Motion2VecSets: Non-Rigid Shape Reconstruction and Tracking with 4D Latent Set Diffusion

Jiapeng Tang, Wei Cao, Biao Zhang, Chang Luo, Yaoyao Liu, Matthias Nießner

Abstract—We introduce Motion2VecSets, a 4D diffusion model for dynamic surface mesh generation from various ambiguous observations, including a sequence of RGB images, sparse and partial point clouds, and low-resolution voxel grids. While recent methods using neural field representations have shown success in modeling non-rigid objects, conventional feed-forward architectures struggle with noisy, partial, or sparse observations due to their deterministic nature. To address the inherent one-to-many mapping problem, we introduce a diffusion model that explicitly learns the shape and motion distribution of non-rigid objects through an iterative denoising process of compressed latent representations. The diffusion-based priors provide more plausible and diverse reconstructions under ambiguous conditions. Instead of relying on global latent codes, we represent 4D dynamics using latent sets. This novel 4D representation captures local shape and deformation patterns, leading to more accurate non-linear motion capture and significantly improving generalization capacity to unseen motions and identities. For temporally coherent tracking, we jointly denoise latent sets across frames and enable cross-frame information exchange. To reduce computational cost, we design an interleaved spatial-temporal attention block that alternately aggregates deformation latents along spatial and temporal dimensions. Extensive experiments on datasets of humans, animals, and articulated objects demonstrate that Motion2VecSets outperforms prior methods in reconstructing and tracking non-rigid deformations from various imperfect observations. Our implementation is available at https://vveicao.github.io/projects/Motion2VecSets/.

Index Terms—Diffusion Models, Non-Rigid Object Reconstruction and Tracking, 3D and 4D Surface Generation, Mesh Deformation.

1 Introduction

Reconstructing dynamic object surfaces and motions from diverse observations, such as point clouds, voxel grids, and images, is a core research area of computer vision and computer graphics. It plays a vital role in practical applications like virtual and augmented reality, computer games, movie effects, and robotic manipulation. Classic methods for 3D mesh reconstruction and generation trace back to foundational techniques such as Marching Cubes [1], Poisson Surface Reconstruction [2], KinectFusion [3], and Multi-view Stereo [4]. For non-rigid object reconstruction and tracking, approaches like DynamicFusion [5], VolumeDeform [6], and DoubleFusion [7] jointly perform surface fusion and motion tracking by leveraging handcrafted deformation priors, such as As-Rigid-As-Possible (ARAP) [8] and Embedded Deformation [9]. While these methods are effective for capturing short-term motions, they often struggle to handle complex or highly non-linear deformations commonly encountered in real-world scenarios.

Recently, there have been notable advances in learning-based 3D and 4D object modeling. Early efforts employed parametric models [10], [11], [12], [13], [14] tailored to specific object categories. However, their reliance on a fixed mesh topology limits their ability to capture the complex 4D dynamics of general non-rigid objects. Model-free methods [15], [16], [17] overcome this constraint by using coordinate-based MLPs [18] to model deformations with an arbitrary topology and an unstructured geometry, showing promising results for large non-rigid motions. Despite these advances, current methods still face challenges under ambiguous input conditions, such as noisy, sparse, or partial observations, where the reconstruction problem becomes ill-posed due to multiple plausible solutions. Moreover, they represent dynamics as a sequence of single latent codes and thus struggle to capture shape and motion priors accurately. These issues become even more severe with unseen identities, due to the limited generalization capacity of the global latent representation.

To address the aforementioned challenges, we propose Motion2VecSets, a 4D diffusion model designed for dynamic surface mesh reconstruction from various ambiguous observations, including sparse, noisy, or partial point clouds, RGB images, and coarse voxel grids. It explicitly learns a probabilistic distribution of non-rigid surface geometry and temporal dynamics via an iterative denoising process, enabling more realistic and diverse reconstructions, especially in the presence of uncertain inputs. Learning such a 4D diffusion requires a compact yet expressive representation to encode the underlying shape and motion. Inspired by the observation that objects with diverse topologies often exhibit similar local geometry and deformation patterns, we represent dynamic surfaces as a sequence of latent sets:

- *Jiapeng Tang, Chang Luo, and Matthias Nießner are with the Technical University of Munich, Germany. Jiapeng Tang and Matthias Nießner are also with the Munich Center for Machine Learning (MCML).*
- *Biao Zhang is with King Abdullah University of Science and Technology.*
- *Wei Cao and Yaoyao Liu are with the University of Illinois Urbana-Champaign.*
- *The first four authors contributed equally to this work.*
- *Jiapeng Tang (jiapeng.tang@tum.de) and Wei Cao (weicao3@illinois.edu) are the corresponding authors.*

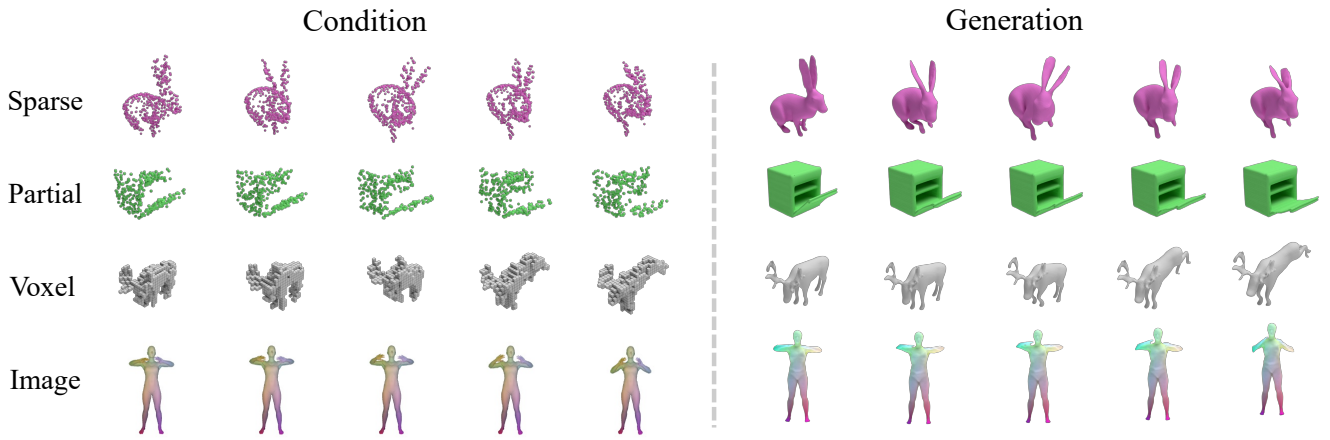


Fig. 1: We present *Motion2VecSets*, a 4D diffusion model capable of reconstructing dynamic surface meshes with complex geometries and robust motion tracking from ambiguous 4D observations, including sequences of sparse and partial point clouds, coarse voxel grids, and RGB images.

one for capturing the geometry of the initial frame and the others for modeling its temporal evolution. This design not only preserves local shape and motion details but also improves the generalization capacity to unseen identities and motions. We employ a shape latent set diffusion model to reconstruct the 3D surface mesh at the initial frame. For non-rigid deformation tracking, we introduce a synchronized deformation latent set diffusion, which jointly denoises deformation latent sets across all time frames. This enforces spatio-temporal consistency and enables coherent surface tracking throughout the sequence. To address the high memory cost associated with simultaneous deformation diffusion across time, we propose an interleaved space-time attention block as the core unit of the denoiser. This module alternates between aggregating latent features along the spatial and temporal dimensions, achieving efficient yet expressive modeling. As illustrated in Fig. 1, our Motion2VecSets can reconstruct plausible, high-fidelity non-rigid surfaces with complex structures and exhibits robust motion tracking performance under a variety of ambiguous input settings.

TABLE 1: Comparison of different 3D and 4D reconstruction methods.

