RaBit: Para*m*etric Modeling of 3D *Bi*ped Car*t*oon Characters with a Topological-consistent Dataset

Zhongjin Luo[1*] Shengcai Cai[1,2*] Jinguo Dong[1] Ruibo Ming[3]
Liangdong Qiu[1] Xiaohang Zhan[2] Xiaoguang Han[1#]

[1]The Chinese University of Hong Kong, Shenzhen [2]Huawei Technologies [3]Tsinghua University

Figure 1. We present ***3DBiCar***, the first large-scale repository of 3D biped cartoon characters. It contains 1,500 topologically consistent, textured, and skinned 3D high-quality meshes manually created by professional artists, covering 15 species. Further, we propose *RaBit*, the first cartoon character parametric model simultaneously parameterizing shape, pose, and texture.

Abstract

Assisting people in efficiently producing visually plausible 3D characters has always been a fundamental research topic in computer vision and computer graphics. Recent learning-based approaches have achieved unprecedented accuracy and efficiency in the area of 3D real human digitization. However, none of the prior works focus on modeling 3D biped cartoon characters, which are also in great demand in gaming and filming. In this paper, we introduce 3DBiCar, the first large-scale dataset of 3D biped cartoon characters, and RaBit, the corresponding parametric model. Our dataset contains 1,500 topologically consistent high-quality 3D textured models which are manually crafted by professional artists. Built upon the data, RaBit is thus designed with a SMPL-like linear blend shape model and a StyleGAN-based neural UV-texture generator, simultaneously expressing the shape, pose, and texture. To demonstrate the practicality of 3DBiCar and RaBit, various applications are conducted, including single-view reconstruction, sketch-based modeling, and 3D cartoon animation. For the single-view reconstruction setting, we find a straightforward global mapping from input images to the output UV-based texture maps tends to lose detailed appearances of some local parts (e.g., nose, ears). Thus, a novel part-sensitive texture reasoner is designed to make all important local areas perceived. Experiments further demonstrate the effectiveness of such a novel design, both qualitatively and quantitatively. We will release both 3DBiCar and RaBit to the research community.

1. Introduction

With the rapid development of digitization, creating high-quality 3D articulated characters is highly demanded in game platforms, film industries, and metaverse scenarios. However, even for expert artists, creating a 3D character is labor-intensive and time-consuming. Therefore, reducing the cost of producing visually plausible 3D characters is essential in the field of computer vision and graphics.

Recently, researchers have made great progress in digitizing realistic human characters. The emergence and popular-

ity of various 3D sensing devices make capturing 3D data from the real world convenient, prompting a growing number of 3D real-people scanned datasets [3,7,12,42,44,51,52,54]. Based on these large-scale datasets, several powerful parametric models [5,12,34,38] have been developed to facilitate the reconstruction and analysis of human shapes, actions, and interactions. With the help of parametric models, deep learning techniques have shown the potential to efficiently infer accurate 3D digital humans from single-view images [26,38] or even sparse sketches [9,24]. Most recently, there are some works [35,40] that devote to exploring the intelligent generation of cartoon-like character heads. However, none of the prior works focuses on the modeling of 3D full-body biped cartoon characters, which are also in great demand in the area of gaming (e.g., Animal Crossing), filming (e.g., Zootopia), and virtualizing (e.g., Metaverse). In this work, we raise a new problem to the community: *How to quickly produce 3D biped cartoon characters from easy-to-obtain inputs (e.g., a single image)?*

Revisiting the road map of realistic human digitization, the first step to tackling the above problem is building a high-quality 3D biped cartoon characters dataset. We thus introduce *3DBiCar*, the first large-scale publicly available 3D biped cartoon character dataset following three criteria: *1)* ***Diversity.*** *3DBiCar* spans a wide range of 3D biped cartoon characters, containing 1,500 high-quality 3D models covering 15 species, as shown in the Fig. 3. *2)* ***Richness.*** Each model in *3DBiCar* owns not only a detailed shape but also a texture UV-map, which are matched with a reference image. Additionally, each character is attached with two models, one with T-pose and another with the reference pose. *3)* ***Topological-consistency.*** More importantly, each 3D model is created by carefully deforming a pre-defined template mesh. Thus all 3D characters in *3DBiCar* are unified in topology, paving the way to learn a skinned parametric model. Fig. 1 shows some representative models of the proposed dataset.

Based on *3DBiCar*, we further propose a generative model, dubbed *RaBit*, for 3D biped cartoon character generation. It combines a linear blend shape model with a neural texture generator and simultaneously parameterizes the shape, pose, and texture to a low-dimensional parametric space. For shape and pose modeling, numerous methods have shown principal component analysis's (PCA's) advantage in building decent statistical shape models [5, 12, 31, 38]. Inspired by SMPL [34], we utilize the traditional PCA technique to parameterize shape. Due to the variety and complexity of cartoon texture, directly adopting PCA for texture modeling fails to reconstruct details and falls into blurry results. In *RaBit*, we tackle this problem by introducing a StyleGAN-based generator.

