mathbf{v}'_s, \mathbf{v}'_{s\to t} \in \mathcal{N}(l_e), \tag{9}$$

where $\mathbf{v}'_s$ and $\mathbf{v}'_{s\to t}$ represent the eye vertices of the source face and deformed face, l_e represents the eye landmarks of human and $\mathcal{N}_s(l_e)$ represents the neighbor vertices of l_e.

Controller-Based Mapping

We propose the controller-based mapping (CBM) to map the semantic information across species taking advantage of EC. CBM builds multi-density correspondences between the source controller and the target controller according to the importance of semantic information, *i.e.*, object regions with finer semantic information such as ears and nose are mapped with denser constraints while regions with plain structures such as cheeks and forehead receive less attention. We use the pre-defined landmarks on the human face and animal face to determine the density of semantic information on the source controller and target controller, respectively. Each landmark on the face corresponds to an area on the controller, which is regarded as rich in semantic information and thus is restricted by CBM during morphing. In this way, the human-to-animal shape morphing is conducted in a semantic-adaptive way.

There are two steps of CBM. In step one, for each landmark of the source human, CBM finds an area with a set of vertices on the source controller $\mathcal{C}_s$ which has the closest distance to the landmark. Since the source controller shares the same shape information with the source human, the vertices set on the source controller shares the same semantic information with the landmarks of the source human. Also, CBM finds such vertices set of the target controller $\mathcal{C}_t$ for each landmark of the target animal. We call these two vertices sets of controllers as the source set $\mathcal{S}_s$ and target set $\mathcal{S}_t$. In step two, after Φ_D predicts the deformation of $\mathcal{C}_s$, the corresponding deformed set $\mathcal{S}_{s\to t}$ on the deformed controller is obtained from $\mathcal{S}_s$. Then a constraint $\mathcal{L}_{\text{CBM}}$ between $\mathcal{S}_{s\to t}$ and $\mathcal{S}_t$ is designed to build semantic consistency between human and animal faces. The CBM loss is designed as:

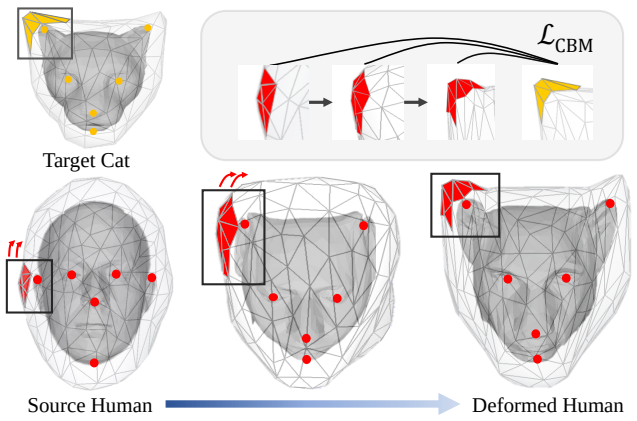

Figure 4: Controller-based mapping for human and cat faces with six landmarks. The red area on the source controller is deformed to align with the yellow area on the target controller gradually by $\mathcal{L}_{\text{CBM}}$ to realize the correspondence of the human left ear to the cat left ear.

$$\mathcal{L}_{\text{CBM}} = \mathcal{L}_{\text{CD}}(\mathcal{S}_{s\to t}, \mathcal{S}_t). \tag{10}$$

where we select six landmarks here including two ears, two eyes, nose and chin. The example of mapping the right ear from the human face to the animal face is shown in Fig. 4.

The facial landmarks are constrained across species through EC, not directly on the landmarks itself. This allows the surrounding areas of the landmarks on the source human to also correspond to those of the target mesh, so that makes the deformation result smoother and avoids human facial feature distortion.