Method	#Latents	Model-Free	4D Recon.	Inputs			Probabilistic Outputs
				Points	Images	Voxels	
SMPL [10]	–	✗	✓	–	–	–	–
MANO [12]	–	✗	✓	–	–	–	–
DeepSDF [19]	Single	✓	✗	✓	✗	✗	✗
OccNet [20]	Single	✓	✗	✓	✓	✓	✓
ConvOccNet [21]	Multiple	✓	✗	✓	✗	✓	✓
3DShape2VecSet [22]	Multiple	✓	✗	✓	✓	✗	✓
Oflow [15]	Single	✓	✓	✓	✓	✗	✓
LPDC [16]	Single	✓	✓	✓	✗	✗	✗
CaDex [17]	Single	✓	✓	✓	✗	✗	✗
DNF [23]	Single	✓	✓	✗	✗	✗	✗
Ours	Multiple	✓	✓	✓	✓	✓	✓

A preliminary version of this work was presented at CVPR 2024, where we introduced Motion2VecSets [24] for dynamic shape reconstruction from point cloud sequences. In this extended version, we expand both the method and experimental scope. Motion2VecSets now supports 4D mesh reconstruction from a wide range of imperfect 4D observations, including noisy, sparse, or partial point clouds, coarse voxel grids, and RGB images. A comparison of 3D and 4D reconstruction methods under different input settings is provided in Table 1. We expand our experiments to show that our Motion2VecSets can handle a broader range of non-rigid objects, including humanoids, animals, and articulated objects. We further broaden our evaluation to demonstrate the generality of our framework across diverse non-rigid object categories, including humans, animals, and articulated objects. Additionally, we show that Motion2VecSets can synthesize natural surface motions when conditioned on sparse 3D handle trajectories. Finally, we conduct robustness experiments under varying levels of noise and sparsity in point cloud inputs, demonstrating that our method consistently outperforms state-of-the-art approaches. Our contributions can be summarized as follows:

- We propose a novel **4D latent diffusion model** for dynamic surface mesh generation, enabling realistic and consistent shape deformation over time.
- We introduce a novel **4D neural representation based on latent sets**, coupled with transformer architectures. This representation improves the capacity to model complex geometry and motion, and enhances generalization to unseen identities and dynamics.
- We design an **Interleaved Spatio-Temporal Attention** mechanism to denoise multi-frame bundled deformation latent sets. This ensures coherent spatio-temporal structure across frames while maintaining high computational efficiency.
- We demonstrate the effectiveness of our method in **non-rigid shape reconstruction and tracking** from diverse and imperfect 4D observations, including a sequence of sparse and partial point clouds, RGB images, and low-resolution voxel grids.

Extensive comparisons with state-of-the-art methods show that Motion2VecSets achieves superior performance in dynamic surface reconstruction across various object categories, including humanoids, animals, and articulated shapes. It outperforms prior methods on commonly used benchmarks such as Dynamic FAUST [25], DeformingThings4D-Animals [26], and Shape2Motion [27].

2 Related Works

In this section, we review these closely related works, including 3D shape representation and reconstruction, non-

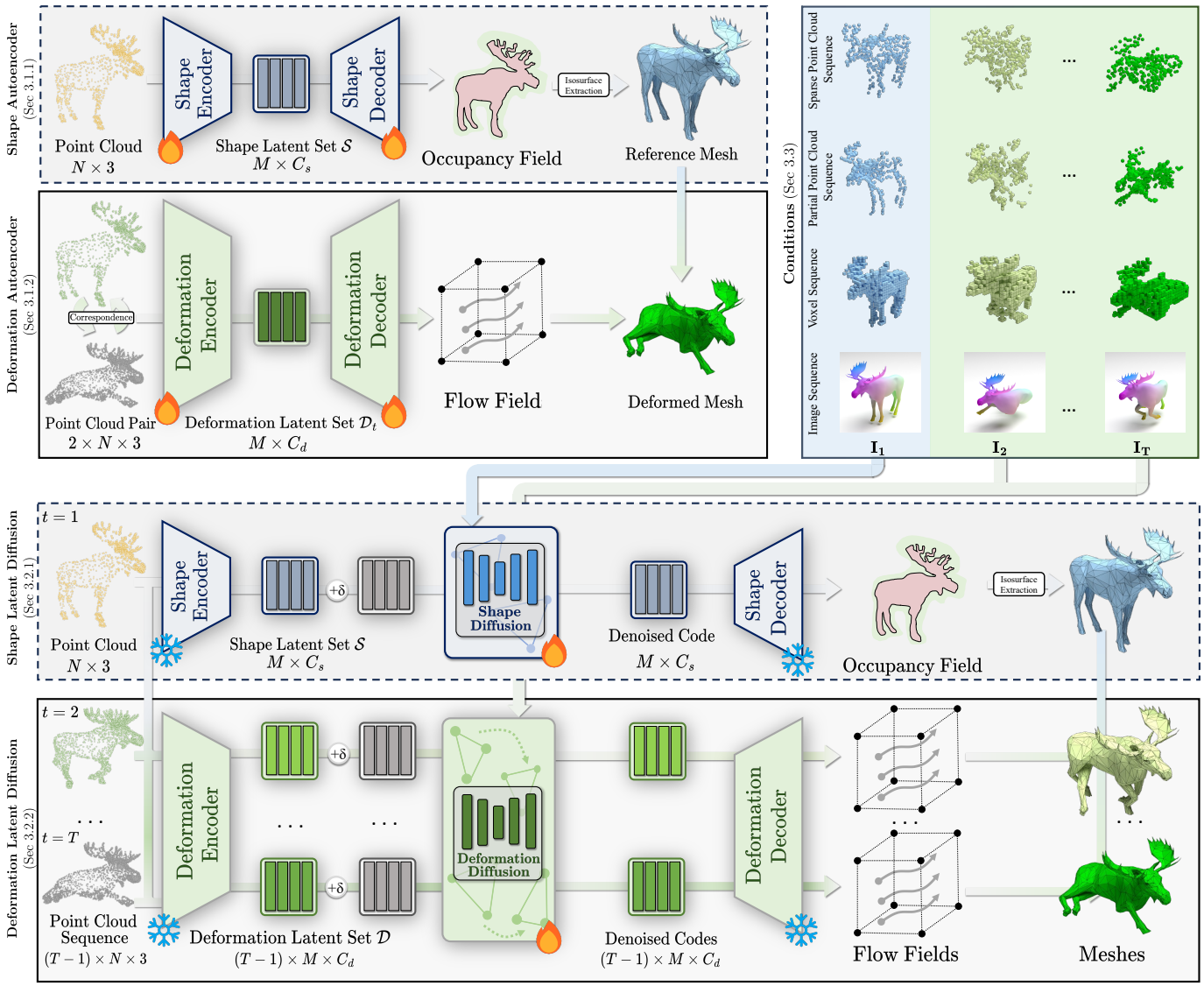


Fig. 2: **Overview Pipeline of Motion2VecSets.** Our method consists of two stages: 4D latent set representation learning and 4D latent set diffusion. In the first stage, the *Shape Autoencoder* encodes the first reference frame into a shape latent set $\mathcal{S} \in \mathbb{R}^{M \times C_s}$ The *Deformation Autoencoder* encodes each pair of point clouds formed between the first and a subsequent frame into a deformation latent set $\mathcal{D}^t \in \mathbb{R}^{M \times C_d}$. Stacking all $\mathcal{D}^t$ across time yields a motion latent tensor $\mathcal{D} \in \mathbb{R}^{(T-1) \times M \times C_d}$. In the second stage, given a sequence of multi-modal observations $\{\mathbf{I}_t\}_{t=1}^{T}$ (e.g., images, voxels, or point clouds), we use modality-specific encoders to extract conditioning embeddings $\mathcal{C}$. Conditioning on $\mathcal{C}_1$, the *Shape Latent Diffusion* denoises a noisy shape latent to reconstruct the occupancy field of the reference frame via the frozen shape decoder. A surface mesh can be obtained via iso-surface extraction. Conditioning on $\mathcal{C}_2, \ldots, \mathcal{C}_T$, the *Synchronized Deformation Latent Diffusion* jointly denoises the deformation latents for all subsequent frames. The denoised latents are then decoded by the frozen deformation decoder into flow fields, which deform the reference mesh over time.

rigid deformation and tracking, diffusion models, and 3D/4D shape generation.