To explore the practical usage of *3DBiCar* and *RaBit*, we first conduct the application of single-view reconstruction. Considering prior works for SMPL-based human geometry generation from single-view images [6, 26, 30], we build a baseline method with our dataset and the parametric model. We select one regression-based method for pose and shape inference. For texture inference, we find directly applying a global texture-generator tends to make the results lose detailed appearances, especially for some local but important regions (e.g., nose and ears). Thus, a novel part-sensitive reasoner is proposed, which utilizes separate encoder-decoder architectures to deal with different local regions. We term our baseline method for single-view reconstruction as *BiCarNet*. Moreover, two further applications, i.e., sketch-based modeling and 3D character animation, are also explored. Experimental results on these applications demonstrate that it is already able to result in reasonable outputs. We hope that our work opens a door for researchers to explore bipedal character digitization with the proposed dataset.

To summarize, our contributions include:

- We introduce *3DBiCar*, the first large-scale 3D biped cartoon character dataset. It contains 1,500 high-quality textured 3D models with a consistent mesh topology.
- We propose the first 3D full-body cartoon parametric model *RaBit* for biped character modeling. We will release both *3DBiCar* and *RaBit* for future research.
- We carefully designed *BiCarNet*, the first method to reconstruct 3D textured biped cartoon characters from a single-view image. A novel part-sensitive reasoner is invented for detailed texture generation.
- Two other applications, i.e., sketch-based modeling and 3D character animation, are also successfully conducted. We strongly believe our work opens a door for future research.

2. Related Work

3D Character Datasets. In general, 3D character datasets could be categorized as real-captured and computer-designed datasets. For capturing character data from the real world, the availability of 3D scanning devices has enabled researchers to collect many scanned 3D human-related datasets, mainly focusing on human faces [10] and bodies [13]. For human faces, FaceWarehouse [12] collects large-scale 3d faces with high diversity in age, ethnicity, and expression. FaceScape [51] further builds a large-scale detailed 3D face dataset with high resolution in texture and mesh. For human bodies, CAESAR dataset [42] opens up the learning of the human body and is widely used for body shape modeling for its shape diversity and satisfying resolution of meshes. Many following works [3, 7, 44, 52, 54] extend [42] in shape, pose, and texture, on quantity and quality. Although these real-captured datasets are widely used in realistic human digitalization, they are unsuitable for imaginary character generation.

For designing character data with computers, researchers try to perform deformation on real 3D human faces or bodies to construct exaggerated shapes programmatically. [11, 24, 45, 49] Still, their results lack diversity and are far from satisfactory. To address this issue, 3DCaricShop [40] proposes a large-scale 3D exaggerated faces dataset. SimpModeling [35] constructs a large man-made animalmorphic head dataset. Although using 3DCaricShop and SimpModeling could facilitate the generation of unreal character heads, it still remains a problem to synthesize full-body cartoon characters due to the lack of corresponding body data. In our work, we tackle this problem by introducing a large 3D biped cartoon character dataset, *3DBiCar*. It contains 1,500 high-quality 3D full-body textured models and spans a wide range of biped cartoon characters.

Parametric Shape Modeling. Parametric models of shapes are widely used in 3D digitizations. Blanz *et al.* [5] pioneer parametric modeling using PCA and release a 3D statistical morphable face model (3DMM). Their parameterization models textured faces and provides a set of controls for intuitively manipulating shapes, expressions, and textures. Since then, PCA-based parameterizing has gradually dominated the area of statistical shape modeling over the past decades. Following 3DMM, researchers model the whole head to represent the neck region and 3D head rotations. [12, 31] Allen *et al.* [2] open up the study of full body parameterization. However, they focus only on body shape and omit the body pose. SCAPE [3] represents body shape and pose in terms of triangle deformations, while SMPL [34] models a whole range of natural shapes and poses based on vertex displacements. SMPL-X [38] integrates SMPL [34] with FLAME [31] head model and the MANO [43] hand model for expressive capturing of bodies, hands and faces together. With recent advances in deep learning, researchers turn to explore nonlinear shape models using neural networks [1, 4, 8, 41, 50, 55]. However, since these non-linear modeling methods are inferior in simplicity, robustness and availability, PCA-based methods remain prevalent in the research community. In this paper, we also adopt PCA into *RaBit* to model the geometry of 3D biped cartoon characters.

Based on the above parametric models, researchers have made remarkable progress in human digitization, such as reconstruction from simple inputs (e.g., a single image or sparse sketches) [6, 15, 24, 26, 39] and real-time pose retargeting [14, 29, 48]. For instance, SMPLify [6] estimates 3D body shape and pose parameters automatically from 2D joints with multiple ellipsoids. HMR [26] proposes an end-to-end framework for reconstructing a full 3D mesh of a human body from a single RGB image. DeepSketch2Face [24] proposes a sketch-based system for inferring 3D face models from 2D sketches with the help of parametric models and CNN-based deep regression networks. TCMR [14] presents a temporally consistent mesh recovery system for recovering smooth 3D human motion from monocular videos. To probe the capability of *RaBit* to downstream applications, we conduct various utilization, such as single-view cartoon character reconstruction, sketch-based character modeling, and 3D cartoon animation. Experimental results demonstrate the practicality of *3DBiCar* and *RaBit*.