In total, our training loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{EC}} + \mathcal{L}_{deform} + \mathcal{L}_{\text{CBM}} + \mathcal{L}_{eye} + \mathcal{L}_{smooth}, \tag{11}$$

where we have omitted a set of hyper-parameters that balance the importance of different loss terms here, and provide detailed descriptions in the supplementary materials.

Sparse-to-dense Inference Strategy

A fine-grained dense face with wrinkle details usually has at least tens of thousands of vertices, which cannot be directly fed into the network for training and testing. Down-sampling the dense face and then up-sampling the deformation result

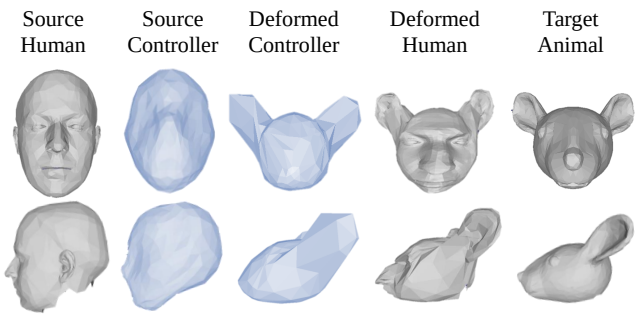

Figure 5: Explicit controllers and morphing results for human faces and animal faces from different direction. The controllers are well-structured to realize fine deformation.

directly causes the loss of face details. In order to achieve the deformation of the dense human face while preserving details, we propose a sparse-to-dense strategy to use the down-sampled sparse face to control the original dense face. The dense source human face mesh $\mathcal{M}$ is first down-sampled to a sparse mesh $\mathcal{M}_s$ by leveraging mesh down-sampling (Ranjan et al. 2018). Then $\mathcal{M}$ and $\mathcal{M}_s$ are bound by the weight computed following Eq. 2. After the network outputs the deformed mesh $\mathcal{M}_{s\to t}$ for $\mathcal{M}_s$, the deformed result of source $\mathcal{M}$ is computed following Eq. 1. Under this strategy, our model can handle fine-grained dense human face meshes with any number of vertices and still retain feature details.

Experiments

Datasets and Implementation Details

COMA (Ranjan et al. 2018) is a 3D dataset which contains 20466 head meshes of 12 different subjects. **The Headspace Dataset** (Dai et al. 2020) contains 3D human faces of 1519 people with detailed BMP textures. The training data of our model are human-animal face mesh pairs. For human faces, our training set including 10 subjects is randomly selected from 12 subjects in COMA (totally 17410 samples), and the test set contains 2 subjects by merging the others (2 subjects with 1000 samples) in COMA and extra 100 subject-dependent samples in Headspace. For animal faces, we choose one cat mesh, one mouse mesh and one monkey mesh. More details are in the supplementary materials.

Implementation Details. The network layers of Φ_E, Φ_G, and Φ_D are the same as encoders and the decoders used in Neural Cages (Wang et al. 2020) which are the simplified versions of AtlasNet (Groueix et al. 2018).

Evaluation Metrics

We evaluate the cross-species 3D face morphing task from two aspects. That is **Shape Morphing Error (SME)** between the source and target, and **Feature Preserving Error (FPE)** of human faces, which are two conflicting goals.

SME is defined as the sum of two metrics (*i.e.*, species classification error and shape alignment error). The species classification error is measured by an animal classifier based on ResNet-18 (He et al. 2016), which achieves a 95.6% average accuracy trained with over 10,000 images of three animals. The other shape alignment error is the Chamfer distance between the source and the target.

