DiffSurf: A Transformer-based Diffusion Model for Generating and Reconstructing 3D Surfaces in Pose

Yusuke Yoshiyasu[1] and Leyuan Sun[1]

National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Japan {yusuke-yoshiyasu,son.leyuansun}@aist.go.jp

Abstract. This paper presents DiffSurf, a transformer-based denoising diffusion model for generating and reconstructing 3D surfaces. Specifically, we design a diffusion transformer architecture that predicts noise from noisy 3D surface vertices and normals. With this architecture, DiffSurf is able to generate 3D surfaces in various poses and shapes, such as human bodies, hands, animals and man-made objects. Further, DiffSurf is versatile in that it can address various 3D downstream tasks including morphing, body shape variation and 3D human mesh fitting to 2D keypoints. Experimental results on 3D human model benchmarks demonstrate that DiffSurf can generate shapes with greater diversity and higher quality than previous generative models. Furthermore, when applied to the task of single-image 3D human mesh recovery, DiffSurf achieves accuracy comparable to prior techniques at a near real-time rate. https://github.com/yusukey03012/DiffSurf

Keywords: Diffusion model · 3D surface · Human mesh recovery

1 Introduction

Creating and reconstructing 3D shape models in various shapes and poses is a significant challenge in computer vision and computer graphics, with extensive applications in gaming, augmented reality (AR) and virtual reality (VR). For such 3D content-based applications, a surface mesh is the most commonly used shape representation because of its efficiency during graphics rendering and user-friendliness for artists.

Over the past few years, diffusion models [28, 65] have revolutionized the content creation paradigm in the image domain, particularly in the task of image generation from text prompts. Diffusion models can generate high-quality and diverse data by learning to reverse the diffusion process. This process gradually constructs desired data samples from noise whose dimensionality is higher than that of previous generative models such as generative adversarial networks (GANs). Additionally, diffusion models have been applied to the generation of 3D data, such as object point clouds [49, 51, 91], textured 3D models [41, 63, 81], scene radiance field [6] and 3D human pose [25]. Recent 3D shape diffusion models are able to generate 3D surfaces with complex geometry and topology by incorporating implicit functions, such as signed distance fields (SDF), and extracting their zero level-set surface using the marching cubes algorithm [13, 71].
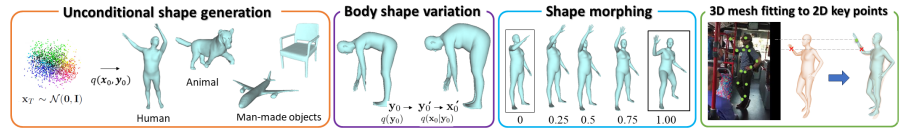
Fig. 1: DiffSurf addresses the unconditional generation of 3D surfaces in diverse poses. It can generate 3D surfaces of various objects types such as humans, mammals and man-made objects. Downstream tasks, including unconditional generation, morphing and fitting to 2D key points can be addressed with pre-trained DiffSurf models.

Yet, there remain several challenges in generating 3D surfaces based on diffusion models. **Firstly**, the current approaches do not consider pose. In the case of 3D human and animal generation in different poses, point-to-point correspondences between shapes are important but they are lost when employing the current 3D shape diffusion models. There exist recent works of diffusion models for generating 3D human poses and shapes conditioned on images [14, 25, 40, 42] but generation of diverse body shapes and poses are not considered. **Secondly**, we want our generative model to handle a wide range of objects, such as human bodies, mammals and man-made objects. **Thirdly**, a framework that can deal with a wide variety tasks, e.g. interpolating two shapes, altering pose and manipulating shape, is highly sought after.

In this paper, we propose a transformer-based diffusion model for generating and reconstructing 3D surfaces (dubbed DiffSurf). To address the aforementioned challenges, we design a diffusion transformer architecture that predicts noise from noisy 3D surface vertex coordinates. By representing a surface with points and normals, processing them in diffusion transformer and then employing up-samplers dedicated for topologically fixed and varied cases, DiffSurf is able to handle diverse body poses, various object types and multiple different tasks as illustrated in Figs. 1 and 2. To our knowledge, DiffSurf is the first diffusion model that addresses generation of 3D surfaces in diverse body poses and shapes. The contributions of this paper are summarized as follows:

1. DiffSurf, a denoising diffusion transformer model that can generate 3D surfaces in various body shapes and poses.
2. It can generate 3D shapes of diverse object types, such as human bodies, mammals and man-made objects, using a diffusion transformer model that leverages point-normal representation.
3. It provides methodologies for addressing various 3D processing and image-to-3D downstream tasks by effectively utilizing pre-trained DiffSurf models based on score distillation sampling (SDS) and classifier-free guidance (CFG).

2   Related Work

Generative models for 3D shape and pose Previous generative models for 3D shape and pose generation predominantly utilize variational autoencoders (VAEs) [3, 18, 24, 30, 76, 77, 89, 93], generative adversarial networks (GANs) [20, 34] or normalizing flows [7, 82, 90]. SMPL-X [58] employs a VAE to learn a pose prior and enforces constraints on body joint angles. COMA [64] and CAPE [52] designed their VAEs based

on graph neural networks to model facial expressions and clothing geometric deforma-tions, respectively. Kanazawa et al. [34] introduced the human mesh recovery (HMR) technique that estimates a posed human body model from a single image by employing GANs to to provide body pose and shape priors. Other approaches focus on designing and obtaining latent encoders or representation using GANs [11, 12, 20]. Recent tech-niques for 3D human pose and shape estimation employ normalizing flows [7,39,82,90] in attempting to learn 3D priors from motion capture datasets. Pose-NDF instead rep-resents and learns a pose space using neural distance fields [79].

Diffusion models and 3D generation In 3D generation, Luo et al. [49] proposed the first 3D point cloud generation method based on diffusion models by extending Denois-ing Diffusion Probabilistic Models (DDPM) [28]. LION [91] and SLIDE [51] adopted Latent Diffusion Models (LDM) [65] in point cloud generation, aiming to reduce point cloud resolution for more efficient training and sampling. A learning-based surface re-construction technique called ShapeAsPoint (SAP) [61] is then employed to convert the generated point clouds into volume functions and subsequently into meshes. To gener-ate 3D surfaces, recent approaches use implicit functions in diffusion process such as SDF, and extract a mesh from 3D volume using marching cubes [13, 71, 88]. MeshDif-fusion [47] uses DMTet [70] that combines a tetra mesh and SDF to represent the object shape to generate topologically and geometrically complex objects. Mo et al. proposed DiT-3D [56] which extends diffusion transformer [4, 5, 59, 60] to voxelized 3D point clouds, accomplishing the generation of 3D objects. PolyDiff introduced a 3D diffu-sion model that can work with polygonal meshes [1]. Point-e [57] and Shape-e [33] extend [49] to colored point clouds using transformers.

3D shape and pose from image Human mesh recovery approaches [78] predict a 3D human body mesh from a single image or video frames, which can be roughly di-vided into parametric [8, 34, 92] and vertex-based approaches [16, 38, 43]. Parametric approaches regress the body shape and pose parameters of human body models like SMPL or SMPL-X. On the other hand, vertex-based approaches directly regress from an image to 3D vertex coordinates. Transformer architectures, which are known for their ability to capture long-range dependencies, have been employed in vertex-based human mesh recovery and shown strong performances [15,43,44,86,87]. The reconstruction of mammals from images has also been addressed in the field [67,83,94]. Research on 3D human body pose and shape generation via diffusion models has recently commenced but most being task-specific and conditional on 2D data. They include 3D human pose estimation methods from 2D keypoints [25,69], parametric human mesh recovery tech-niques [14,46] and vertex-based human mesh recovery methods [40,42].