2.1 3D Shape Representation and Reconstruction

Over the past decade, learning-based 3D reconstruction methods have explored a variety of shape representations, including voxels [28], [29], [30], [31], point clouds [32], [33], [34], octrees [35], [36], [37], and meshes [38], [39], [40], [41], [42], [43]. Meshes are the most natural and compact representation of 3D surfaces and are widely used in downstream applications such as rendering, simulation, and animation. However, learning to directly predict high-quality mesh topology and geometry remains challenging due to their irregular structure and sensitivity to discretization artifacts. Alongside general-purpose 3D reconstruction methods, parametric models have proven effective for modeling specific shape categories, such as the human body (e.g., SMPL [10], STAR [11]), face (e.g., FLAME [13]), hand (e.g., MANO [12]), and animal (e.g., SMAL [14]). These models offer strong performance for tasks like tracking and animation within their predefined domains. However, their reliance on fixed mesh templates limits their ability to capture large topological variations and complex non-rigid deformations observed in general object categories. To address these limitations, recent advances in neural implicit function fields [18], [44], [45] have gained widespread at-

tention. These approaches represent 3D geometry as continuous fields using coordinate-based MLPs, such as occupancy fields [18], signed distance functions (SDFs) [45], and implicit function decoders [44]. Such representations provide high flexibility in modeling arbitrary topologies and fine-grained geometric details, and can reconstruct surfaces at theoretically infinite resolution. Subsequent works have extended these ideas with improved network architectures and training strategies [46], [47], [48], [49], [50], [51], further pushing the boundaries of high-fidelity 3D reconstruction.

2.2 Non-Rigid Deformation and Tracking

Early methods for non-rigid surface tracking, such as DynamicFusion [5], DoubleFusion [7], and VolumeDeform [6], jointly perform surface fusion and motion tracking from RGB-D sequences. These approaches rely on predefined deformation priors, typically assuming local rigidity to regularize motion. Common regularization strategies include As-Rigid-As-Possible (ARAP) [8] and Embedded Deformation (ED) [9], which enable real-time performance and smooth deformation tracking under limited motion. However, such hand-crafted priors struggle to generalize to complex, large-scale, or non-linear deformations commonly found in real-world scenarios. To address these limitations, recent works have explored data-driven approaches. Notably, Neural Shape Deformation Priors (NSDP) [52] introduce learned deformation priors from training data, demonstrating improved capability in capturing highly non-linear and topology-aware deformations. Recent advancements in 3D shape representation have been effectively extended to the 4D domain, enabling a more comprehensive modeling of object dynamics over time, such as OFlow [15], NPMs [53], LCRODE [54], LPDC [16], CaDeX [17]. Despite their success, these methods rely on a single global latent code [16], [17], [53] to encode temporal shape variations, which limits their ability to capture fine-grained local deformations and generalize across diverse identities or motion patterns. In contrast, we represent recurring local geometric and dynamic patterns across different objects using a set of latent codes. Our method captures complex surfaces and motions with higher accuracy and generalizes well to unseen identities and motions.

2.3 Diffusion Models

Diffusion models [55], [56] have recently emerged as a prominent class of generative frameworks based on iterative denoising processes. They have achieved state-of-the-art results across diverse domains, including image synthesis [57], [58], [59], [60], [61], video generation [62], [63], [64], audio generation [65], [66], and text creation [67], [68], [69]. To improve scalability and reduce the computational cost associated with high-resolution synthesis, Latent Diffusion Models (LDMs) [60] encode data into compact latent spaces, enabling efficient high-resolution synthesis. For video generation, early adaptations like Video Diffusion Models [64] and Imagen Video [70] apply denoising independently per frame, often resulting in temporal inconsistency and motion artifacts, while also incurring high costs for long sequences. To address this, recent approaches introduce motion-aware mechanisms: Stable Video Diffusion [62] incorporates temporal attention in latent space, and AnimateDiff [63] leverages pre-trained motion modules to animate diffusion outputs coherently. Our synchronized deformation diffusion shares a similar motivation with recent video diffusion models. We introduce an interleaved attention mechanism that alternates between spatial and temporal dimensions, substantially reducing computational complexity and memory consumption while preserving temporal coherence.

2.4 3D and 4D Generation

Recent advancements in generative models have been extended to 3D and 4D content generation, demonstrating strong performance across a variety of tasks, including shape generation [22], [71], [72], [73], scene synthesis [74], texture and material generation [75], [76], [77], human generation [78], [79], [80], and human scene interaction [81], [82], [83], [84].

Early efforts [71], [72], [85] applied the diffusion process directly to point cloud denoising of a fixed size. Other approaches optimize 3D representations, such as NeRF [86] or 3D Gaussian Splatting [87], by leveraging pre-trained text-to-image foundation models [62] via Score Distillation Sampling (SDS) [88] and its variants. To enable feed-forward mesh generation without costly per-shape optimization, subsequent methods [22], [73], [89] compress 3D shapes into latent representations, which are then decoded into neural fields, where the denoising process is performed in latent space. LaGeM [90] and Structured 3D Latents [91] further organize latent codes hierarchically to capture structural information at multiple semantic levels. Recent large-scale models such as Direct3D [92] and Hunyuan3D [93], [94] scale latent diffusion to high-resolution textured asset generation, using transformer-based architectures and multimodal training on massive 3D datasets.

For non-rigid object generation, NAP [95] formulates it as a graph generation task, including both nodes and edges. Similarly, methods such as RigAnything [96] and MagicArticulate [97] enable explicit skeleton and skinning weight prediction from raw meshes. However, they still primarily focus on static generation, without modeling temporal consistency across deformations. A separate line of work targets human motion generation [23], [98], [99], [100]. In contrast, our focus lies in modeling detailed surface deformations rather than highly abstracted skeletal motions. Additionally, our method addresses general non-rigid objects beyond the human domain. Most relevant to our work, DNF [23] disentangles geometry and motion using a dictionary-based shared latent field and learns unconditional 4D object generation. However, unlike DNF, our approach focuses on conditional 4D reconstruction and tracking from diverse and ambiguous inputs such as sparse point clouds, coarse voxels, and RGB images. Other works, such as L4GM [101], Diffusion4D [102], and Consistent4D [103], focus on optimizing NeRF or 3D Gaussian Splatting (3DGS) representations from monocular videos. In contrast, our work emphasizes high-quality surface geometry reconstruction and robust tracking, rather than volumetric radiance modeling or image-based view synthesis.

3 Approach

The objective is to generate surface meshes with dense temporal correspondences from various imperfect inputs, including coarse voxel grids, noisy or partial point clouds, and RGB images. The reconstructed mesh sequence is denoted as $\{\mathcal{M}^t\}_{t=1}^T = \{\mathcal{V}^t, \mathcal{F}^t\}_{t=1}^T$, where $\mathcal{V}^t$ and $\mathcal{F}^t$ represent the vertices and faces at each time step t. We aim to address this problem via neural fields. An occupancy field and a series of flow fields parameterize an object mesh sequence:

$$\begin{aligned} \text{Occupancy}: &\quad \mathbb{R}^3 \to \mathbb{R}, \\ \text{Flow}_{1\to 2}: &\quad \mathbb{R}^3 \to \mathbb{R}^3, \\ &\vdots \\ \text{Flow}_{1\to T}: &\quad \mathbb{R}^3 \to \mathbb{R}^3. \end{aligned} \tag{1}$$

The occupancy field describes the implicit surface of the first frame $\mathcal{M}^1 = \{\mathcal{V}^1, \mathcal{F}^1\}$, which can be easily extracted using iso-surface extraction algorithms. The flow fields are densely defined on the surface of the initial frame and output a vector on the target shape. Subsequent mesh frames can be recovered by querying the flow field. The symbol $\text{Flow}_{1\to t}$ gives the flow field between frame 1 and frame t. They are used to deform $\mathcal{V}^1$ to $\{\mathcal{V}\}_{t=2}^T$. The face connectivity $\mathcal{F}^t$ is inherited from $\mathcal{F}^1$.