Parametric Texture Modeling. Traditionally, textures are modeled as a linear subspace using the similar idea of body blendshape models. Blanz *et al*. [5] represent the face appearance in per-vertex colors and parameterize texture as a linear model based on PCA. Dai *et al.* [16] store texture information in a UV space where the texture resolution is unconstrained by mesh resolution. Moschoglou *et al.* [36] formulate a robust matrix factorization problem to learn the parametric representation of facial UV maps from a collection of training textures. However, these linear texture models may lead to a sub-optimal appearance output [18, 23] due to the weak assumption of Gaussian and tend to produce blurry results.

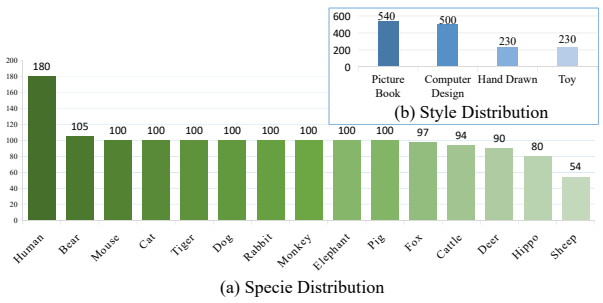
With recent advances in deep learning, researchers turn to utilize deep neural networks to model texture. A number of deep generative models [17, 19–22, 37, 46] have been proposed to parameterize texture into a latent space. For example, GANFIT [21] utilizes GAN-based neural networks to train a generator of facial texture in UV space for 3D face reconstruction. StylePeople [22] incorporates neural texture synthesis, mesh rendering, and neural rendering into the joint generation process to train a neural texture generator for the task of single-view human reconstruction. GET3D [20] introduces a texture-field generative model that directly generates explicit textured 3D meshes, ranging from cars, chairs, animals, motorbikes, and human characters to buildings. These methods have shown the promising capacity of neural generators to represent texture. In our work, we adopt a GAN-based neural texture generator into *RaBit* to provide high-quality texture modeling. Furthermore, in the task of single-view reconstruction, we design a novel part-sensitive texture reasoner to make all important local appearances perceived.

3. Dataset

Considerable progress has been made in digitizing realistic and articulated human characters. However, efficiently creating visually plausible biped cartoon characters remains demanding and challenging, mainly due to the lack of data. In this work, we propose to fill this gap by introducing *3DBiCar*, the first large-scale full-body 3D biped character data. We build *3DBiCar* following three rules:

Diversity. *3DBiCar* spans a wide range of 3D biped cartoon characters, containing 1,500 high-quality 3D models. First, we carefully collect images of 2D full-body biped cartoon characters with diverse identities, shape, and textural styles from the Internet, resulting in 15 character species and 4 image styles, as shown in Fig. 3. Then we recruit six pro-

Figure 2. The **gallery** of the representative examples sampled from *3DBiCar*. Each collected reference image is followed by the T-pose model and the posed model, created by professional artists. *3DBiCar* contains 1,500 topologically consistent, textured and skinned 3D high-quality models with paired 2D images, which covers 15 species and 4 image styles.



(a) Specie Distribution

Figure 3. **Data distribution.** Chart (a) illustrates the number of 15 species of bipedal cartoon characters in *3DBiCar*. Chart (b) shows the number of four styles of reference images collected in our dataset.

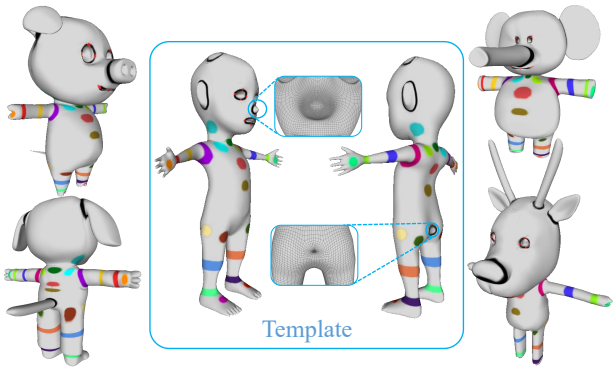


Figure 4. **Template.** The models in the center are the predefined template mesh with landmarks. It can bee seen that we refine the structure on specific regions, where a complex nose or tail may exist. The colored regions and delineated lines denote the landmarks. These landmarks represent specific components of the character's body, such as elbow and eye socket. During model crafting, artists are required to deform the template model while keeping the landmarks in the position where the original body components are.

fessional artists to create 3D corresponding character models according to the collected reference images. The modeling result is required to be matched with the reference images as much as possible. The representative image-model pairs sampled from our dataset are shown in Fig. 2.

Topological-consistency. The key to building a linear parametric shape model is keeping a unified mesh topology. Traditional human parametric models utilize a template mesh to register different human body scans with 3D landmarks to keep topologically uniform. Inspired by this, we first create a template mesh with several 3D colored landmarks as shown in Fig. 4. All six artists are required to craft 3D models by deforming the above-predefined template under the constraints of these obvious landmarks. We set up a review committee of 10 to check these models based on the landmarks, ensuring the consistency of mesh topology. The landmarks could also

be used to compute the position of models' joints for body posing or character animation. The topological consistency of *3DBiCar* paves the way to learn a skinned parametric model, which we will discuss in Section 4.