3   Background: Diffusion models

Diffusion models establish a Markov chain of diffusion steps by gradually adding ran-dom noise to data (i.e., a forward diffusion process) and learn to reverse this process to construct desired data samples from the noise (i.e., a reverse diffusion process). Con-sidering a data point sampled from a data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$, the forward process produces a sequence of noisy samples $\mathbf{x}_1 \ldots \mathbf{x}_T$ by adding a small amount of Gaussian

noise to the sample in T steps:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{1}$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

where $\{\beta_t\}_1^T \in (0,1)$ is the noise variance schedule.

Data generation can be initiated from a Gaussian noise input $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, provided that the aforementioned forward process is reversed to sample from $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$. However, the calculation of $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ depends on the entire dataset and is not straightforward. To address this, a neural network model p_θ is used to approximate these conditional probabilities with a Gaussian model:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma^2\mathbf{I}) \tag{3}$$

where t is a timestep uniformly sampled from $1, 2, \ldots, T$. Then, $\mu_\theta(\mathbf{x}_t, t)$ can be rewritten with noise prediction $\epsilon_\theta(\mathbf{x}_t, t)$ as:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t) \right) \tag{4}$$

where $\alpha_t = 1 - \beta_t$ and $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 - \sqrt{1-\bar{\alpha}_t}\epsilon$ with $\bar{\alpha}_t = \prod_{t=1}^{t} \alpha$. To learn to estimate $\epsilon_\theta(\mathbf{x}_t, t)$ between consecutive samples $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, the training loss is defined as follows:

$$L = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||_2^2 \tag{5}$$

UniDiffuser UniDiffuser [5] introduced an approach to handle multi-modal data distributions in the diffusion process in a unified manner. It defines the conditional expectations in a general form, $\mathbb{E}[\epsilon_x, \epsilon_y|\mathbf{x}_{t^x}, \mathbf{y}_{t^y}]$, for all $0 \leq t^x, t^y \leq T$, where t^x and t^y represent two potentially different timesteps. $\mathbf{x}_{t^x}$ and $\mathbf{y}_{t^y}$ are the corresponding perturbed data. With this formulation, marginal diffusion, conditional diffusion and joint diffusion can be achieved by setting $t^y = T$, $t^y = 0$ and $t^x = t^y = t$, respectively. To this end, a joint noise prediction network $\epsilon_\theta(\mathbf{x}_{t^x}, \mathbf{y}_{t^y}, t^x, t^y)$ is employed to predict noise $\epsilon_\theta = [\epsilon_\theta^x, \epsilon_\theta^y]$ injected into $\mathbf{x}_{t^x}$ and $\mathbf{y}_{t^y}$, which is trained by minimizing the following loss:

$$\mathcal{L}_{\text{uni}} = \mathbb{E}_{\mathbf{x}_0,\mathbf{y}_0,\epsilon_x,\epsilon_y,\mathbf{x}_{t^x},\mathbf{y}_{t^y}} ||[\epsilon^x, \epsilon^y] - \epsilon_\theta(\mathbf{x}_{t^x}, \mathbf{y}_{t^y}, t^x, t^y)||_2^2$$

where $\mathbf{x}_0$ and $\mathbf{y}_0$ are the data points, ϵ_x and ϵ_y are sampled from Gaussian distributions, and t^x and t^y are independently and uniformly sampled from the range $1, 2, \ldots, T$. Furthermore, UniDiffuser is directly applicable to classifier-free guidance (CFG), which is a method introduced to enhance the sample quality of conditional diffusion models [27], without modifying the training loss. For instance, $\mathbf{x}_0$ conditioned on $\mathbf{y}_0$ can be generated as follows:

$$\hat{\epsilon}_\theta^x(\mathbf{x}_t, \mathbf{y}_0, t) = (1+s_g)\epsilon_\theta^x(\mathbf{x}_t, \mathbf{y}_0, t, 0) - s_g\epsilon_\theta^x(\mathbf{x}_t, t, T) \tag{6}$$
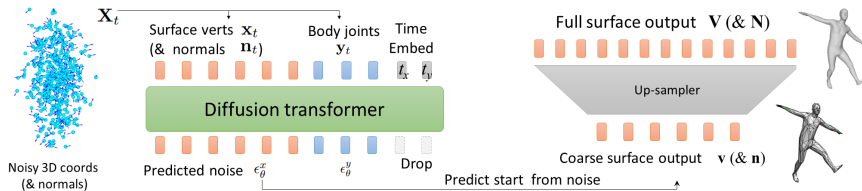
Fig. 2: Overview. DiffSurf consists of a diffusion transformer and an up-sampler. The diffusion transformer takes in the noisy 3D coordinates of surface vertices $\mathbf{x}_t \in \mathbb{R}^{N \times 3}$ and body joints $\mathbf{y}_t \in \mathbb{R}^{J \times 3}$. It processes these two modalities of data along with their corresponding timestep tokens t_x and t_y. The transformer then outputs noise predictions for vertex and joint tokens, ϵ_θ^x and ϵ_θ^y, respectively. For the 3D surface generation of man-made objects, we also input the noisy surface normals $\mathbf{n}_t \in \mathbb{R}^{N \times 3}$ corresponding to vertex tokens into the diffusion transformer. Once the 3D coordinates of surface vertices $\mathbf{v}$ (and normals $\mathbf{n}$) are generated, up-sampling is optionally performed to obtain the full dense surface output $\mathbf{V}$ (and $\mathbf{N}$).

where s_g is a guidance scale and $\epsilon_\theta^x(\mathbf{x}_t, \mathbf{y}_0, t, 0)$ and $\epsilon_\theta^x(\mathbf{x}_t, t, T)$ are the conditional and unconditional models, respectively.

Score Distillation Sampling (SDS) A loss calculation framework called Score Distillation Sampling (SDS) is proposed by DreamFusion [63] for utilizing pre-trained diffusion models in optimizing and regularizing a neural 3D scene model. The scene model is defined by a parametric function of the form $x = g(\phi)$ capable of generating an image x from the desired camera pose. In this context, g is a volumetric renderer such as NeRFs and ϕ is a Multi-Layer Perceptron (MLP) that models a 3D volume. Given a text condition $\mathbf{y}_0$, SDS derives the gradient to update ϕ in such a way that:

$$\nabla_\phi \mathcal{L}_{\text{SDS}}(\phi, g(\theta)) = \mathbb{E}_{t, \epsilon}\left[\omega(t)(\epsilon_\theta(\mathbf{x}_t, \mathbf{y}_0, t) - \epsilon)\frac{\partial \phi}{\partial x}\right]$$

where $\omega(t)$ is a weighting function. In practice, the conditional noise prediction model $\epsilon_\theta(\mathbf{x}_t, \mathbf{y}_0, t)$ is replaced with the classifier free guidance one, $\hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{y}_0, t)$.

4 Method

We introduce DiffSurf, a general network architecture for generating, editing and reconstructing 3D surfaces based on a plain diffusion transformer model, which is extendable to a wide range of object types. In addition, we introduce downstream methodologies to leverage pre-trained DiffSurf models for solving various 3D processing tasks.