Conventional feed-forward deterministic models often struggle in this ill-posed setting, especially when the inputs are sparse, incomplete, or ambiguous, making it difficult to recover accurate dynamic geometry without strong priors. To address these challenges, we propose *4D Latent Set Diffusion*, a generative method that explicitly learns the probabilistic distribution of deformable surface sequences through both shape and motion diffusion priors. This enables sampling diverse, high-quality object surfaces and motion tracking with multiple plausible outcomes. For compact yet expressive 4D representation, we introduce a latent set-based neural representation equipped with transformer architectures. This design preserves rich geometric details while capturing complex deformations in a low-dimensional latent space.

Consistent with other latent diffusion models, our training consists of two distinct stages: autoencoders and diffusion models. We describe the shape and deformation autoencoders in Sec. 3.1 and Fig. 2 *top*. We elaborate the 4D latent set diffusion models in Sec. 3.2 and Fig. 2 *bottom*. We also discuss different condition modalities in Sec. 3.3.

3.1 4D Neural Representation with Latent Sets

Previous works often utilize single global codes [15], [16], [17] to represent 4D sequences, potentially losing significant surface geometry and temporal evolution details. Instead of a global latent, we assign local latent codes to individual local regions, which significantly improves the network's capability to accurately model non-linear motions and generalize to unseen identities and motions. Given that different non-rigid objects share similar local geometry and deformation patterns, the latent sets can also improve the generalization ability to handle unseen motions and identities. Also inspired by [22], we apply the "VecSet" representation to both the shape and deformation field learning. For a mesh sequence of T frames, the 4D latent set is defined as:

$$\underbrace{\mathbb{R}^{M\times C_s}}_{\text{shape}} \times \underbrace{\mathbb{R}^{(T-1)\times M\times C_d}}_{\text{deformation}}, \tag{2}$$

where M is the size of the latent set, and C_s and C_d are the channels of the shape and deformation latent space, respectively. The *shape latent set* is responsible for reconstructing the first frame, serving as the reference frame. The flow fields are decoded by the *deformation latent sets*, which are used to predict the dense correspondence between the initial frame and the subsequent frames.

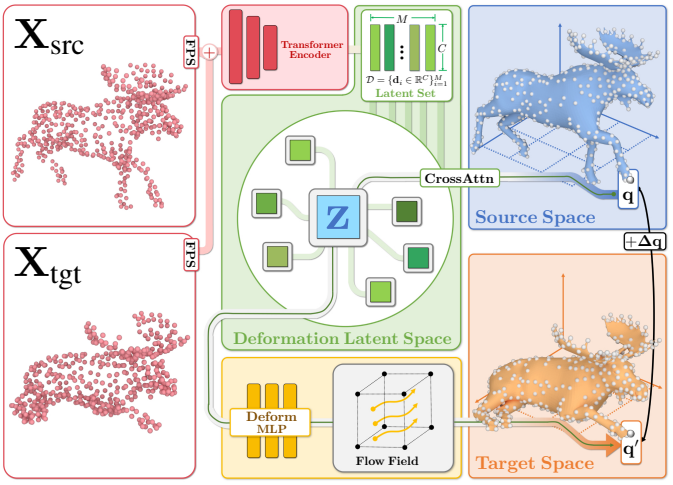


Fig. 3: **Deformation Autoencoder.** Given a pair of point clouds $\mathbf{X}_{\text{src}}$ and $\mathbf{X}_{\text{tgt}}$ from two frames of a dynamic mesh sequence, we initially downsample them using farthest point sampling (FPS). Subsequently, the concatenated points are fed into transformer encoder to generate the Deformation Latent Set $\mathcal{D}$. For a query point $\mathbf{q}$ in the source space, a cross-attention layer is utilized to retrieve the most relevant fused feature $\mathbf{z}$. This selected feature is subsequently fed into the deformation MLP decoder to predict an offset $\mathbf{\Delta q}$, which translates it to $\mathbf{q}'$ in the target space. To reduce the feature diversity of $\mathcal{D}$, KL-regularization is employed.

3.1.1 Shape Latent Set

Following the point query variant of 3DShape2VecSet [22], we compress the 3D surface of the initial frame into a set of latent codes. The input of the autoencoder is a surface point cloud $\mathbf{X}$ of size $N = 2048$. We further downsample the input to a smaller point cloud of size $M = 512$. A cross attention layer is then applied to obtain the compressed *shape latent* $\mathcal{S} \in \mathbb{R}^{M\times C_s}, C_s = 32$. The encoder processes $\mathbf{X}$ to obtain $\mathcal{S}$ as follows:

$$\begin{aligned} \text{index} &= \text{FPS}(\mathbf{X}) && \text{find index} \\ \mathbf{Z} &= \text{PE}(\mathbf{X}) && \text{point embed.} \\ \mathbf{Z}^* &= \text{IndexSelect}(\mathbf{Z}, \text{index}) && \text{subsample point embed.} \\ \mathcal{S} &= \text{KL}(\text{CrossAttn}(\mathbf{Z}, \mathbf{Z}^*)) && \text{compress} \end{aligned} \tag{3}$$

In the decoder, a cross-attention layer is used to fuse the latent codes for occupancy prediction of 3D points through an MLP. To learn the shape latent set, we train the shape auto-encoder by minimizing the binary cross-entropy (BCE) loss between the predicted occupancies of query points randomly sampled in 3D space and actual occupancies :

3.1.2 Deformation Latent Set

To compress temporal shape deformations into compact latent sets, we design a dedicated deformation autoencoder that encodes pairwise deformations between the initial and subsequent frames into downsampled latent sets. To do so, we design a deformation autoencoder which is illustrated in Fig. 3. Unlike Shape Autoencoder that processes single frame, our deformation autoencoder has two input frames: the source $\mathbf{X}_{\text{src}}$ and target point clouds $\mathbf{X}_{\text{tgt}}$ of size $N = 2048$. This pairwise input structure requires a novel encoder architecture, described as follows:

$$
\begin{aligned}
\text{index} &= \text{FPS}(\mathbf{X}_{\text{src}}) && \text{find index} \\
\mathbf{Z}_{\text{src}} &= \text{PE}(\mathbf{X}_{\text{src}}) && \text{source embed} \\
\mathbf{Z}_{\text{tgt}} &= \text{PE}(\mathbf{X}_{\text{tgt}}) && \text{target embed} \\
\mathbf{Z}^*_{\text{src}} &= \text{IndexSelect}(\mathbf{Z}_{\text{src}}, \text{index}) && \text{subsample source embed} \\
\mathbf{Z}^*_{\text{tgt}} &= \text{IndexSelect}(\mathbf{Z}_{\text{tgt}}, \text{index}) && \text{subsample target embed} \\
\mathbf{Z} &= \text{Concat}([\mathbf{Z}_{\text{src}}, \mathbf{Z}_{\text{tgt}}], -1) && \text{concat along channel} \\
\mathbf{Z}^* &= \text{Concat}([\mathbf{Z}^*_{\text{src}}, \mathbf{Z}^*_{\text{tgt}}], -1) && \text{concat along channel} \\
\mathcal{D} &= \text{KL}(\text{CrossAttn}(\mathbf{Z}, \mathbf{Z}^*)) && \text{compress}
\end{aligned}
\tag{4}
$$

We begin by applying Farthest Point Sampling (FPS) to the source point cloud to obtain a set of downsampled indices. These indices are then propagated to the target point cloud using $\text{IndexSelect}(\cdot,\cdot)$, thereby preserving inter-frame correspondences. This operation ensures that features are extracted from consistent local surface regions across frames. While one could employ cross-attention layers to align and fuse features between the source and target through explicit correspondence search, we adopt this simpler yet effective strategy to capture deformation patterns accurately. This design choice not only learns accurate deformation features but also avoids the computational overhead associated with attention-based matching. Similar to the shape autoencoder, we employ a cross-attention mechanism $\text{CrossAttn}(\cdot,\cdot)$ alongside a KL divergence regularizer $\text{KL}(\cdot)$ to obtain compressed deformation latents $M \times C_d, C_d = 32$, effectively capturing localized surface deformation patterns from the source to the target frame. In the decoder, we again leverage cross-attention to reconstruct the flow field, which deforms the source shape that approximates the target shape. The reconstruction loss is defined as the ℓ_2 distance between the predicted and ground-truth target point clouds.