Richness. We provide complete and various forms of data for each character. There are not only the 3D shape meshes and UV-space textures carefully crafted by artists but also collected reference images. For each character, artists are asked first to create a T-pose mesh and then deform it to match the reference pose. Furthermore, all the models are rigged and skinned by the predefined skeleton and skinning weight matrix, which supports further animation production of characters. Note that eyeball meshes are extra modeled to support the facial expression in the future better. Each body mesh comprises 38,726 vertices and 77,448 faces, while each eyeball contains 1,025 vertices and 2,046 faces.

4. Parametric Modeling

We propose the first pa**ra**metric model of 3D **bi**ped car**t**oon characters (*RaBit*), which contains a linear blend model for shapes and a neural generator for textures. *RaBit* simultaneously parameterizes the shape, pose, and texture of the 3D biped character. Specifically, we decompose the parametric space into identity-related body parameter B (Sec. 4.1), non-rigid pose-related parameter Θ (Sec. 4.2) and texture-related parameter T (Sec. 4.3). Overall, a 3D biped character is parameterized as follows:

$$\begin{aligned} M =& F(B, \Theta, T) \\ =& F_T(F_P(F_S(B), \Theta), T), \end{aligned} \tag{1}$$

where F_S, F_P, and F_T are the parametric functions to generate shape, pose, and texture respectively. The following sections will elaborate the details in our parameterization.

4.1. Shape Modeling

Recently, linear shape models dominate the representation of statistical 3D model. Numerous methods [5, 12, 31, 38] have shown PCA's ability in modeling of the human body and face. Inspired by [34], we parameterize our character shape linearly with the following equation,

$$M_S = F_S(B) = \bar{M}_S + \sum_i^{|B|} \beta_i s_i, \tag{2}$$

where $\bar{M}_S$ denotes the mean shape and M_S is the reconstructed shape. The coefficients of linear shape are $\beta_i \in B$. $|B|$ is the number of shape parameters and is set to 100 in our implementation. $s_i \in \mathbb{R}^{3\times N}$ denotes the orthonormal principal components of vertex displacements that capture shape variations in different character identities. Note that we also construct the eyes' meshes to allow possible eyeball-driven demand in the future. The eyes' meshes are computed based on the predefined landmarks shown in Fig. 4. Please refer to the supplementary for detailed implementation of eyes.

4.2. Pose Modeling

RaBit adopts standard vertex-based linear blend skinning with a predefined skeleton and skinning weight matrix, provided by our *3DBiCar*. Our template body mesh has 38,726 vertices and there are $K = 23$ joints in the predefined skeleton. Specifically, pose parameter Θ defines a set of angles $\Theta = [\theta_1, \theta_2, ..., \theta_K] \in \mathbb{R}^{69}$ following [34]. The rotation of node k can be expressed as $\theta_k \in \mathbb{R}^3$ where θ_k can be converted to rotation matrix format $R(\theta_k)$ using Rodrigues' formula. The following equations demonstrate how the pose function F_P changes vertex $v_i \in M_S$ to its corresponding position $v_i' \in M_P$:

$$\bar{v}_i = \sum_{k=1}^{K} w_{k,i} G'_k(\Theta, J) v_i, \tag{3}$$

$$G'_k(\Theta, J) = G_k(\Theta, J) G_k(\Theta', J)^{-1}, \tag{4}$$

$$G_k(\Theta, J) = \prod_{j \in A(k)} \begin{bmatrix} R(\theta_j) & J_j \\ 0 & 1 \end{bmatrix}, \tag{5}$$

where $w_{k,i}$ is the k-th element of the linear blend matrix W for the i-th vertex. $G_k(\Theta, J)$ is the global transformation of joint k, $G'_k(\Theta, J)$ is the global transformation of joint k removing the transformation of rest pose Θ', and $A(k)$ denotes a set including all the ancestors of joint k, and J_j is the location of the j-th joint which locates at the bounding box center of specific body landmark. Thus given the T-pose mesh M_S and specific pose Θ, we can generate the corresponding posed mesh M_P by F_P.

$$M_P = F_P(M_S, \Theta) \tag{6}$$

4.3. Texture Modeling

Although traditional PCA is capable to build a decent statistical shape model, it fails to represent high-frequency details in textures and falls into producing blurry results, due to its weak Gaussian assumption. Recently, GAN-based architectures [19, 21, 22, 27, 28, 47] have shown the notable capability of generating high-fidelity images. Thus, our *RaBit* we resort to StyleGAN2-based techniques for UV texture maps generation, but with a coherent UV unfolding (as shown in Fig. 5) to facilitate the learning of texture compared with [22]. Specifically, the neural texture generator in *RaBit* translates a latent code to a texture map, which could be formulated as follows,

$$G(T) : \mathbb{R}^Z \rightarrow \mathbb{R}^{H\times W\times C}, \tag{7}$$

where $G(T)$ takes a Z-dimensional parameter vector as input and synthesizes a $H \times W \times C$ texture map. Thus given

the posed mesh M_P and a specific texture code T, we can generate the textured mesh, which could be denoted as,

$$M = F_T(M_P, T) = H(M_P, G(T)), \tag{8}$$

where H means the process of applying the texture map to mesh model. In our implementation, the generator follows the architecture of StyleGAN2 [28], while taking a 512-dimensional parameter vector as input and generating a $1024 \times 1024 \times 3$ texture map.