For the generation of posed 3D surfaces, we propose to incorporate the vertex-based mesh recovery paradigm [15, 43, 44] from human mesh recovery into 3D shape generation. Then, point-to-point correspondences between shapes are ensured by learning from training meshes with the same connectivity. This is a straightforward yet effective strategy, which is overlooked by the recent 3D shape generation literature, when unconditionally generating human and animal 3D surfaces in different poses where correspondences across shapes are vital. To the best of our knowledge, DiffSurf is the first 3D diffusion model capable of unconditionally generating 3D surfaces of articulated objects in diverse poses.

As for generating man-made objects, we propose a strategy for capturing better geometry by incorporating surface normals into the diffusion process. This approach not only leads to better generation results but also provides a more informative input for the subsequent surface recon-

Fig. 3: w/o and with surface normals.
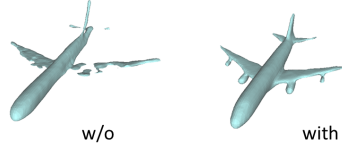
struction post-processing [51, 61, 91] than using point sets alone, as shown in Fig. 3. This is particularly useful for generating 3D surfaces with complex geometry and topological differences, although point-to-point correspondences may be lost in this case. Unlike the recent diffusion models based on implicit functions [13, 56], in the diffusion process, we avoid using a volumetric representation whose complexity grows cubically w.r.t voxel resolution. Instead, we directly work on an explicit point representation through the diffusion transformer to exploit long-range dependencies between points. As a result, DiffSurf only needs a single diffusion model as opposed to previous hierarchical latent diffusion models which rely on two diffusion models [91], making our model computationally more efficient.

4.1   Network architecture

We describe our diffusion transformer architecture for generating 3D surfaces. It draws inspiration from vertex-based human mesh recovery approaches [15, 43, 44] and point cloud latent diffusion models [49, 51, 91], which process explicit 3D point-based representations in neural network models. Specifically, we have designed a UniDiffuser-like multi-modal diffusion transformer architecture [84] that predicts noise from noisy 3D coordinates of surface vertices and body joints, treating them as two distinct modalities. The incorporation of body joints not only facilitates more effective training [43] but also provides landmark controls [51] for manipulating shape and pose. Consequently, we introduce a simple yet versatile transformer architecture for generating and reconstructing 3D surfaces of articulated objects in various shapes and poses, thereby enabling various downstream 3D processing tasks as illustrated in Fig. 2. DiffSurf consists of 1) a diffusion transformer and 2) a mesh up-sampler.

Diffusion transformer blocks The inputs to the diffusion transformer consist of noisy 3D coordinates for a set of joint query tokens $Q_{\mathrm{J}} = \{Q_{\mathrm{J}}^1 \dots Q_{\mathrm{J}}^J\}$ and coarse vertex query tokens $Q_{\mathrm{V}} = \{Q_{\mathrm{V}}^1 \dots Q_{\mathrm{V}}^N\}$, corresponding to an articulated body mesh comprising J joints and N vertices. The input noisy 3D coordinates of surface vertices, body joints and their concatenations are respectively denoted as $\mathbf{x}_t \in \mathbb{R}^{N \times 3}$, $\mathbf{y}_t \in \mathbb{R}^{J \times 3}$ and $\mathbf{X}_t \in \mathbb{R}^{(J+N) \times 3}$. The diffusion transformer processes these two modalities of data and their corresponding timesteps t_x and t_y as tokens. It outputs noise predictions for vertices and joints, ϵ_θ^x and ϵ_θ^y. For the generation of man-made objects, we concatenate the noisy 3D coordinates of vertices $\mathbf{x}_t$ with the corresponding surface normals $\mathbf{n}_t \in \mathbb{R}^{N \times 3}$ to construct an $N \times 6$ matrix, which is then input to the diffusion transformer. Our diffusion transformer consists of 7 layers of transformer blocks and input/output MLP layers. Each transformer block has hidden layers with the dimension of 256 channels. The input MLP layer converts $\mathbf{x}_t$ and $\mathbf{y}_t$ into 256-dimensional embedding features and the output MLP layer converts the features processed by transformer into ϵ_θ^x and ϵ_θ^y.

Up-sampling After a coarse surface comprising N vertices, $\mathbf{v} \in \mathbb{R}^{N \times 3}$ and corresponding surface normals $\mathbf{n} \in \mathbb{R}^{N \times 3}$ are produced from the noise prediction ϵ_θ^x by the diffusion transformer, we apply an upsampling operation. For human and animal generation, where point-to-point correspondences i.e. mesh connectivity is fixed, an upsampling technique based on MLPs similar to [43, 44, 86] is adopted to obtain a dense mesh (see Fig. 2) with M vertices, $\mathbf{V} \in \mathbb{R}^{M \times 3}$. For the surface generation of man-made objects, refinement and upsampling based on the improved PointNet++ model [50] are applied to increase the number of points and normals by a factor of $\times 5$ [51]. Subsequently, a learning-based surface reconstruction technique called Shape-As-Points (SAP) [61] is employed to convert the upsampled points $\mathbf{V}$ and normals $\mathbf{N}$ into a mesh.

4.2 Downstream methodologies

Here, we demonstrate that DiffSurf is capable of performing a series of downstream tasks in 3D surface editing and reconstruction.

Pose conditioned mesh generation DiffSurf can generate a mesh that is conditioned on 3D skeleton landmark locations, such as those obtained using motion capture and image-based 3D pose regressors. Essentially, this process of conditional mesh generation involves feeding 3D body joint locations as queries into the diffusion transformer and setting the timestep to $t_y = 0$. Naively feeding 3D joint locations into DiffSurf results in slight discrepancies between the mesh and the joints. To address this, we leverage CFG (Eq. (6)) to push the mesh toward joint locations and improve alignments between them. We found that setting the CFG weight to around $s_g = 1$ effectively improves alignment while preserving the mesh structure. Excessively increased CFG weights, e.g., $s_g > 3$, can result in distortion of the mesh (as shown in the Appendix).

Body shape variation By feeding a skeleton to DiffSurf as a condition and performing sampling with varying random noise, we can generate meshes in different body shapes, as shown in Fig. 1 (right). However, this approach does not allow for changes in body heights and segment lengths. To address this, we adopt a two-step strategy. The first step involves unimodal generation to create a batch of skeletons with different body poses and styles. The second step then adjusts the segment lengths of one of generated skeletons based on others in the batch. By inputting these modified skeletons into DiffSurf, we can generate meshes in diverse body styles while maintaining the pose.

Shape morphing DiffSurf is capable of morphing between two meshes with different poses and shapes. This is achieved by blending two meshes represented as the Gaussian noise, $\mathbf{x}_T^1$ and $\mathbf{x}_T^2$, through spherical linear interpolation (SLERP) [72], $\hat{\mathbf{x}}_T = \text{SLERP}(\mathbf{x}_T^1, \mathbf{x}_T^2, w)$, where w is an interpolation weight within the range $[0, 1]$. Setting w outside this range $[0, 1]$ e.g., $[-0.25, 1.25]$, results in extrapolation. Given the new noise $\hat{\mathbf{x}}_T$, a mesh is then sampled from it using DiffSurf. It is noteworthy that DiffSurf has the capability to simultaneously handle both shape and pose variations.