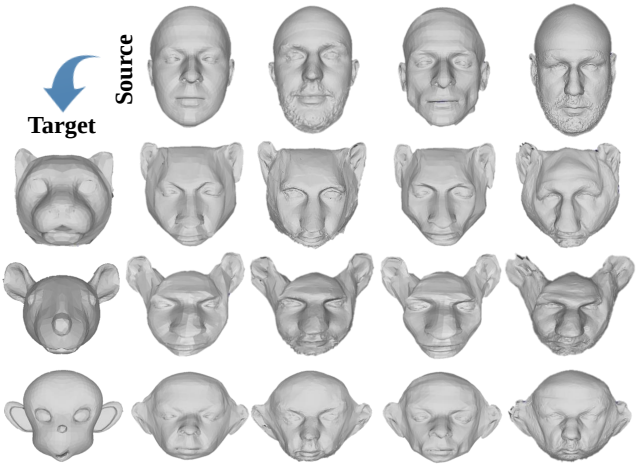

Figure 6: Morphing results of different people to different animals. The identity features of people are well preserved.

FPE is designed to measure the identity difference between source and deformed human using identity loss (Gecer et al. 2019). We render the original face and deformed face to 2D plane with texture, then we use pre-trained face recognition model from ArcFace (Deng et al. 2019) to get the embedding vector and compute the cosine distance. Here we use the 100 textured faces with 18000 vertices for FPE. Note that these metrics do not favor any method, since the Chamfer distance is optimized in all methods and the other metrics are optimized in none methods.

Morphing Results

We conduct a series of qualitative experiments and illustrate the high quality of our deformation results.

Shape Alignment using EC. As shown in Fig. 5, we visualize the controllers and deformed results of morphing a coarse human face into a mouse face. We can see that the deformation of controllers is corresponding to the deformation of the human face in an intuitive manner, which is prone to add mapping constraints and provides more stable platform for sophisticated or meticulous deformation.

Feature Preservation. The deformed results of four people to three animals in Fig. 6 show that AFM is able to preserve the characteristics of different sources, such as the identity, the boniness of cheeks and chin, the jawbone shape and so on. In order to demonstrate the practicability and versatility of our method in high-quality real textured human faces, we produce morphing results with texture rendering of four human faces by mapping the texture of source human to deformed human directly, as shown in Fig. 7. Moreover, the dynamic process of morphing a human face to a cat face is shown in Fig. 8 to demonstrate the wide applicability.

Sparse-to-dense Results. Fig. 9 exhibits the morphing results of two fine-grained human faces with over 18,000 ver-

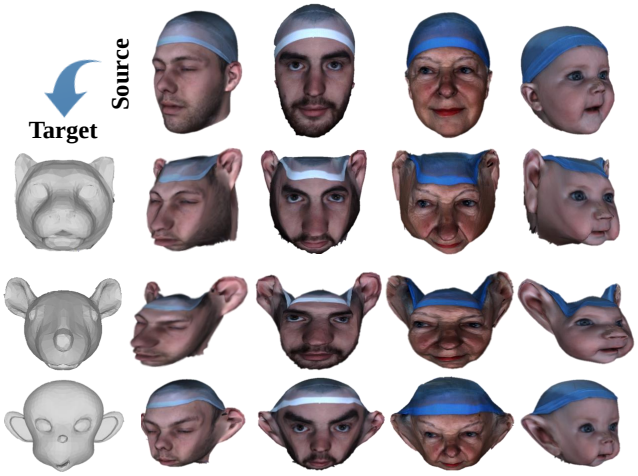

Figure 7: High-quality vivid morphing results produced for fine-grained human faces with texture by AFM.

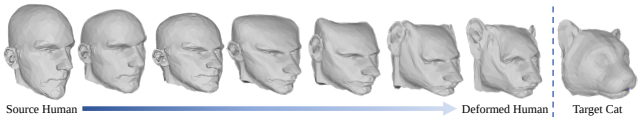

Figure 8: Dynamic process of human-to-cat morphing.