5. Single-View Reconstruction

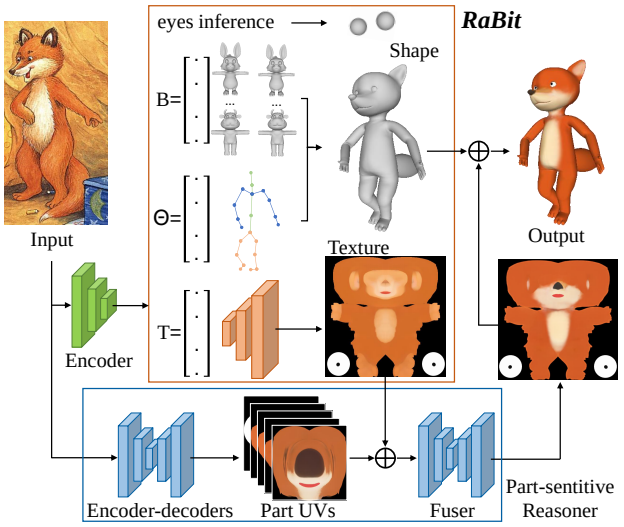


Figure 5. ***BiCarNet***. Given the input image containing a 2D cartoon character, we first map the image to the parametric vectors. The vectors are then fed to our *RaBit* to generate a posed body mesh, two eyeballs, and a global UV texture. A novel part-sensitive reasoner is further introduced to perceive local regions and generate the detailed UV texture next. Finally, a vivid 3D cartoon character is obtained with our *BiCarNet*.

Single-view reconstruction (SVR) is one of the most popular tasks of efficient 3D content generation, and recent work has made noticeable progress on human reconstruction based on parametric model of human characters (e.g., SMPL). To verify the practicality of our proposed *3DBiCar* and *RaBit*, we also conduct SVR for bipled cartoon characters. A baseline learning-based method is presented, which is termed as *BiCarNet*.

5.1. *BiCarNet*

Given a single image of cartoon characters, our *BiCarNet* aims to reconstruct the corresponding 3D shape, pose, and texture. As *Rabit* spans a large space of textured models. Then, the key problem is to build an encoder to map the input image to the parametric space. As shown in the upper part of Fig 5, We adopt the learning block in HMR [26] as our Encoder. Once these parametric vectors are learned, we can feed them to our *RaBit* model to generate a posed body mesh, two eyeballs, and a UV texture (we name it global for convenience to introduce the following method description).

During our preliminary experiments, we find that the shape reconstruction of characters, *i.e*. the eyes and body, is satisfactory, while the inferred UV tends to lose detailed appearances of some small yet significant areas, such as the nose and ears. We thus propose a novel part-sensitive texture reasoner (PSR) to address the above issues, as the lower part of Fig 5 shows. Specifically, we design five individual UV-mappings for these significant parts of the nose, ears, horns, eyes, and mouth. Five lightweight encoder-decoder branches are next introduced to learn the appearances of these local regions from the input image, respectively. The learned part UVs could be remapped to the corresponding area on the global UV map to produce a blended texture. However, a direct blending tends to cause seam artifacts. Thus we further adopt a Fuser to address the artifacts as Fig 5 illustrates. Please refer to the Supplementary for thorough implementations of *BiCarNet*.

5.2. Experiments

Data preparation. We first split *3DBiCar* into a training set (1,050 image-model pairs) and a testing set (450 pairs). To support a stable training of *BiCarNet*, we next generate a large number of synthetic paired data with the help of *RaBit*, which are highly diversified in shape, posture, and texture. To be specific, a lot of 3D textured models are first sampled from the *RaBit* space, which are then rendered into images from different camera views. This finally produce 13,650 pairs for training. Note that, *BiCarNet* takes an image with foreground masked as input. All synthetic images naturally have no background. For real images, all the foreground masks are manually annotated.

Result gallery. Fig. 6 shows representative results generated by *BiCarNet*. As illustrated, our *BiCarNet* can generate vivid 3D cartoon characters loyal to individual cartoon images in shape, pose, and texture. We believe that our work opens the door to producing 3D biped cartoon characters from easy-to-obtain inputs.

Results on Shape Reconstruction. As mentioned above, *BiCarNet* utilizes HMR-like blocks and *RaBit* for shape and pose learning. Currently, other reconstruction methods could also be used for topology-consistent geometry inference, such as GCNN-based methods [33] and UV-based methods [53]. We choose two representative methods for comparison, i.e., Mesh-Graphormer [32, 33] and DecoMR [53]. Mesh-Graphormer combines graph convolutions and self-attentions in a transformer for 3D human reconstruction from a single image. DecoMR reconstructs 3D human mesh from single images by regressing a UV-based location map. Tab. 1 shows the quantitative comparisons of the above three

Figure 6. **Result gallery of *BiCarNet*.** Our *BiCarNet* is capable of generating vivid 3D cartoon characters with only single-view image input.

methods on MPVE, MPJPE, and PA-MPJPE. We also provide qualitative comparisons in Fig. 7. Both quantitative and qualitative results demonstrate that the HMR-like method achieves the highest performance on geometry inference and provides more accurate reconstructions closer to ground truths. As noted, both Mesh-Graphormer and DecoMR outperform HMR for SMPL-based human reconstruction. It is interestingly found that they perform worse in our settings. One possible reason is that our biped cartoon meshes own an extremely larger amount of vertices than SMPL to model more complex geometry. This greatly increases the challenge of vertices regression in Mesh-Graphormer and DecoMR. Thus, in our setting, directly regressing the low-dimension parameters performs better.