Shape refinement Instead of starting from Gaussian noise to generate a mesh, DiffSurf can refine a mesh exhibiting noise and distortions by leveraging the SDS gradients as calculated in Eq. (7). Analogous to mesh fairing [21], a sequence of refined meshes can be constructed by explicitly applying the SDS gradients to a mesh:

$$\mathbf{X}^{l+1} = \mathbf{X}^l - \nabla_\phi \mathcal{L}_{\text{SDS}}(\phi, \mathbf{X}^l) \tag{7}$$

$L_{\text{SDS}} + L_{\text{CP}}$ ··· $+ L_{\text{consist}}$ ··· $+ L_{\text{SDS}}^{\text{edge}} + L_{\text{SDS}}^{\text{lap}}$
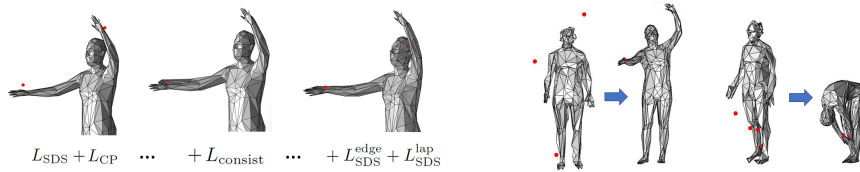
Fig. 4: Control point deformation. Left: comparison of the loss terms. Right: deformation examples obtained using five control points (head, wrists and ankles, marked in red). The process starts with the rest pose and a human mesh is deformed to align with the control points. DiffSurf can deform a mesh into extremely different poses, such as a forward-bending posture.

It should be noted that, in contrast to DreamFusion [63], $\mathbf{X}^l$ is not an output from a neural network model but rather the 3D coordinates of surface vertices and body joints.

Control point deformation DiffSurf facilitates data-driven mesh deformation, similar to prior shape editing techniques [22, 23, 75], by creating a mesh in a plausible shape and pose through the specification of a set of control points and fitting the mesh toward them. Building upon [63], we devise a loss function derived from differential mesh properties of the SDS target mesh to preserve local geometries of the mesh. We minimize the following total loss L_{def} to optimize $\mathbf{X}$:

$$L_{\text{def}}(\mathbf{X}) = L_{\text{SDS}} + L_{\text{SDS}}^{\text{edge}} + L_{\text{SDS}}^{\text{lap}} + L_{\text{consist}} + L_{\text{CP}} \tag{8}$$

where L_{SDS}, $L_{\text{SDS}}^{\text{edge}}$ and $L_{\text{SDS}}^{\text{lap}}$ represent losses defined by the distances between the SDS targets and predictions for the 3D vertex coordinates, edges and Laplacian coordinates of the coarse mesh, respectively. L_{consist} maintains the consistency between the joint and mesh predictions, defined by the distances between the optimized joints and the regressed joints, which are calculated from the coarse mesh vertices using the joint regressor. The regressed joints $\mathbf{j}_{\text{reg}} \in \mathbb{R}^{J \times 3}$ are calculated from the mesh vertices using the joint regressor, $\mathbf{j}_{\text{reg}} = \mathcal{J}\mathbf{V}$, where $\mathcal{J} \in \mathbb{R}^{J \times M}$ is a joint regressor matrix [48, 66, 94]. L_{CP} quantifies the distances between the optimized joint locations and the control points (see the Appendix for more details).

Figure 4 (left) shows the comparisons of the loss terms. Using L_{SDS} and L_{CP}, Diff-Surf is able to fit a human mesh towards control points, but there remain some distances between them. Incorporating L_{consist} improves the fit but introduces distortions around the control points. Adding $L_{\text{SDS}}^{\text{edge}}$ and $L_{\text{SDS}}^{\text{lap}}$ remedies this issue by considering the differential properties of the coarse mesh to preserve its local geometry.

2D keypoint fitting While previous research [8, 39] has addressed the challenge of fitting a parametric model to 2D landmarks, we propose an alternative vertex-based fitting approach for this task. Consequently, our method is applicable to both parametric and vertex-based mesh recovery approaches and improves their mesh recovery results.

Starting from the initial solution derived from a mesh recovery approach, DiffSurf optimizes mesh vertices and body joints to improve their alignment with 2D keypoint locations. The loss function for 2D keypoint fitting is defined by modifying Eq. (8) slightly as follows:

$$L_{\text{fit}}(\mathbf{X}) = L_{\text{SDS}} + L_{\text{SDS}}^{\text{edge}} + L_{\text{SDS}}^{\text{lap}} + L_{\text{consist}} + L_{\text{2D}} \tag{9}$$

where L_{2D} measures the discrepancies between the ground truth and predicted 2D keypoints. These 2D keypoint predictions are obtained by projecting 3D joint positions using the camera parameter predictions from the mesh recovery approach. In addition, as DiffSurf is trained on the dataset with global position and rotation aligned at the root, we rigidly align the mesh prediction with the canonical orientation before subtracting the SDS gradients, such that: $T(\mathbf{X} - \nabla_\phi \mathcal{L}_{SDS}(\phi, T^{-1}(\mathbf{X})))$, where T is the global transformation of the mesh recovery result w.r.t the canonical orientation. The global rotation can be the root orientation predicted by the mesh recovery approach when the global pose is available. Otherwise, a coordinate frame defined from the body joint predictions can be used to obtain T e.g., the x-axis and y-axis are defined from the unit vectors emanating from the pelvis to the neck and from the right hip to the left hip.

Mesh generation from 3D keypoints DiffSurf is capable of reconstructing a 3D mesh from 3D joint locations by employing the pre-trained DiffSurf model for pose-conditioned surface generation. To achieve this, we first predict 3D body joint locations from an image using a 3D pose regressor, which produces the 3D positions of 14 body joints. Subsequently, these 3D joint positions are inputted into DiffSurf to perform conditional mesh generation with Eq. (6). Similar to the SDS-based 2D keypoint fitting, this approach requires aligning the mesh with the canonical orientation prior to mesh sampling. It is noteworthy that this modular human mesh recovery design, based on DiffSurf, decouples an image-based 3D pose regressor from a mesh generator, enabling its training even when image-mesh paired data are not available. The architecture of our 3D pose regressor used here is transformer-based (see the Appendix).

5   Experiments

5.1   Dataset and metrics

3D generation We trained our DiffSurf models separately on publicly available 3D datasets: SURREAL [80], AMASS [53], FreiHAND [17], BARC [67], Animal3D [83] and ShapeNet [10]. We follow 3D-CODED [26] for the SURREAL train/test split definition. The global positions of meshes used in the training are aligned at the root position and oriented to face forward. The body joints are obtained from meshes using joint regressors [48,94] for humans and animals. For ShapeNet objects, we feed sparse latent points generated by SLIDE [51] as body joint tokens to transformer.

Evaluation of 3D human generation was conducted on the SURREAL testset (200 meshes) and DFAUST [9] (800 meshes). The standard metric used for evaluating 3D generation is the 1-NNA metric [85], which quantifies the distributional similarity between generated shapes and the validation set. This metric assesses both the quality and diversity of the generated results. For human generation, given the differences in global orientations between validation and training meshes, we first perform a rigid alignment of the predicted mesh with the validation meshes before calculating the 1-NNA metric. We refer to this modified metric as Rigid Aligned 1-NNA (RA-1-NNA).