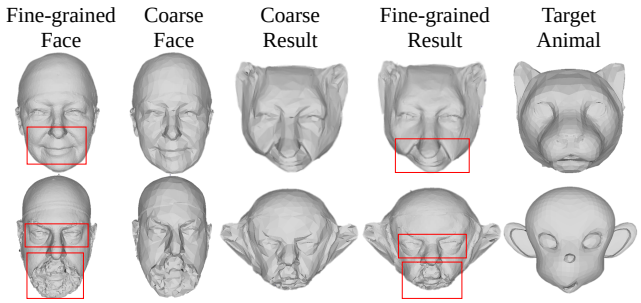

Figure 9: Morphing results for fine-grained human faces with 18,000 vertices by sparse-to-dense strategy. Our AFM well preserves the shape features such as the nasolabial fold of the woman, and the beard and eye bags of the man.

tices by sparse-to-dense strategy. It can be easily seen that AFM is able to process fine-grained meshes, where the number of vertices is far larger than that of training data.

Analysis on similar deformation methods. We compare our method with other existing similar methods, *e.g.*, optimization-based method (non-rigid ICP (Amberg, Romdhani, and Vetter 2007)), regression-based method (3DN (Wang et al. 2019b), and KeypointDeformer (Jakab et al. 2021)). For non-rigid ICP, the same six landmarks as AFM are used. For the regression-based methods, we retrain them on our datasets with off-the-shelf codes and their provided training protocols. Fig. 10 shows the qualitative results, and Fig. 11 and Tab. 1 exhibit the quantitative results. Nonrigid ICP creates artifacts and yields feature distortion. 3DN only focuses on the shape alignment, and completely corrupts the human facial features. KeypointDeformer causes semantic mismatch (right ear) as some keypoints of their predictions are wrong. We experimentally find that our method outperforms them on face morphing.

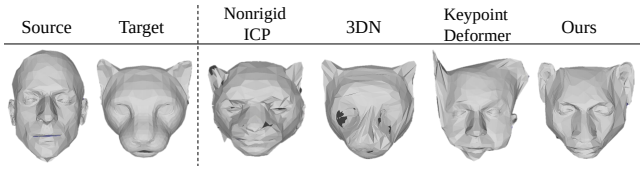

Figure 10: Comparison of our method with nonrigid ICP (Amberg, Romdhani, and Vetter 2007) 3DN (Wang et al. 2019b) and KeypointDeformer (Jakab et al. 2021). Our method outperforms them on face morphing.

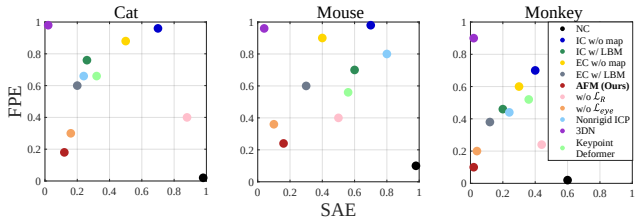

Figure 11: Quantitative comparison in terms of Shape Alignment Error (SAE) and Feature Preserving Error (FPE). The point nearest to the bottom left is the best.

Moreover, extended experiments on other tasks (*i.e.*, human bodies, hand, person-to-person, etc.) shows the generalization and wide applicability of AFM. Please refer to supplementary materials for more results.

Ablation Studies

In this section, we develop two experimental methods termed Implicit Controller (IC) and Landmark-based Mapping (LBM) to demonstrate the effectiveness of AFM.

Baseline Using Neural Cages. Neural Cages (Wang et al. 2020) (NC) performs direct deformation on objects and preserves surface details well. We regard the human facial features as surface details and regard animal faces as target shapes. We set the vertex number of NC's cage to 322 (same with us) and use the other default training setting of NC to train on our human-animal face pairs as our baseline.

IC and LBM. The difference between IC and EC is that there are no explicit constraints for IC when training, *i.e.*, $\mathcal{L}_{\text{EC}} = 0$. LBM is also a mapping method to maintain semantic consistency, while it directly computes the $\mathcal{L}_2$ loss between landmarks on the deformed and target faces instead of controllers compared with CBM. The loss of LBM is formulated as:

$$\mathcal{L}_{\text{LBM}} = \mathcal{L}_2(\mathcal{M}_{s\to t}[l_s], \mathcal{M}_t[l_t]), \tag{12}$$

where l_s and l_t represent the landmark indexes of source and target. It is worth noting that the basis for CBM is EC, thus the only choice for mapping of IC is LBM.