Method	MPVE ↓	MPJPE ↓	PA-MPJPE ↓
DecoMR [53]	85.74	81.23	47.23
Mesh-Graphormer [33]	63.31	47.15	34.12
Ours (HMR [26] + *RaBit*)	**51.46**	**37.80**	**25.97**

Table 1. **Quantitative results of shape reconstruction.** Our method achieves the best results in terms of MPVE, MPJPE and PA-MPJPE. Note that all metrics are measured in a unit 10^{-3}m.

Results on Texture Inference. To demonstrate the capability of our proposed GAN-based texture generator, we first compare our *RaBit*-based texture inference (i.e., *BiCarNet*) with PCA-based inference. Specifically, for PCA-based method, we utilize the same learning architecture to map the input image into the PCA-based texture space. Furthermore, to evaluate the effectiveness of the proposed texture inference modules, we also conduct ablative analysis on *BiCarNet* without Fuser and *BiCarNet* without Part-sensitive Reasoner (PSR). Table 2 shows the quantitative results of different texture inference methods on MSE, PSNR and FID and our proposed method achieves the highest scores compared with all other methods. Moreover, Fig. 8 illustrates the qualitative

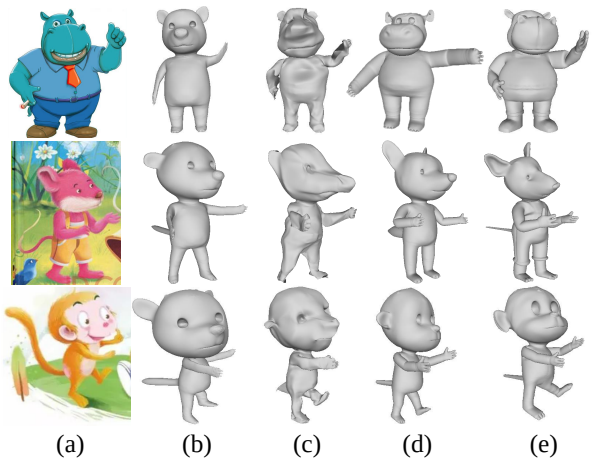


Figure 7. **Qualitative results of shape reconstruction.** From left to right, each row contains (a) the input image, reconstructed meshes of (b) Mesh Graphormer, (c) DecoMR, (d) our method, and (e) the ground truth mesh.

results of these methods for texture inference. Fig. 9 shows the results without and with Fuser, which demonstrates our fusion module can deal with unnature seam-like artifacts. We can observe that the part-sensitive texture reasoner and the Fuser help to capture the local regions of characters and recover their detailed appearances.

Method	MSE($\times 10^{-1}$)↓	PSNR($\times 10^{2}$) ↑	FID ↓
PCA	0.2309	0.2254	0.4642
BiCarNet	**0.1093**	**0.2458**	**0.1133**
BiCarNet w/o Fuser	0.1108	0.2397	0.1407
BiCarNet w/o PSR	0.1346	0.2361	0.4024

Table 2. **Quantitative results on texture inference.** PCA denotes linear modeling method for texture and the last two rows indicate the results of *BiCarNet* respectively without two designed module. Our *BiCarNet* outperforms others methods in all metrics.

6. More Applications

6.1. Sketch-based Modeling

Customizing 3D biped cartoon characters usually requires a heavy workload with commercial tools, even for experienced artists. Sketch-based modeling enables amateur users to get involved in 3D shape customization in a simple and intuitive fashion. In this section, we build a sketch-based modeling application with the help of *3DBiCar* and *RaBit*.

We first sample a series of shape vectors randomly and feed them to *RaBit* to generate 3D cartoon characters with diversified shapes, resulting in 12,000 T-pose models. Then we apply suggestive contour [24] to render the front-view sketches with different abstraction levels and obtain 108,000

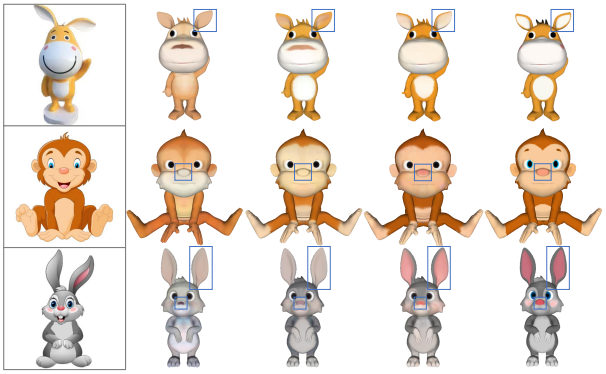


Figure 8. **Qualitative comparisons on texture inference.** The input image (a) is followed by the textured models from (b) PCA, (c) *BiCarNet* w/o PSR, (d) *BiCarNet* and (e) the ground truth. Note that we use the same shape and focus on the difference of textures.