To ensure that deformation latents consistently represent the same local surfaces or object parts across time, we reuse the FPS downsampling indices from the initial frame for all subsequent frames. This design enables direct stacking of deformation latents along the temporal dimension. Benefiting from the correspondence-preserving property of our deformation autoencoder, we can apply attention operations across frames to exchange deformation information, thereby enhancing temporal coherence. This temporal alignment is also crucial for the efficiency of our denoising network, as discussed later in Sec. 3.2. With a mild notational simplification, we denote the resulting motion latent for a sequence as $\mathcal{D} \in \mathbb{R}^{(T-1)\times M \times C_d}$, obtained by stacking the per-frame deformation latents.

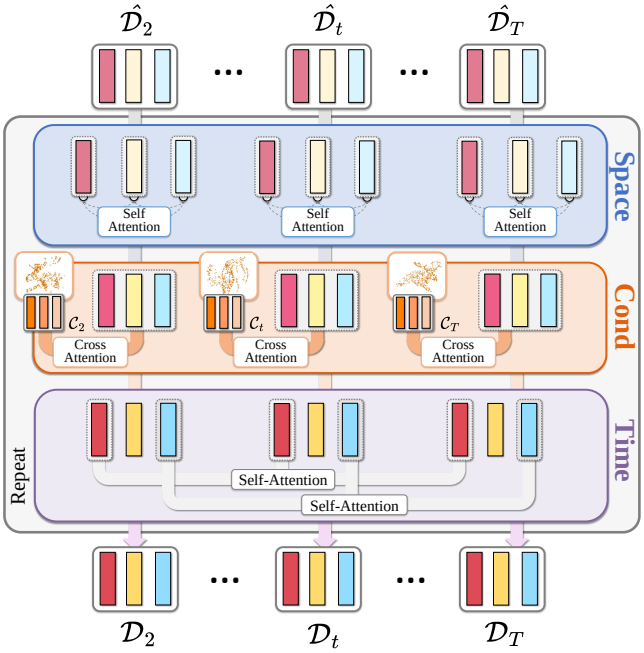


Fig. 4: **Synchronized Deformation Set Diffusion.** Given noised deformation vector sets $\{\hat{\mathcal{D}}_t\}_{t=2}^T$ (top) from a sequence, each set denoted as $\hat{\mathcal{D}}_t = \{\hat{\mathbf{d}}_t^1, ..., \hat{\mathbf{d}}_t^M\}$ of timestep $t \in [2, T]$, we use repeated Interleaved Spatio-Temporal Attention Blocks (ISTA) as our denoising network. In each ISTA block, we first pass them to the space self-attention layer (Space Attention) to aggregate latent features $\hat{\mathcal{D}}^t$ across different spatial locations within each frame to explore spatial contexts. Next, we inject conditional information extracted from imperfect inputs via cross-attention (Condition Attention) between conditional codes $\mathcal{C}_t$ and noised deformation codes $\hat{\mathcal{D}}_t$ at each frame. The inputs could be sparse or partial point clouds, images, or voxel grids. Here we use point clouds as an example. Lastly, to enhance temporal coherence, a time self-attention layer (Time Attention) is used to aggregate latent codes from the same position but from different frames, i.e. $\{\hat{\mathbf{d}}_t^i\}_{t=2}^T$. Repeat this ISTA block and we finally get denoised deformation latent sets $\{\mathcal{D}_t\}_{t=2}^T$ (bottom). Within each layer, different colored latents represent the dynamics of distinct local regions, while the same colored latents represent the dynamics of a local region at different time steps.

3.2 4D Latent Set Diffusion

3.2.1 Shape Diffusion

Following the diffusion paradigm in EDM by Karras et al. [104], we aim to minimize the expected ℓ_2-denoising error. This is achieved by adding the noise ϵ sampled from the Gaussian distribution to the shape latent set $\mathcal{S}$, and then feeding the noise-added code $\hat{\mathcal{S}} = \mathcal{S} + \epsilon$ to the denoiser (to avoid confusing, we also use $\mathcal{S}$ to represent its matrix form $\mathbb{R}^{N\times C_s}$). The whole process is denoted as:

$$
\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})} \left\| \text{ShapeDenoiser}\left(\hat{\mathcal{S}}, \sigma, \mathcal{C}\right) - \mathcal{S} \right\|_2^2 \tag{5}
$$

Here, σ represents the noise level. $\mathcal{C}$ is the conditioning latents extracted from the first input frame $\mathbf{I}_1$ which can be a voxel grid, a point cloud, or an image (see Sec. 3.3).

3.2.2 Synchronized Deformation Diffusion

To adapt these 3D models [11], [12], [13], [14] directly to 4D, the most straightforward approach is frame-by-frame processing, which may lead to discontinuities in temporal and spatial correspondence. Another approach is to aggregate all spatial-temporal latents, which would significantly increase the time complexity to $O(T^2M^2)$ for a sequence of T frames and M deformation latents. However, our 4D latent set representation allows us to bypass the need for full attention across spatial and temporal domains. As discussed in Sec. 3.1, the deformation latents at the same indices across different frames correspond to the deformation behaviors of the same local surface region. Leveraging this property, we implement an alternating way for latent feature aggregation, systematically switching between the spatial and temporal domains. This method not only preserves the spatio-temporal consistency, but also reduces the computational complexity to $O(TM^2)$ in the spatial domain and $O(MT^2)$ in the temporal domain. The details of synchronized deformation diffusion are described as follows. Given a sequence of input observations $\mathcal{I} = \{\mathbf{I}_t\}_{t=1}^{T}$, we pair subsequent frames with the first frame , *i.e.*, $\{\mathbf{I}_1, \mathbf{I}_t\}_{t=2}^{T}$. These pairs are encoded into a series of conditional latents $\mathcal{C}_t = \{\mathbf{c}_i \in \mathbb{R}^C\}_{i=1}^{L}, C = 32$ via conditioning networks which will be described in Sec. 3.3. Then these conditional latents, together with the diffused shape latent set $\mathcal{S}$ in Sec. 3.2.1, are injected into the denoising network as the condition providing guidance for the network to handle ambiguous inputs, like partial point clouds.

Interleaved Spatio-Temporal Attention. Fig. 4 depicts the denoiser network of our proposed synchronized deformation latent set diffusion. The basic unit is the designed Interleaved Spatio-temporal Attention Block (ISTA). We first linearly project the shape latents (with channel dimension C_s) and deformation latents (with channel dimension C_d) into a shared embedding space of dimension C. This results in the following inputs to the denoising network: noisy motion latents of shape $[B, T-1, M, C]$, shape latents of shape $[B, M, C]$, and conditioning latents of shape $[B, T-1, L, C]$, where B is the batch size. Each Interleaved Spatio-Temporal Attention (ISTA) block consists of three attention layers:

- *Space Self-Attention Layer*: performs self-attention along the spatial dimension (M). To enable this, the input tensor is reshaped to $[B \times (T-1), M, C]$, allowing the network to model spatial dependencies within each frame independently.
- *Conditional Cross-Attention Layer*: This layer injects conditioning signals into the denoising network via cross-attention. Prior to the attention computation, both the input deformation latents and conditioning latents are reshaped to $[B \times (T-1), L, C]$. In addition to the primary conditioning, we further incorporate shape latents $\mathcal{S}$ by applying cross-attention between $\mathcal{S}$ and the deformation features of each temporal frame. It enables the network to modulate the motion dynamics based on the underlying static geometry.
- *Time Self-Attention Layer*: This layer performs self-attention along the temporal axis ($T-1$), allowing the network to aggregate motion features over time. To enable this, the input is reshaped to $[B \times M, T-1, C]$ before applying the attention operation. By attending across frames for each spatial location, this layer effectively captures temporal dependencies and consolidates deformation features of corresponding local regions throughout the sequence.