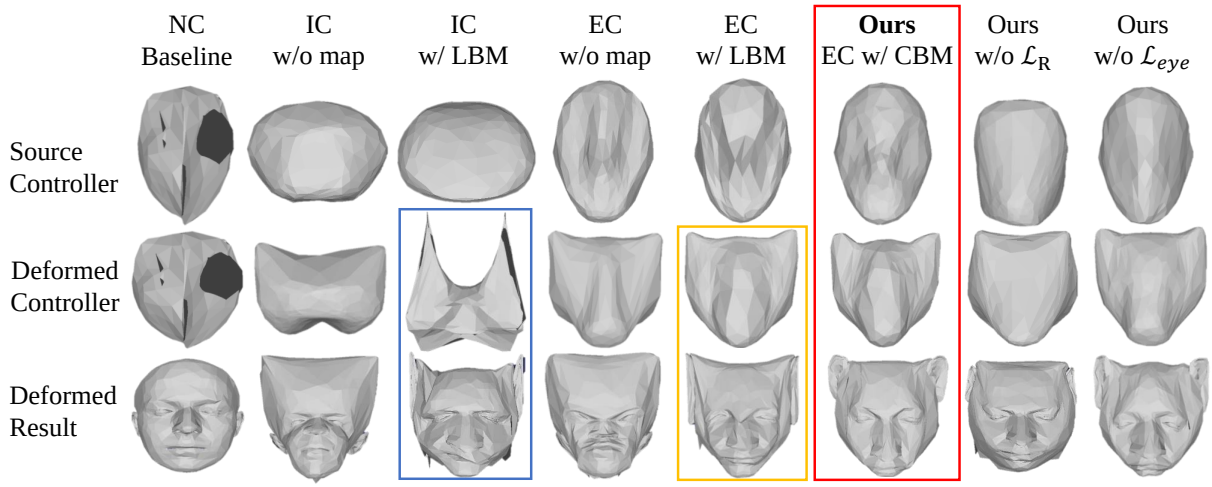

Figure 12: Qualitative comparison of ablation studies. The results demonstrate the necessity and effectiveness of EC and CBM.

	EC	IC	CBM	LBM	MS	HRS	Score
Nonrigid ICP	/	/	/	/	6.8	3.8	5.30
3DN	/	/	/	/	9.1	0.1	4.60
KeypointDeformer	/	/	/	/	5.8	4.6	5.20
NC	✗	✓	✗	✗	1.1	8.5	4.80
IC w/o map	✗	✓	✗	✗	1.9	1.6	1.75
IC w/ LBM	✓	✗	✗	✓	6.2	3.4	4.80
EC w/o map	✓	✗	✗	✗	2.6	2	2.30
EC w/ LBM	✓	✗	✗	✓	7.1	4.6	5.85
AFM (Ours)	✓	✗	✓	✗	8.4	8.1	**8.25**

Table 1: Comparison on techniques and user studies scores (mean) in terms of Morphing Score (MS) and Human Recognition Score (HRS).

Ablation studies settings. To illustrate the effectiveness of EC and CBM, we design four settings on human-to-cat morphing: 1) IC w/o mapping (*i.e.*, w/o CBM and w/o LBM), 2) IC w/ LBM, 3) EC w/o mapping (*i.e.*, w/o CBM and w/o LBM), 4) EC w/ LBM. Notice that the comparison of their techniques is shown in Tab. 1. For losses, we let $\mathcal{L}_{\mathrm{R}} = 0$ and $\mathcal{L}_{eye} = 0$ to illustrate their effectiveness.