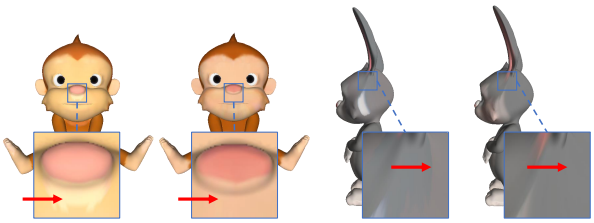


Figure 9. **Qualitative ablation on Fuser in Texture inference.** Left: *BiCarNet* w/o Fuser. Right: *BiCarNet* with Fuser.

sketch-model pairs. Given a sketch as input, we employ ResNet-50 and three MLPs as encoder and decoder to map the input sketch to 100-dimensional shape parameters. The generated shape parameters are next fed to *RaBit* to reconstruct the corresponding 3D model. Please refer to the supplemental materials for more details. Note that users only need to depict a 3D character with T-pose on a 2D canvas while the output characters of our method are animation-ready and can be directly applied to other commercial tools. Fig. 10 displays the sketches created by users with little knowledge of modeling as well as the corresponding models generated by our method. It can be seen that our sketch-based modeling application offers a smart approach for amateur users to create 3D biped cartoon characters with diversified shapes. We will further explore the support of shape reconstruction from sketches with arbitrary poses, and texture painting in the future.

6.2. 3D Character Animation

In previous sections, we have demonstrated the effectiveness of *3DBiCar* and *RaBit* in supporting 3D biped cartoon characters generation from easy-to-obtain inputs, i.e., single-view images and sketches. This section further demonstrates the usability of our generated models for character animation.

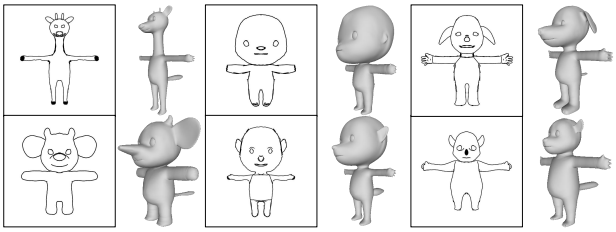


Figure 10. **Result gallery of sketch-based modelling.** The sketch created by amateur users denotes on the left and generated models on the right.

Following the recent advance of human recovering method [48] and parametric model [34], we first extract the human from videos frames and then adopt a temporal-aware encoder to recover the sequence of human poses. Next, a motion retargeting method [25] is used to convert the poses on the human skeleton to the motion of our cartoon characters. As shown in Fig. 11, animation-ready characters based on our *RaBit* can be directly applied to 3D animation. Please refer to the supplementary for animation videos.

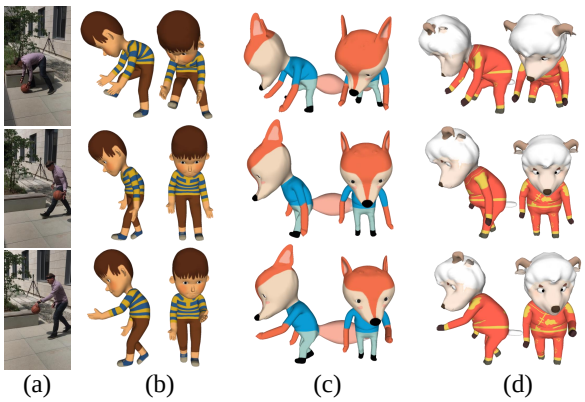


Figure 11. **Transferring motion of a human video to animate characters.** (a) denotes the input frames. (b), (c), and (d) indicate three corresponding posed cartoon characters.

7. Conclusion

In this work, we introduce *3DBiCar*, the first large-scale 3D biped cartoon character dataset. It contains 1,500 textured and skinned models with a consistent mesh topology. Based on *3DBiCar*, we propose the first 3D full-body cartoon parametric model *RaBit* for biped character modeling. Furthermore, we propose a baseline method *BiCarNet* for reconstructing 3D textured models from a single image with cartoon characters. To better capture texture for local regions, a novel part-sensitive texture reasoner is also contributed. Experimental results demonstrate the capability of *3DBiCar* and *RaBit* as well as the effectiveness of *BiCarNet*. Last but not least, two further applications, i.e., sketch-based model-

ing and 3D character animation, demonstrate the usability and practicality of our dataset and parametric model. We hope that our work can open a door for further researches in the area of cartoon character digitalization.