In the denoising phase, we regard the entire sequence of deformation codes as a motion latent set and jointly denoise them together to learn temporal motion priors. We add a Gaussian noise ϵ to the motion latent set $\mathcal{D}$ to obtain its noise-corrupted version $\hat{\mathcal{D}}$. The denoising objective is thus formulated as:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})} \left\| \text{MotionDenoiser}\left(\hat{\mathcal{D}}, \sigma, \mathcal{S}, \mathcal{C}\right) - \mathcal{D} \right\|_2^2 \quad (6)$$

Here, $\mathcal{C}$ is the conditioning latents $\{\mathcal{C}_2, \mathcal{C}_3, ..., \mathcal{C}_T\}$.

3.3 Conditioning Network

In this section, we describe how we extract conditioning embeddings $\mathcal{C}$ from ambiguous inputs $\mathcal{I}$, such as point clouds, voxel grids, images, and sparse handle movements. Given *raw* conditioning signal $\mathcal{I}$ of T frames

$$\begin{aligned} \{\tilde{\mathcal{C}}_1, \ldots, \tilde{\mathcal{C}}_T\} &= \text{Enc}(\mathcal{I}_1, \cdots, \mathcal{I}_T), && \text{per frame embeds} \\ C_1 &= \tilde{\mathcal{C}}_1, \\ C_2 &= \text{Concat}([\tilde{\mathcal{C}}_1, \tilde{\mathcal{C}}_2], -1), && \text{merge with ref} \\ &\vdots \\ C_T &= \text{Concat}([\tilde{\mathcal{C}}_1, \tilde{\mathcal{C}}_T], -1), && \text{merge with ref.} \end{aligned} \quad (7)$$

For reference shape generation in Sec. 3.2.1, we condition the network on $\mathcal{C}_1$. For motion generation in Sec. 3.2.2, we condition on $\mathcal{C}_t$, which represents the relative deformation features of frame t with respect to the first reference frame.

Point clouds. The input point clouds ($L = 300$ or 512 points) may be sparse, partial, or noisy. We encode them into feature embeddings using a transformer encoder, which can obtain L latents.

Voxel grids. A voxel grid represents 3D shapes using binary occupancy values. We process voxel grids at a resolution of 32^3 using a series of 3D convolutional layers to extract volumetric features of resolution 8^3, which are then flattened across spatial dimensions into a sequence of $L = 512$ latents.

RGB images. We use RGB images of resolution $224 \times 224 \times 3$ as inputs. Each image is processed using a pretrained DINO-v2 ViT-B/14 backbone [105], which encodes it into a feature map of size $16 \times 16 \times 768$. This feature map is then flattened into a sequence of $L = 257$ latents, comprising 256 patch embeddings and one global embedding.

4 Experiments

Datasets: We conducted experiments on three 4D datasets. The first, the Dynamic FAUST (D-FAUST) [25], focuses on human body dynamics, including 10 subjects and 129 sequences. It is split into training (70%), validation (10%), and test (20%) subsets, following [15]. The second, the DeformingThings4D-Animals (DT4D-A) [26], includes 38 identities with a total of 1227 animations, divided into training (75%), validation (7.5%), and test (17.5%) subsets following [17]. The training and validation sets use motion

TABLE 2: Quantitative comparisons of 4D shape reconstruction from **sparse and noisy** point cloud sequences on the DT4D-A [26], D-FAUST [25], and S2M [106] datasets.

Dataset	Method	Unseen Motion			Unseen Individual		
		IoU ↑	CD ↓	Corr ↓	IoU ↑	CD ↓	Corr ↓
DT4D-A [26]	OFlow [15]	70.6%	0.104	0.204	57.3%	0.175	0.285
	LPDC [16]	72.4%	0.085	0.162	59.4%	0.149	0.262
	CaDex [17]	80.3%	0.061	0.133	64.7%	0.127	0.239
	Ours	**88.9%**	**0.050**	**0.061**	**83.7%**	**0.058**	**0.074**
D-FAUST [25]	OFlow [15]	81.5%	0.065	0.094	72.3%	0.084	0.117
	LPDC [16]	84.9%	0.055	0.080	76.2%	0.071	0.098
	CaDex [17]	89.1%	0.039	0.070	80.7%	0.055	0.087
	Ours	**90.7%**	**0.033**	**0.047**	**83.7%**	**0.045**	**0.064**
S2M [106]	OFlow [15]	/	/	/	53.2%	0.223	0.204
	LPDC [16]	/	/	/	56.7%	0.173	0.163
	CaDeX [17]	/	/	/	58.9%	0.118	0.160
	Ours	/	/	/	**75.9%**	**0.070**	**0.095**

TABLE 3: Quantitative comparisons of 4D shape completion from **monocular noisy depth scans** on the DT4D-A [26], D-FAUST [25], and S2M [106] datasets.

Dataset	Method	Unseen Motion			Unseen Individual		
		IoU ↑	CD ↓	Corr ↓	IoU ↑	CD ↓	Corr ↓
DT4D-A [26]	OFlow [15]	64.2%	0.305	0.423	55.1%	0.408	0.538
	LPDC [16]	62.2%	0.339	0.427	51.6%	0.467	0.488
	CaDex [17]	70.8%	0.254	0.499	59.2%	0.379	0.498
	Ours	**73.3%**	**0.177**	**0.404**	**66.3%**	**0.193**	**0.438**
D-FAUST [25]	OFlow [15]	76.9%	0.084	0.165	66.4%	0.109	0.194
	LPDC [16]	68.3%	0.138	0.167	59.6%	0.156	0.204
	CaDex [17]	80.7%	0.074	0.123	70.4%	0.096	0.157
	Ours	**83.8%**	**0.054**	**0.111**	**74.4%**	**0.075**	**0.140**
S2M [106]	OFlow [15]	/	/	/	53.6%	0.232	0.228
	LPDC [16]	/	/	/	54.5%	0.217	0.196
	CaDeX [17]	/	/	/	56.3%	0.145	0.186
	Ours	/	/	/	**63.6%**	**0.128**	**0.175**

sequences of seen individuals. The test set is divided into two parts: unseen motions and unseen individuals. The last dataset, the Shape2Motion (S2M) [106], consists of articulated objects across 8 diverse categories, including laptops, doors, staplers, eyeglasses, washing machines, refrigerators, and ovens. Each object category includes one or more articulation controls. For this dataset, we follow the experimental setup proposed by CaDeX [17] and A-SDF [27], where the articulation is rotated to generate 30 frames per object. Among all generated sequences within a given category, 80% are selected as training objects, while the remaining sequences are allocated to the test split.