Fig. 11 shows the quantitative results of the above methods using SAE and FPE. Each 2D point in the figure represents one method, and the coordinates represent the value of two metrics, where the origin is the ideal result. This figure confirms that our AFM achieves the balance of structure change and human feature preservation better. Fig. 12 shows the qualitative results. Compared with the baseline, our method realizes the semantic-adaptive fine deformation of the human face. Compared with IC w/o mapping and EC w/o mapping, the results show that the human facial features undergo severe distortion without mapping. The results for losses show that $\mathcal{L}_{\mathrm{R}}$ greatly enhances the effect of alignment, and $\mathcal{L}_{eye}$ strengthens the preservation of human eyes.

Effectiveness of EC and CBM. As illustrated in Fig. 12 (blue box), the controller of IC w/ LBM is irregular and the deformed face is full of spiky artifacts. As shown in the yellow box, EC w/ LBM corrupts face features and creates self-intersections and non-physical distortions. Especially for the ears, the vertices of the face which do not belong to the ears are stretched together. It indicates that EC and CBM help preserve human facial features better during mapping.

User study. We also compare the various methods via a human subjective evaluation for the 9 methods listed in Tab. 1. Specifically, we introduce two types of user studies, **Morphing Score (MS)** and **Human Recognition Score (HRS)**. For MS, we randomly use 50 groups of meshes, each consists of the source, the target, and the results of all methods. For each group, the participants are asked to score the results by the degree of the morphing as MS, from 0 to 10 (0 is the worst, 10 is the best). For HRS, we randomly select deformed results of 50 different people for each method respectively, and show participants a deformed face and 10 source faces (including one real source and nine randomly selected different source) each time. The participants are then asked to choose the right source. We multiply the correct rate of the participants' choices by 10 as HRS. We conduct the user study with over 100 participants and calculate the mean value of two scores of each method as our user study scores. The results listed in Tab. 1 show that our method outperforms the others in the preference of participants.

The above experiments demonstrate the necessity and effectiveness of AFM on cross-species 3D face morphing.

Conclusion

In this paper, we propose an alignment-aware 3D face morphing framework called AFM for a new potential task: cross-species 3D face morphing. Explicit Controller and Controller-based Mapping are proposed to build the semantic-adaptive correspondences between human and animal faces which helps preserve human features better. Comprehensive experimental results demonstrate the effectiveness and exquisite performance of our method.

Acknowledgments

This work was supported by National Science Foundation of China (U20B2072, 61976137).

References

Amberg, B.; Romdhani, S.; and Vetter, T. 2007. Optimal Step Nonrigid ICP Algorithms for Surface Registration. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Bai, Z.; Cui, Z.; Liu, X.; and Tan, P. 2021. Riggable 3D Face Reconstruction via In-Network Optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6216–6225.

Bailey, S. W.; Omens, D.; Dilorenzo, P.; and O'Brien, J. F. 2020. Fast and deep facial deformations. *ACM Transactions on Graphics (TOG)*, 39(4): 94–1.

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187–194.

Calderon, S.; and Boubekeur, T. 2017. Bounding proxies for shape approximation. *ACM Transactions on Graphics (TOG)*, 36(4): 1–13.

Chandran, P.; Bradley, D.; Gross, M.; and Beeler, T. 2020. Semantic Deep Face Models. In *2020 International Conference on 3D Vision (3DV)*, 345–354.

Chaudhuri, B.; Vesdapunt, N.; Shapiro, L.; and Wang, B. 2020. Personalized Face Modeling for Improved Face Reconstruction and Motion Retargeting. In *European Conference on Computer Vision*, 142–160. Springer.

Chaudhuri, B.; Vesdapunt, N.; and Wang, B. 2019. Joint face detection and facial motion retargeting for multiple faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9719–9728.

Chen, L.; Huang, J.; Sun, H.; and Bao, H. 2010. Cage-based deformation transfer. *Computers & Graphics*, 34(2): 107–118.