Human mesh recovery We trained our 3D pose regressor using publicly available datasets, adopting the mixed dataset training strategies as outlined in [36, 43]. The datasets used include Human3.6M [29], MPI-INF-3DHP [55], COCO [45], MPII [2]
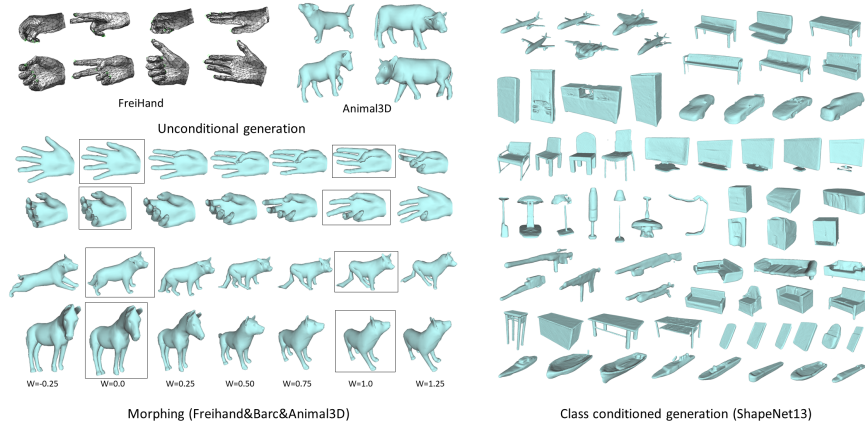
FreiHand   Animal3D

Unconditional generation

W=-0.25   W=0.0   W=0.25   W=0.50   W=0.75   W=1.0   W=1.25

Morphing (Freihand&Barc&Animal3D)   Class conditioned generation (ShapeNet13)

Fig. 5: Left: Unconditional generation and morphing of hands, dogs and animals. Right: class conditioned generation of man-made objects.

and LSPET [31]. For training, we utilized the 3D joint labels from Human 3.6M and 2D keypoint labels from all the dataset. Additionally, we conduct another training experiment using 3D body joint labels obtained from pseudo 3D meshes produced by EFT [32] on the in-the-wild image datasets. Unlike recent mesh transformer approaches [15,43], we did not use 3DPW [54] as training data for fine-tuning on 3DPW, but instead only performed evaluations on its test set.

We used the following three standard metrics for evaluation: MPJPE, PA-MPJPE and MPVE. Mean-Per-Joint-Position-Error (MPJPE) measures the Euclidean distances between the ground truth and the predicted joints. The PA-MPJPE metric, where PA stands for Procrustes Analysis, measures the error of the reconstruction after removing the effects of scale and rotation. Mean-Per-Vertex-Error (MPVE) measures the Euclidean distances between the ground truth and the predicted vertices.

5.2   Training and sampling

The training of DiffSurf involves two steps: training of the diffusion model and the up-samplers are done separately. We use pre-trained up-sampler models for fixed and varied topology cases (see the Appendix for the details on the network architectures and their training). Our diffusion transformer model is trained with a batch size of 256 for 400 epochs on 4 NVIDIA V100 GPUs for the SURREAL dataset, and for 200 epochs on 8 NVIDIA A100 GPUs for the AMASS dataset. It takes about 1 day for both cases. For the BARC, Animal3D and ShapeNet objects, we extend the training of diffusion transformer to 4000-8000 epochs because they contain fewer meshes than SURREAL and AMASS. The dataset statistics are provided in Appendix. We use the Adam optimizer for training our models, while reducing the learning rate by a factor of 10 after $1/2$ of the total training epochs beginning from 1×10^{-4}. For the training objective of DiffSurf, we adopt the v-prediction parameterization [62,68] and employ the DDIM [73] sampler along with a sigmoid variance scheduler. We set the diffusion time step to $T = 1000$ and tested sampling steps in the range [1-250].

Fig. 6: Example results of human mesh recovery by DiffSurf on the 3DPW dataset. Compared to METRO, DiffSurf produces less distorted results, especially in occluded situations.

5.3 Downstream applications

Unconditional shape generation Figures 1 (left) and 5 (left) depict unconditional 3D mesh generation results of humans, hands, dogs, mammals and man-made objects based on DiffSurf. DiffSurf trained on FreiHAND can generate 3D hand meshes in a variety of poses, including thumbs up, victory (peace), open and close. The results obtained using the BARC and Animal3D dataset indicate that DiffSurf can handle a range of dog breeds, from small to large, as well as different species. Our approach can achieve class-conditioned generation of ShapeNet 13 objects by inputting class labels to the transformer as in U-ViT, which includes generation of topologically different shapes such as the lamp examples. These results demonstrate the ability of DiffSurf to generate 3D meshes in diverse shapes and poses.

Pose conditioned generation and body shape variations Figure 1 and the Appendix demonstrate the outcomes of pose-conditional mesh generation and body shape variation using DiffSurf. When provided with different 3D joint locations while sampling from the same mesh noise input, DiffSurf can generate various poses of the same body mesh. Variations in body shape are realized by altering the mesh noise input.

Shape morphing In Figs. 1, 5 (left), 7 (right) and the Appendix, we present morphing results produced by DiffSurf, where two meshes are interpolated in the Gaussian noise space. In contrast to linear interpolation of 3D vertex coordinates in the 3D Euclidean space, which results in a straight-line interpolation trajectory leading to the artifacts such as arm shrinkage and hand expansions, DiffSurf yields visually plausible interpolation outcomes (see Fig. 7 right). Since LIMP is an approach that learns from a small amount of meshes (in static poses), its pose representation capability is limited. As shown in Fig. 5 (left), this approach can morph between two objects with different shapes, such as horse and dog, which showcases the ability of DiffSurf for handling body shapes and poses together and its potential for becoming a viable alternative to prior nonlinear and data-driven morphing techniques [22,23].

Control point deformation Figure 4 (right) illustrates the results of control point deformation. In these examples, five control points are specified at the head, wrists and ankles and the deformation begins with the rest pose. Through gradient-based optimization and progressively decreasing diffusion time steps, the mesh is refined to a pose that conforms to the control points without distortions.

Human mesh recovery from image Figure 6 visualizes the mesh recovery results on the 3DPW dataset obtained by DiffSurf. Even though DiffSurf is trained without image-mesh paired data, it produces visually pleasing results without noticeable artifacts.
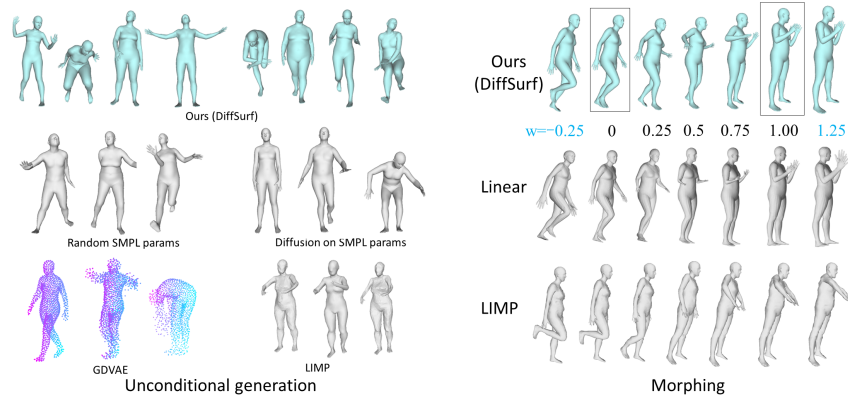
Fig. 7: Qualitative comparisons with previous techniques are presented. Left: comparison of unconditional generation results. Right: comparison of morphing results against linear interpolation in 3D Euclidean space and LIMP.