Baselines: We compare against state-of-the-art methods in 4D reconstruction, including OFlow [15], LPDC [16], CaDex [17]. **OFlow** assigns each 4D point both an occupancy value and a motion velocity vector, utilizing a Neural-ODE framework [107] for learning deformations. **LPDC** employs an MLP to parallelly learn correspondences among occupancy fields across different time steps via explicitly learning continuous displacement vector fields from spatio-temporal shape representation. **CaDeX** introduces a canonical map factorization and utilizes invertible deformation networks to maintain homeomorphisms. For fair comparisons, we follow their original training paradigms.
Evaluation Metrics: The Intersection over Union (IoU) evaluates the volume overlap between predicted and ground truth meshes. The Chamfer Distance (CD) calculates the average nearest-neighbor distance between two point sets. ℓ_2-distance measures the Euclidean distance between corresponding points on the predicted and ground truth meshes.
Implementations: Our training pipeline consists of two stages. In the first stage, we train the shape and deformation auto-encoders. For the shape auto-encoder, we use a learning rate of 10^{-4} and a KL-divergence loss weight of 10^{-3}. For the deformation auto-encoder, the learning rate is also 10^{-4}, with a KL-divergence loss weight of 10^{-6}. Both models are trained for 100 epochs with a batch size of 24. In the second stage, we train the diffusion models. The learning rate is set to 10^{-4} for both the shape and deformation diffusion networks. We train the shape diffusion model for 50 epochs with a batch size of 8, and the deformation diffusion model with a batch size of 4. Both the shape and deformation diffusion models follow the noise scheduling strategy of EDM [104]. During training, the noise level σ is sampled from a log-normal distribution $\mathcal{N}(\mu = -1.2, \sigma = 1.2)$.

TABLE 4: Quantitative comparisons of 4D shape super-resolution from **voxel grid sequences** with resolution 32^3 on the DT4D-A [26] dataset.

Input	Method	Unseen Motion			Unseen Individual		
		IoU ↑	CD ↓	Corr ↓	IoU ↑	CD ↓	Corr ↓
4D Voxel of DT4D-A	OFlow [15]	68.4%	0.237	0.385	58.6%	0.311	0.412
	LPDC [16]	69.3%	0.231	0.311	59.2%	0.321	0.343
	CaDeX [17]	77.1%	0.152	0.269	65.8%	0.228	0.331
	Ours	**84.3%**	**0.066**	**0.126**	**75.6%**	**0.088**	**0.187**

TABLE 5: Quantitative comparisons of 4D shape reconstruction from **RGB image sequences** on the D-FAUST [25] dataset.

Input	Method	Unseen Motion			Unseen Individual		
		IoU ↑	CD ↓	Corr ↓	IoU ↑	CD ↓	Corr ↓
Image sequence of D-FAUST	OFlow [15]	51.5%	0.284	0.319	31.0%	0.431	0.442
	LPDC [16]	52.3%	0.255	0.282	35.2%	0.371	0.397
	CaDeX [17]	53.8%	0.229	0.284	38.7%	0.288	0.352
	Ours	**69.9%**	**0.119**	**0.206**	**51.6%**	**0.189**	**0.299**

During inference, we use 18 denoising steps with noise levels linearly scheduled from $\sigma_{\max} = 80$ to $\sigma_{\min} = 0.002$. Note that the shape and deformation auto-encoders can be trained in parallel, as can the shape and deformation diffusion models. The shape and deformation autoencoders are individually trained for approximately 10 hours, while the shape and deformation diffusion models require around 20 and 30 hours, respectively. All models are trained on two NVIDIA H200 GPUs.

4.1 4D Shape Reconstruction from Point Clouds

We initially assessed our model's ability for 4D reconstruction from sparse and noisy point cloud sequences Following the setup in OFlow [15], our network processed sequences of $T = 17$ continuous frames.

4.1.1 *Sparse & Noisy Point Cloud Sequences*

Each frame represents a sparse point cloud, with $L = 300$ for D-FAUST [25] and S2M [106] or $L = 512$ for DT4D-A [26]. We also simulate noisy observations by adding Gaussian noise (σ = 0.05).

Quantitatively, our model consistently outperforms previous methods on the D-FAUST [25], DT4D-A [26], and S2M [106] datasets, as shown in Table 2. The performance gain is especially notable on the unseen individual split of DT4D-A, which features diverse topologies across various

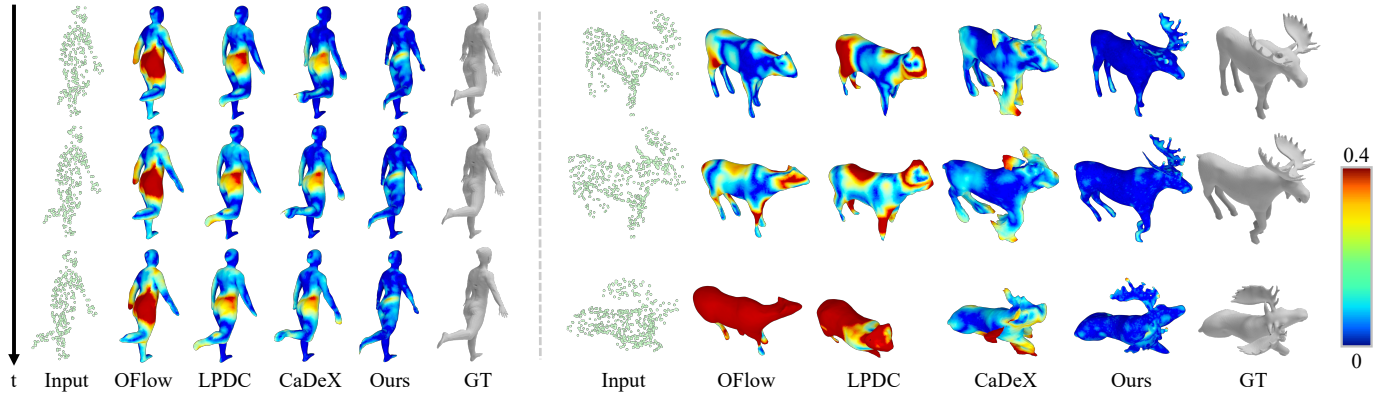


Fig. 5: Qualitative comparisons of 4D shape reconstruction from **sparse and noisy** point clouds on the D-FAUST [25] (left) and DT4D-A [26] (right) datasets. We visualize the Chamfer Distance between reconstruction and ground truth as error maps. Our method reconstructs more accurate surface geometries and motion dynamics.

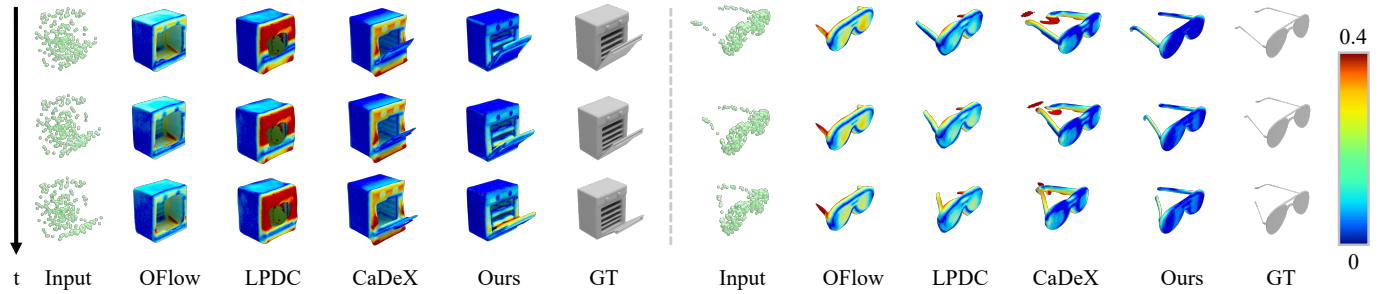


Fig. 6: Qualitative comparisons of 4D shape reconstruction from **sparse and noisy** point clouds on the S2M [106] dataset. Our method reconstructs more accurate surface geometries and motion dynamics.

animal species. Our method improves IoU by 19% and 17% over the previous best models on DT4D-A and S2M, respectively. Additionally, it reduces both Chamfer Distance and ℓ_2-correspondence error to less than half of those reported by existing approaches on D-FAUST and DT4D-A.

Qualitatively, as shown in Fig. 5 and Fig. 6, our model excels at reconstructing complete shapes with minimal Chamfer Distance errors. This advantage is especially clear in challenging regions such as fast-moving body parts (e.g., human feet), highly articulated structures (e.g., animal heads), and fine details (e.g., eyeglass temples).