Chen, Z.; and Kim, T.-K. 2021. Learning Feature Aggregation for Deep 3D Morphable Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 13164–13173.

Clay, V.; König, P.; and König, S. U. 2019. Eye tracking in virtual reality. *Journal of Eye Movement Research*, 12(1).

Dai, H.; Pears, N.; Smith, W.; and Duncan, C. 2020. Statistical Modeling of Craniofacial Shape and Texture. *International Journal of Computer Vision*, 128(2): 547–571.

Deng, J.; Guo, J.; Niannan, X.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Gao, Z.; Zhang, J.; Guo, Y.; Ma, C.; Zhai, G.; and Yang, X. 2020. Semi-Supervised 3D Face Representation Learning From Unconstrained Photo Collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Gecer, B.; Ploumpis, S.; Kotsia, I.; and Zafeiriou, S. 2019. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1155–1164.

Geng, Z.; Cao, C.; and Tulyakov, S. 2019. 3d guided fine-grained face manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9821–9830.

Groueix, T.; Fisher, M.; Kim, V. G.; Russell, B. C.; and Aubry, M. 2018. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 216–224.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Ho, L.-H.; Sun, H.; and Tsai, T.-H. 2019. Research on 3D Painting in Virtual Reality to Improve Students' Motivation of 3D Animation Learning. *Sustainability*, 11(6): 1605.

Hoppe, H. 1997. View-dependent refinement of progressive meshes. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 189–198.

Huang, J.; Shi, X.; Liu, X.; Zhou, K.; Wei, L.-Y.; Teng, S.-H.; Bao, H.; Guo, B.; and Shum, H.-Y. 2006. Subspace gradient domain mesh deformation. In *ACM SIGGRAPH 2006 Papers*, 1126–1134.

Jakab, T.; Tucker, R.; Makadia, A.; Wu, J.; Snavely, N.; and Kanazawa, A. 2021. KeypointDeformer: Unsupervised 3D Keypoint Discovery for Shape Control. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12783–12792.

Jiang, Z.-H.; Wu, Q.; Chen, K.; and Zhang, J. 2019. Disentangled Representation Learning for 3D Face Shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Joshi, P.; Meyer, M.; DeRose, T.; Green, B.; and Sanocki, T. 2007. Harmonic coordinates for character articulation. *ACM Transactions on Graphics (TOG)*, 26(3): 71–es.

Ju, T.; Schaefer, S.; and Warren, J. 2005. Mean value coordinates for closed triangular meshes. In *ACM Siggraph 2005 Papers*, 561–566.

Ju, T.; Zhou, Q.-Y.; van de Panne, M.; Cohen-Or, D.; and Neumann, U. 2008. Reusable skinning templates using cage-based deformations. *ACM Transactions on Graphics (TOG)*, 27(5): 1–10.

Le, B. H.; and Deng, Z. 2017. Interactive Cage Generation for Mesh Deformation. In *Proceedings of the 21st ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D '17. New York, NY, USA: Association for Computing Machinery. ISBN 9781450348867.

Lee, G.-H.; and Lee, S.-W. 2020a. Uncertainty-Aware Mesh Decoder for High Fidelity 3D Face Reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6100–6109.

Lee, G.-H.; and Lee, S.-W. 2020b. Uncertainty-Aware Mesh Decoder for High Fidelity 3D Face Reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Li, R.; Bladin, K.; Zhao, Y.; Chinara, C.; Ingraham, O.; Xiang, P.; Ren, X.; Prasad, P.; Kishore, B.; Xing, J.; and Li, H. 2020. Learning Formation of Physically-Based Face Attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, S.; Li, T.; Chen, W.; and Li, H. 2019. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, 7708–7717.

Ouzounis, C.; Kilias, A.; and Mousas, C. 2017. Kernel projection of latent structures regression for facial animation retargeting. *arXiv preprint arXiv:1707.09629*.