Table 1: Comparisons with baseline models on unconditional human generation. The RA-1-NNA metric [%] assesses the diversity and quality of generated results. A lower value on this metric signifies superior performance. SD indicates the standard deviation.

	SURREAL	DFAUST
Random SMPL (SD×0.2)	81.1	92.8
Random SMPL (SD×1.0)	67.2	71.4
Random SMPL (SD×3.0)	71.6	95.4
VPoser [58]	60.7	70.2
Parametric Diffusion	59.6	76.2

*Trained on AMASS

Test / Train		SURREAL	DFAUST
GDVAE [3]	SURREAL	93.8	98.1
LIMP [18]	FAUST	81.3	93.3
DiffSurf	SURREAL	54.4	69.6
DiffSurf	AMASS	**54.0**	**69.5**

5.4 Comparisons

Unconditional human mesh generation Here, DiffSurf is compared against five baselines. We employ parametric baseline approaches: Random SMPL which draws SMPL body shape/pose parameters randomly from $\times[0.2, 1.0, 3.0]$ the standard deviations of AMASS parameter collections to generate human body meshes; VPoser [58] which learns pose priors with VAEs; parametric diffusion transformer that generates SMPL parameters. We also compared DiffSurf with the previous generative models for surfaces based on VAEs: GDVAE [3] and LIMP [18].

Figure 7 presents a qualitative comparison, while Table 1 provides quantitative comparisons using the RA-1-NNA metric. As shown in Table 1, naively producing body meshes from random SMPL parameters proved to be far inferior to our approach. Diff-Surf also outperforms VPoser and parametric diffusion, which use rotational parametrization of pose that is usually difficult to learn with neural networks. Since GDVAE relies on a point cloud representation, the results exhibit outliers especially around hands and feet (see Fig. 7). LIMP is based on mesh representation and preserves mesh structure by maintaining both extrinsic and intrinsic surface properties. However, LIMP's diversity in body poses and shapes appears to be constrained, likely due to its training strategy relying on a limited dataset. As both methods are based on MLP-based VAEs, their generated sample quality is not as high as that produced by DiffSurf.

Human mesh recovery from image Table 2 presents a comparison of our method with previous human mesh recovery approaches, which are divided into parametric and

vertex-based categories, on the 3DPW and Human3.6M datasets. Note that none of methods used the 3DPW dataset in training. DiffSurf achieves top-level performances among vertex-based approaches. Our method is also comparable to recent diffusion based methods [14, 42], even though ours is not trained end-to-end on image-mesh paired dataset. Since DiffSurf does not explicitly relate its generation to an image, performance is affected by random input mesh noise ϵ_θ^x that possibly alters body styles and twisting joint angles of generations. We show in the Appendix how multiple hypotheses on the input mesh noise can further possibly improve DiffSurf's performance.

3D Human mesh fitting to 2D keypoints Table 3 shows a comparison of optimization approaches that fits a mesh with ground truth (GT) 2D keypoints. With our SDS-based approach, the MPVPE and PA-MPJPE errors for both parametric and vertex-based mesh recovery techniques decrease, outperforming the previous fitting approaches that utilized GT keypoints [8, 32, 39, 74]. In fact, PA-MPJPE dropped by approximately 2pts and 9pts for HMR-EFT [32] and METRO [43], respectively. These results suggest that DiffSurf can potentially aid in the creation of an image-mesh paired dataset with improved alignment between images and meshes.

Table 2: Comparisons with other 3D human mesh recovery approaches on 3DPW. No fine-tuning on 3DPW performed.

	Method	3DPW		Human 3.6M	
		MPVE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
Parametric	SPIN [37]	116.4	59.2	62.5	41.1
	ProHMR [39]	—	59.8	—	41.2
	OCHHuman [35]	107.1	58.3	—	—
	DiffHMR [49]	110.9	56.5	—	—
	HMR-EFT [32]	—	54.3	—	—
	PARE [36]	**97.9**	**50.9**	76.8	50.6
Vertex-based	METRO [43]	119.1	63.0	54.0	36.7
	Pose2Mesh [16]	106.3	58.3	64.9	46.3
	GATOR [87]	104.5	56.8	64.0	44.7
	HMDiff [42]	—	—	49.3	**32.4**
	DiffSurf	108.0	53.7	**48.9**	36.1
	DiffSurf-EFT	102.6	52.6	50.1	36.9

Table 3: Comparisons with previous fitting approaches on 3DPW. GT 2D keypoints are used.

Method	MPVE ↓	PA-MPJPE ↓
SMPLify [8]	—	106.1
LearnedGD [74]	—	55.9
ProHMR + fitting [39]	—	55.1
HMR-EFT + EFT fitting [32]	—	53.7
HMR-EFT [32] + SDS fitting	98.6	52.1
METRO [43] + SDS fitting	103.0	54.2
DiffSurf + SDS fitting	**93.7**	**48.7**

5.5 Ablation studies

Network architectures To conduct the ablation study on our diffusion transformer model's components, we modified the elements within it and compared their performances. The basic network architecture from which we started is an adaptation of U-ViT [4], tailored to handle 3D mesh and body joint tokens (see Table 4a, top row). We investigated the impacts of long skip connections, different methods of incorporating time embedding, position embedding constructions and the effect of varying the network layer types (transformer or MLPs). As shown in Table 4a, the switch in the layer type from transformer to MLPs yields the most significant impact, indicating that the most critical component of DiffSurf is the transformer layer. As opposed to U-ViT [4] for image generation, the use of long skip connection does not contribute to improving 3D human mesh generation. This may be attributed to the fact that the current form of the diffusion transformer in DiffSurf primarily processing coarse-level mesh vertices. Expressive human body generation that incorporates fine-grained details like finger poses and facial expressions could potentially benefit from long-skip connections.

Table 4: Ablation studies: (a) Network components. Errors are measured for unconditional human generation using the RA-1-NNA metric on the SURREAL dataset; (b) Sampling time steps. Errors are measured for 3D human mesh recovery on the 3DPW dataset with a CFG weight $s_g = 1.0$; (c) CFG scale factor s_g. The error is measured by PA-MPJPE↓ on the 3DPW and H3.6M datasets with 10 DDIM sampling time steps.

(a) Ablation study on network components.

Layer	Pos emb	Time emb	Long skip	1NNA ↓
Transformer	Learned	Token	Yes	55.6
Transformer	Learned	Token	**No**	**54.4**
Transformer	Learned	**Add**	Yes	55.4
Transformer	**3D Template**	Token	Yes	54.8
MLP	Learned	Token	Yes	92.6

(b) Ablation study on sampling time steps.

	1	3	5	10	20	30	100
MPVE ↓	230.4	115.8	107.3	**105.4**	105.8	106.1	106.9
PA-MPJPE ↓	143.9	60.2	55.6	54.3	**54.1**	54.2	54.7
fps	35.9	32.05	25.25	21.2	13.6	8.81	3.38

(c) Ablation study on CFG scale factor.