References

[1] Victoria Fernández Abrevaya, Stefanie Wuhrer, and Edmond Boyer. Multilinear autoencoder for 3d face model learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2018. 3

[2] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. *ACM transactions on graphics (TOG)*, 22(3):587–594, 2003. 3

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 2, 3

[4] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018. 3

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 2, 3, 5

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 2, 3

[7] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3794 –3801, Columbus, Ohio, USA, June 2014. 2

[8] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7213–7222, 2019. 3

[9] Kirill Brodt and Mikhail Bessmeltsev. Sketch2pose: estimating a 3d character pose from a bitmap sketch. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 2

[10] Alan Brunton, Augusto Salazar, Timo Bolkart, and Stefanie Wuhrer. Review of statistical shape spaces for 3d data with comparative analysis for human faces. *Computer Vision and Image Understanding*, 128:1–17, 2014. 2

[11] Hongrui Cai, Yudong Guo, Zhuang Peng, and Juyong Zhang. Landmark detection and 3d face reconstruction for caricature using a nonlinear parametric model. *Graphical Models*, 115:101103, 2021. 3

[12] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 2, 3, 5

[13] Zhi-Quan Cheng, Yin Chen, Ralph R Martin, Tong Wu, and Zhan Song. Parametric modeling of 3d human body shape—a survey. *Computers & Graphics*, 71:88–100, 2018. 2

[14] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1964–1973, 2021. 3

[15] Vasileios Choutas, Lea Müller, Chun-Hao P Huang, Siyu Tang, Dimitrios Tzionas, and Michael J Black. Accurate 3d body shape regression using metric and semantic attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2718–2728, 2022. 3

[16] Hang Dai, Nick Pears, William AP Smith, and Christian Duncan. A 3d morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE international conference on computer vision*, pages 3085–3093, 2017. 3

[17] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7093–7102, 2018. 3

[18] Bernhard Egger, Dinu Kaufmann, Sandro Schönborn, Volker Roth, and Thomas Vetter. Copula eigenfaces with attributes: semiparametric principal component analysis for a combined color, shape and attribute model. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics*, pages 95–112. Springer, 2016. 3

[19] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 3, 5

[20] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *Advances In Neural Information Processing Systems*, 2022. 3

[21] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1155–1164, 2019. 3, 5

[22] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5151–5160, 2021. 3, 5

[23] Fang Han and Han Liu. Semiparametric principal component analysis. *Advances in Neural Information Processing Systems*, 25, 2012. 3

[24] Xiaoguang Han, Chang Gao, and Yizhou Yu. Deepsketch2face: a deep learning based sketching system for 3d face and caricature modeling. *ACM Transactions on graphics (TOG)*, 36(4):1–12, 2017. 2, 3, 7

[25] Ming-Kai Hsieh, Bing-Yu Chen, and Ming Ouhyoung. Motion retargeting and transition in different articulated figures. In *Ninth International Conference on Computer Aided Design and Computer Graphics (CAD-CG'05)*, pages 6–pp. IEEE, 2005. 8

[26] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2, 3, 6, 7

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 5, 6

[29] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 3

[30] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2

[31] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 3, 5

[32] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 6

[33] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 6, 7

[34] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 3, 5, 8

[35] Zhongjin Luo, Jie Zhou, Heming Zhu, Dong Du, Xiaoguang Han, and Hongbo Fu. Simpmodeling: Sketching implicit field to guide mesh modeling for 3d animalmorphic head design. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 854–863, 2021. 2, 3

[36] Stylianos Moschoglou, Evangelos Ververas, Yannis Panagakis, Mihalis A Nicolaou, and Stefanos Zafeiriou. Multi-attribute robust component analysis for facial uv maps. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1324–1337, 2018. 3

[37] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2019. 3

[38] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2, 3, 5

[39] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 3

[40] Yuda Qiu, Xiaojie Xu, Lingteng Qiu, Yan Pan, Yushuang Wu, Weikai Chen, and Xiaoguang Han. 3dcaricshop: A dataset and a baseline method for single-view 3d caricature face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10245, 2021. 2, 3

[41] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 704–720, 2018. 3

[42] Kathleen M Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, and Scott Fleming. Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary. Technical report, Sytronics Inc Dayton Oh, 2002. 2

[43] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017. 3

[44] Alexandre Saint, Eman Ahmed, Abd El Rahman Shabayek, Kseniya Cherenkova, Gleb Gusev, Djamila Aouada, and Bjorn Ottersten. 3dbodytex: Textured 3d body dataset. In *2018 International Conference on 3D Vision (3DV)*, pages 495–504, 2018. 2

[45] Matan Sela, Yonathan Aflalo, and Ron Kimmel. Computational caricaturization of surfaces. *Computer Vision and Image Understanding*, 141:1–17, 2015. 3

[46] Ron Slossberg, Gil Shamai, and Ron Kimmel. High quality facial surface and texture synthesis via generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[47] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5

[48] Zhouping Wang and Sarah Ostadabbas. Live stream temporally embedded 3d human body pose and shape estimation. *arXiv preprint arXiv:2207.12537*, 2022. 3, 8

[49] Qianyi Wu, Juyong Zhang, Yu-Kun Lai, Jianmin Zheng, and Jianfei Cai. Alive caricature from 2d to 3d. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7336–7345, 2018. 3

[50] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 3

[51] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 598–607, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. 2

[52] Yipin Yang, Yao Yu, Yu Zhou, Sidan Du, James Davis, and Ruigang Yang. Semantic parametric reshaping of human body models. In *2014 2nd International Conference on 3D Vision*, volume 2, pages 41–48, 2014. 2

[53] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7054–7063, 2020. 6, 7

[54] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[55] Yi Zhou, Chenglei Wu, Zimo Li, Chen Cao, Yuting Ye, Jason Saragih, Hao Li, and Yaser Sheikh. Fully convolutional mesh autoencoder using efficient spatially varying kernels. *Advances in Neural Information Processing Systems*, 33:9251–9262, 2020. 3