Pan, J.; Han, X.; Chen, W.; Tang, J.; and Jia, K. 2019. Deep mesh reconstruction from single RGB images via topology modification networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 9964–9973.

R, M. B.; Tewari, A.; Seidel, H.-P.; Elgharib, M.; and Theobalt, C. 2021. Learning Complete 3D Morphable Face Models From Images and Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3361–3371.

Ramachandran, S.; Ghafourzadeh, D.; de Lasa, M.; Popa, T.; and Paquette, E. 2018. Joint planar parameterization of segmented parts and cage deformation for dense correspondence. *Computers Graphics*, 74: 202–212.

Ranjan, A.; Bolkart, T.; Sanyal, S.; and Black, M. J. 2018. Generating 3D faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 704–720.

Ribera, R. B. i.; Zell, E.; Lewis, J. P.; Noh, J.; and Botsch, M. 2017. Facial Retargeting with Automatic Range of Motion Alignment. 36(4).

Sacht, L.; Vouga, E.; and Jacobson, A. 2015. Nested cages. *ACM Transactions on Graphics (TOG)*, 34(6): 1–14.

Sander, P. V.; Gu, X.; Gortler, S. J.; Hoppe, H.; and Snyder, J. 2000. Silhouette clipping. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 327–334.

Sederberg, T. W.; and Parry, S. R. 1986. Free-form deformation of solid geometric models. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, 151–160.

Sidhu, V.; Tretschk, E.; Golyanik, V.; Agudo, A.; and Theobalt, C. 2020. Neural Dense Non-Rigid Structure from Motion with Latent Space Constraints. In *European Conference on Computer Vision*, 204–222. Springer.

Smith, W. A. P.; Seck, A.; Dee, H.; Tiddeman, B.; Tenenbaum, J. B.; and Egger, B. 2020. A Morphable Face Albedo Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Sumner, R. W.; and Popović, J. 2004. Deformation transfer for triangle meshes. *ACM Transactions on graphics (TOG)*, 23(3): 399–405.

Sung, M.; Jiang, Z.; Achlioptas, P.; Mitra, N. J.; and Guibas, L. J. 2020. DeformSyncNet: Deformation Transfer via Synchronized Shape Deformation Spaces. *arXiv preprint arXiv:2009.01456*.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Niessner, M. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Tretschk, E.; Tewari, A.; Zollhöfer, M.; Golyanik, V.; and Theobalt, C. 2020. DEMEA: Deep Mesh Autoencoders for Non-rigidly Deforming Objects. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 601–617. Cham: Springer International Publishing. ISBN 978-3-030-58548-8.

Wang, W.; Ceylan, D.; Mech, R.; and Neumann, U. 2019a. 3DN: 3D Deformation Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, W.; Ceylan, D.; Mech, R.; and Neumann, U. 2019b. 3DN: 3D Deformation Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, Y.; Aigerman, N.; Kim, V. G.; Chaudhuri, S.; and Sorkine-Hornung, O. 2020. Neural Cages for Detail-Preserving 3D Deformations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Xia, J. C.; and Varshney, A. 1996. Dynamic view-dependent simplification for polygonal models. In *Proceedings of Seventh Annual IEEE Visualization'96*, 327–334. IEEE.

Yao, G.; Yuan, Y.; Shao, T.; and Zhou, K. 2020. Mesh Guided One-Shot Face Reenactment Using Graph Convolutional Networks. MM '20, 1773–1781. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Zhang, J.; Chen, K.; and Zheng, J. 2020. Facial Expression Retargeting from Human to Avatar Made Easy. *IEEE Transactions on Visualization and Computer Graphics*.

Zhang, X.; Wang, J.; Liu, Y.; Lu, G.; and Zhang, X. 2020. Kinetic Sculpture Design Using the Dynamic Cage. In *Journal of Physics: Conference Series*, volume 1519, 012013.

Zhou, Y.; Deng, J.; Kotsia, I.; and Zafeiriou, S. 2019. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1097–1106.