	$s_g = 0.0$	0.1	0.3	0.5	1.0
3DPW	**52.6**	52.6	52.9	53.3	54.3
H3.6M	37.9	37.0	36.2	**36.1**	36.7

Sampling timesteps Table 4b presents an ablation study on the sampling timesteps for 3D human mesh recovery from an image. It is observed that an increase in the sampling timesteps leads to a decrease in MPVPE and PA-MPJPE, reaching optimal performance at approximately 10-20 steps. We also measured the inference time with respect to sampling timesteps. In this configuration, DiffSurf operates at nearly real-time speed, approximately 20 FPS, when performing DDIM sampling with 10 steps. Furtheremore, the sampling speed of unconditional human generation with 431 vertices were 37, 9 and 1.5 fps for 10, 50 and 250 DDIM sampling steps, respectively. The sampling speed of man-made objects generation (2048 points) were 17.8, 3.5 and 1.8 fps for 10, 50 and 100 DDIM sampling steps, respectively, which is roughly 7× faster than LION [91] with DDIM sampling. Our experiments were conducted on an NVIDIA A100 GPU with the Flash Attention layer enabled [19].

CFG scale factor Table 4c presents the ablation study on the CFG scaling factor s_g for 3D human mesh recovery from an image. For Human 3.6M, a value up to around 1 leads to improvements. On the other hand, we observed that increasing the CFG factor negatively impacts mesh reconstruction on the 3DPW dataset. This discrepancy is likely due to the difference in the accuracy of 3D pose regressors and the frequency of occlusions associated with each dataset. In general, the 3D joint estimation results on Human3.6M are more accurate and reliable than those on 3DPW. This is because the Human3.6M dataset is captured in a controlled experimental environment and is included in the training, whereas 3DPW is an in-the-wild dataset and not used as the training data. DiffSurf provides a method to consider the accuracy and reliability of the 3D joint prediction by balancing between diffusion model priors and 3D body joint conditions through the CFG scaling factor.

6   Conclusion

We presented DiffSurf, a denoising diffusion transformer model for generating 3D surfaces in diverse body shapes and poses. DiffSurf can generate 3D surfaces for a wide range of object types and solve various downstream 3D processing tasks. In future work, we aim to extend DiffSurf towards the generation of expressive human body meshes with fine-grained details, such as facial expressions and finger poses. It would also be intriguing to design a foundational 3D generative model by increasing the capacity of DiffSurf and training it on a larger-scale 3D data.

Acknowledgements

References

1. Alliegro, A., Siddiqui, Y., Tommasi, T., Nießner, M.: Polydiff: Generating 3d polygonal meshes with diffusion models (2023) 3

2. Andriluka, M., Pishchulin, L., Gehler, P., Bernt, S.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014) 9

3. Aumentado-Armstrong, T., Tsogkas, S., Jepson, A., Dickinson, S.: Geometric disentanglement for generative latent shape models. In: ICCV. pp. 8180–8189 (2019) 2, 12

4. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: CVPR (2023) 3, 13

5. Bao, F., Nie, S., Xue, K., Li, C., Pu, S., Wang, Y., Yue, G., Cao, Y., Su, H., Zhu, J.: One transformer fits all distributions in multi-modal diffusion at scale (2023) 3, 4

6. Bautista, M.A., Guo, P., Abnar, S., Talbott, W., Toshev, A., Chen, Z., Dinh, L., Zhai, S., Goh, H., Ulbricht, D., Dehghan, A., Susskind, J.: Gaudi: A neural architect for immersive 3d scene generation. arXiv (2022) 1

7. Biggs, B., Ehrhart, S., Joo, H., Graham, B., Vedaldi, A., Novotny, D.: 3D multibodies: Fitting sets of plausible 3D models to ambiguous image data. In: NeurIPS (2020) 2, 3

8. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: ECCV. pp. 561–578. Springer (2016) 3, 8, 13

9. Bogo, F., Romero, J., Pons-Moll, G., Black, M.J.: Dynamic FAUST: Registering human bodies in motion. In: CVPR (Jul 2017) 9

10. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository (2015) 9

11. Chen, H., Tang, H., Shi, H., Peng, W., Sebe, N., Zhao, G.: Intrinsic-extrinsic preserved gans for unsupervised 3d pose transfer. In: ICCV. pp. 8610–8619 (2021) 3

12. Cheng, S., Bronstein, M.M., Zhou, Y., Kotsia, I., Pantic, M., Zafeiriou, S.: Meshgan: Nonlinear 3d morphable models of faces. CoRR **abs/1903.10384** (2019) 3

13. Cheng, Y.C., Lee, H.Y., Tulyakov, S., Schwing, A.G., Gui, L.Y.: SDFusion: Multimodal 3d shape completion, reconstruction, and generation. In: CVPR. pp. 4456–4465 (2023) 1, 3, 6

14. Cho, H., Kim, J.: Generative approach for probabilistic human mesh recovery using diffusion models (2023) 2, 3, 13

15. Cho, J., Youwang, K., Oh, T.H.: Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In: ECCV (2022) 3, 5, 6, 10

16. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: ECCV (2020) 3, 13

17. Christian, Z., Duygu, C., Jimei, Y., Russel, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: ICCV (2019) 9

18. Cosmo, L., Norelli, A., Halimi, O., Kimmel, R., Rodolà, E.: Limp: Learning latent shape representations with metric preservation priors. In: ECCV. p. 19–35 (2020) 2, 12

19. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In: NeurIPS (2022) 14

20. Davydov, A., Remizova, A., Constantin, V., Honari, S., Salzmann, M., Fua, P.: Adversarial parametric pose prior. In: CVPR. pp. 10987–10995 (jun 2022) 2, 3
21. Desbrun, M., Meyer, M., Schröder, P., Barr, A.H.: Implicit fairing of irregular meshes using diffusion and curvature flow. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. p. 317–324. SIGGRAPH '99 (1999) 7
22. Fröhlich, S., Botsch, M.: Example-driven deformations based on discrete shells. Computer Graphics Forum **30**(8), 2246–2257 (2011) 8, 11
23. Gao, L., Lai, Y.K., Yang, J., Zhang, L.X., Xia, S., Kobbelt, L.: Sparse data driven mesh deformation. IEEE Transactions on Visualization and Computer Graphics **27**(3), 2085–2100 (2021) 8, 11
24. Georgakis, G., Li, R., Karanam, S., Chen, T., Košecká, J., Wu, Z.: Hierarchical kinematic human mesh recovery. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) ECCV. pp. 768–784. Springer International Publishing (2020) 2
25. Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. In: CVPR (June 2023) 1, 2, 3
26. Groueix, T., Fisher, M., Kim, V.G., Russell, B., Aubry, M.: 3d-coded : 3d correspondences by deep deformation. In: ECCV (2018) 9
27. Ho, J.: Classifier-free diffusion guidance. ArXiv **abs/2207.12598** (2022) 4
28. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv preprint arxiv:2006.11239 (2020) 1, 3
29. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE TPAMI **36**(7), 1325–1339 (2014) 9
30. Jiang, B., Zhang, J., Cai, J., Zheng, J.: Disentangled human body embedding based on deep hierarchical neural network (2020) 2
31. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR. pp. 1465–1472 (2011) 10
32. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In: 3DV (2020) 10, 13
33. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions (2023) 3
34. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018) 2, 3
35. Khirodkar, R., Tripathi, S., Kitani, K.: Occluded human mesh recovery. In: CVPR. pp. 1715–1725 (June 2022) 13
36. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3D human body estimation. In: ICCV. pp. 11127–11137 (2021) 9, 13
37. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019) 13
38. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019) 3
39. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: ICCV (2021) 3, 8, 13
40. Li, L., Zhuo, L., Zhang, B., Bo, L., Chen, C.: Diffhand: End-to-end hand mesh reconstruction via diffusion models (2023) 2, 3
41. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR (2023) 1
42. Lin, G.F., Jia, G., Hossein, R., Jun, L.: Distribution-aligned diffusion for human mesh recovery. In: ICCV (2023) 2, 3, 13
43. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: CVPR (2021) 3, 5, 6, 7, 9, 10, 13

44. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: ICCV (2021) 3, 5, 6, 7

45. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014) 9

46. Liu, Y., Yang, J., Gu, X., Guo, Y., Yang, G.Z.: Egohmr: Egocentric human mesh recovery via hierarchical latent diffusion model. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 9807–9813 (2023) 3

47. Liu, Z., Feng, Y., Black, M.J., Nowrouzezahrai, D., Paull, L., Liu, W.: Meshdiffusion: Score-based generative 3d mesh modeling. In: International Conference on Learning Representations (2023) 3

48. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM TOG **34**(6), 248:1–248:16 (2015) 8, 9

49. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: CVPR (June 2021) 1, 3, 6, 13

50. Lyu, Z., Kong, Z., Xu, X., Pan, L., Lin, D.: A conditional point diffusion-refinement paradigm for 3d point cloud completion (2022) 7

51. Lyu, Z., Wang, J., An, Y., Zhang, Y., Lin, D., Dai, B.: Controllable mesh generation through sparse latent point diffusion models. CVPR (2023) 1, 3, 6, 7, 9

52. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3d people in generative clothing. In: CVPR (Jun 2020) 2

53. Mahmood, N., Ghorbani, N., F. Troje, N., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: ICCV (2019) 9

54. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018) 10

55. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3DV. IEEE (2017) 9

56. Mo, S., Xie, E., Chu, R., Hong, L., Nießner, M., Li, Z.: Dit-3d: Exploring plain diffusion transformers for 3d shape generation. arXiv preprint arXiv: 2307.01831 (2023) 3, 6

57. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts (2022) 3

58. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019) 2, 12

59. Peebles, W., Radosavovic, I., Brooks, T., Efros, A., Malik, J.: Learning to learn with generative models of neural network checkpoints. arXiv preprint arXiv:2.12892 (2022) 3

60. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022) 3

61. Peng, S., Jiang, C.M., Liao, Y., Niemeyer, M., Pollefeys, M., Geiger, A.: Shape as points: A differentiable poisson solver. In: NeurIPS (2021) 3, 6, 7

62. Phil, W.: denoising-diffusion-pytorch. https://github.com/lucidrains/denoising-diffusion-pytorch (2023) 10

63. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv (2022) 1, 5, 8

64. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3D faces using convolutional mesh autoencoders. In: ECCV. pp. 725–741 (2018) 2

65. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (June 2022) 1, 3

66. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM TOG **36**(6) (Nov 2017) 8

67. Rueegg, N., Zuffi, S., Schindler, K., Black, M.J.: Barc: Learning to regress 3d dog shape from images by exploiting breed information. In: CVPR (2022) 3, 9
68. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models (2022) 10
69. Shan, W., Liu, Z., Zhang, X., Wang, Z., Han, K., Wang, S., Ma, S., Gao, W.: Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. arXiv preprint arXiv:2303.11579 (2023) 3
70. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2021) 3
71. Shim, J., Kang, C., Joo, K.: Diffusion-based signed distance fields for 3d shape generation. In: CVPR. pp. 20887–20897 (2023) 1, 3
72. Shoemake, K.: Animating rotation with quaternion curves. In: Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques. p. 245–254. SIGGRAPH '85, Association for Computing Machinery, New York, NY, USA (1985) 7
73. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv:2010.02502 (October 2020) 10
74. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) ECCV. pp. 744–760 (2020) 13
75. Sumner, R.W., Zwicker, M., Gotsman, C., Popovic, J.: Mesh-based inverse kinematics. ACM TOG **24**(3), 488–495 (2005) 8
76. Sun, X., Feng, Q., Li, X., Zhang, J., Lai, Y.K., Yang, J., Li, K.: Learning semantic-aware disentangled representation for flexible 3d human body editing. In: CVPR (2023) 2
77. Tan, Q., Gao, L., Lai, Y.K., Xia, S.: Variational autoencoders for deforming 3d mesh models. In: CVPR. pp. 5841–5850 (2018) 2
78. Tian, Y., Zhang, H., Liu, Y., Wang, L.: Recovering 3d human mesh from monocular images: A survey. arXiv preprint arXiv:2203.01923 (2022) 3
79. Tiwari, G., Antic, D., Lenssen, J.E., Sarafianos, N., Tung, T., Pons-Moll, G.: Pose-ndf: Modeling human pose manifolds with neural distance fields. In: ECCV (October 2022) 3
80. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: CVPR (2017) 9
81. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023) 1
82. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: CVPR. pp. 6184–6193 (2020) 2, 3
83. Xu, J., Zhang, Y., Peng, J., Ma, W., Jesslen, A., Ji, P., Hu, Q., Zhang, J., Liu, Q., Wang, J., et al.: Animal3d: A comprehensive dataset of 3d animal pose and shape. arXiv preprint arXiv:2308.11737 (2023) 3, 9
84. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. In: ICCV. pp. 7754–7765 (October 2023) 6
85. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: Pointflow: 3d point cloud generation with continuous normalizing flows. arXiv (2019) 9
86. Yoshiyasu, Y.: Deformable mesh transformer for 3d human mesh recovery. In: CVPR. pp. 17006–17015 (2023) 3, 7
87. You, Y., Liu, H., Li, X., Li, W., Wang, T., Ding, R.: Gator: Graph-aware transformer with motion-disentangled regression for human mesh recovery from a 2d pose. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023) 3, 13

88. Yu, Z., Dou, Z., Long, X., Lin, C., Li, Z., Liu, Y., Müller, N., Komura, T., Habermann, M., Theobalt, C., et al.: Surf-d: High-quality surface generation for arbitrary topologies using diffusion models. arXiv preprint arXiv:2311.17050 (2023) 3

89. Yuan, Y.J., Lai, Y.K., Yang, J., Duan, Q., Fu, H., Gao, L.: Mesh variational autoencoders with edge contraction pooling. In: CVPRW. pp. 274–275 (2020) 2

90. Zanfir, A., Bazavan, E.G., Xu, H., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In: ECCV. pp. 465–481 (2020) 2, 3

91. Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K.: Lion: Latent point diffusion models for 3d shape generation. In: Advances in Neural Information Processing Systems (NeurIPS) (2022) 1, 3, 6, 14

92. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: ICCV (2021) 3

93. Zhou, K., Bhatnagar, B.L., Pons-Moll, G.: Unsupervised shape and pose disentanglement for 3d meshes. In: ECCV (August 2020) 2

94. Zuffi, S., Kanazawa, A., Jacobs, D., Black, M.J.: 3D menagerie: Modeling the 3D shape and pose of animals. In: CVPR (Jul 2017) 3, 8, 9