Our model's strength lies in the 4D latent set diffusion, which enables accurate sampling of local geometry and deformation patterns. While methods like LPDC [16] and OFlow [15] perform well on human datasets with consistent topology, they struggle with the diverse shapes and scales in animal data. Global latent-based approaches also fail to reliably track fast-moving articulated parts, such as hinges. In contrast, our method effectively models complex 4D dynamics across a wide range of non-rigid objects.

4.1.2 Monocular Depth Sequences

To simulate sparse and partial real-world scans, we generated monocular depth sequences by rendering from a fixed camera angle. The size of the input point cloud and the frame length are the same as Sec. 4.1.1.

Quantitative results in Table 3 demonstrate that on the S2M [106] dataset, our method achieves an IoU of 63.6%, outperforming the strongest baseline (CaDeX) by 7.3%, and reducing the Chamfer Distance from 0.145 to 0.128. Similar improvements are observed on the DT4D-A [26] dataset, where our method improves IoU by 7.1% over baselines for unseen individuals. Even on relatively simpler datasets such as D-FAUST [25], our method consistently improves IoU by 4.0% over CaDeX and further reduces the Chamfer Distance.

Qualitative results in Fig. 7 and Fig. 8 further verify our quantitative findings. Baseline methods often produce fragmented surfaces and exhibit inconsistent deformations, particularly in human feet or animal heads. In contrast, our method generates more complete and temporally consistent reconstructions. This advantage is more obvious on the S2M [106] dataset, where our approach preserves fine-scale structural details across time for articulated objects.

The comparisons show that our method reconstructs more complete surfaces with more accurate motion tracking, highlighting the effectiveness of our 4D latent set diffusion in handling ambiguous inputs such as partial scans.

4.2 4D Shape Super-resolution from Voxel Grids

We also conduct comparisons on the task of 4D shape super-resolution from a sequence of coarse voxel grids on the DT4D-A [26] dataset. This task poses significant challenges due to spatial quantization and limited geometric details in the low-resolution volumetric grids. For all baselines, we re-implement all baselines with global latent codes derived from volumetric features as conditions. Quantitative results in Table 4 show that our method outperforms all baselines on the DT4D-A [26] dataset. It achieves an IoU of 84.3%, surpassing CaDeX [17] by over 7%, and reduces the Chamfer Distance by nearly 50%. Qualitative results are shown in Fig. 9, where baseline methods often produce fragmented surfaces and suffer from temporal inconsistencies in complex animal motions. In contrast, our model generates

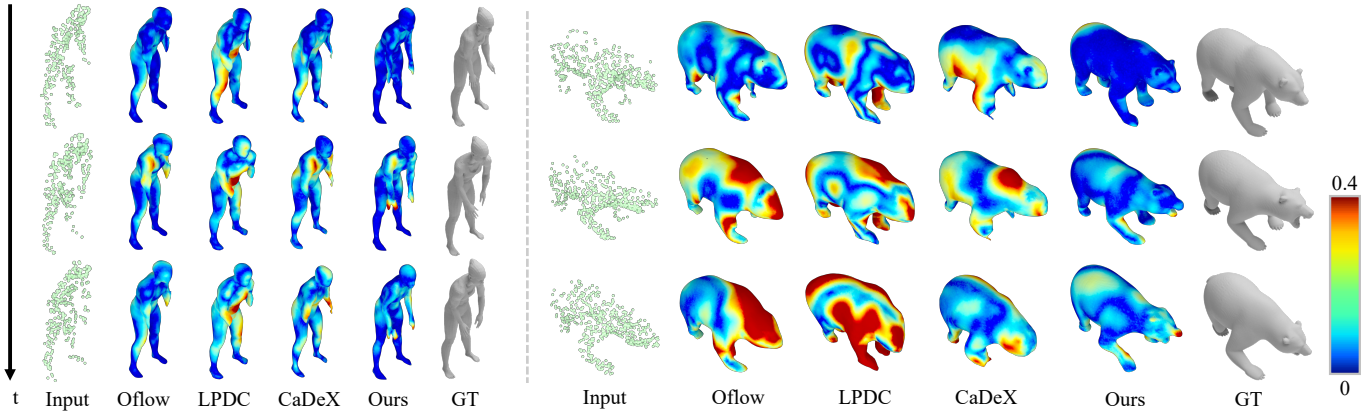


Fig. 7: Qualitative comparisons of 4D shape completion from **monocular noisy depth scans** on the D-FAUST [25] (left) and DT4D-A [26] (right) datasets. Our method reconstructs more accurate surface geometries and motion dynamics.

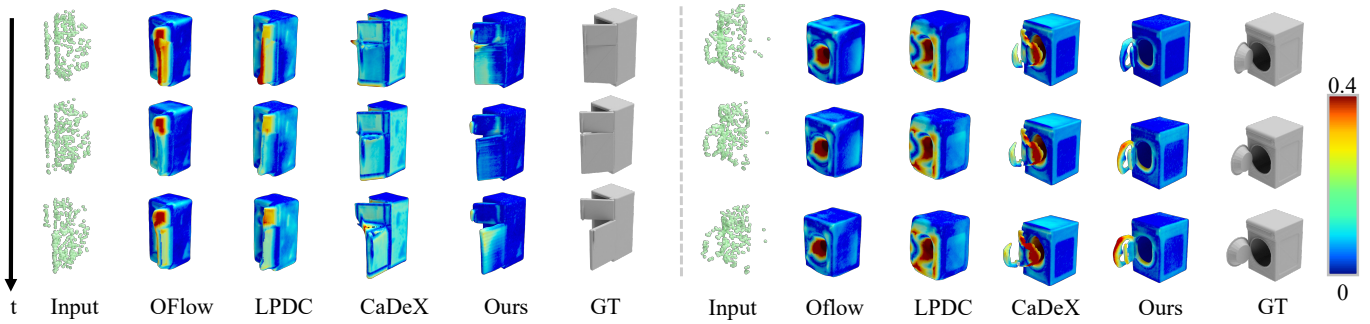


Fig. 8: Qualitative comparisons of 4D shape completion from **monocular noisy depth scans** on the S2M [106] dataset. We visualize the Chamfer Distance between the reconstruction and the ground truth as error maps. Our method exhibits lower reconstruction errors and achieves more plausible motion tracking.

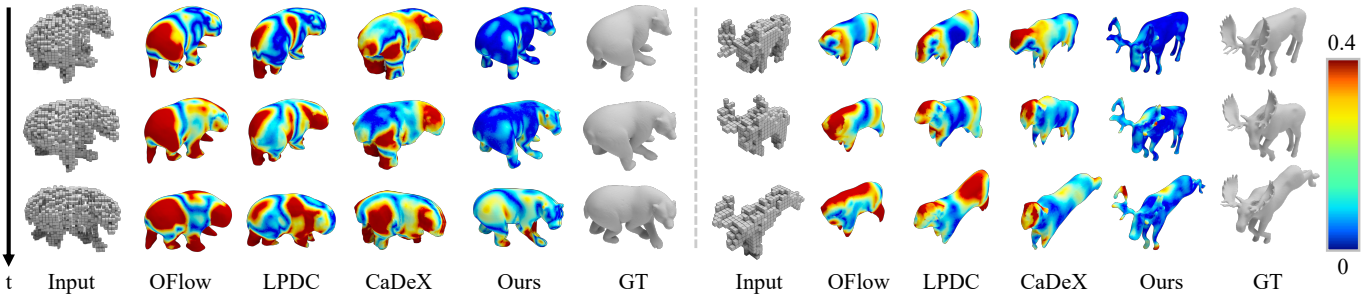


Fig. 9: Qualitative comparisons of 4D shape super-resolution from **voxel grid sequences** with resolution 32^3 on the DT4D-A [26] dataset. Our method exhibits lower reconstruction errors and achieves more plausible tracking.

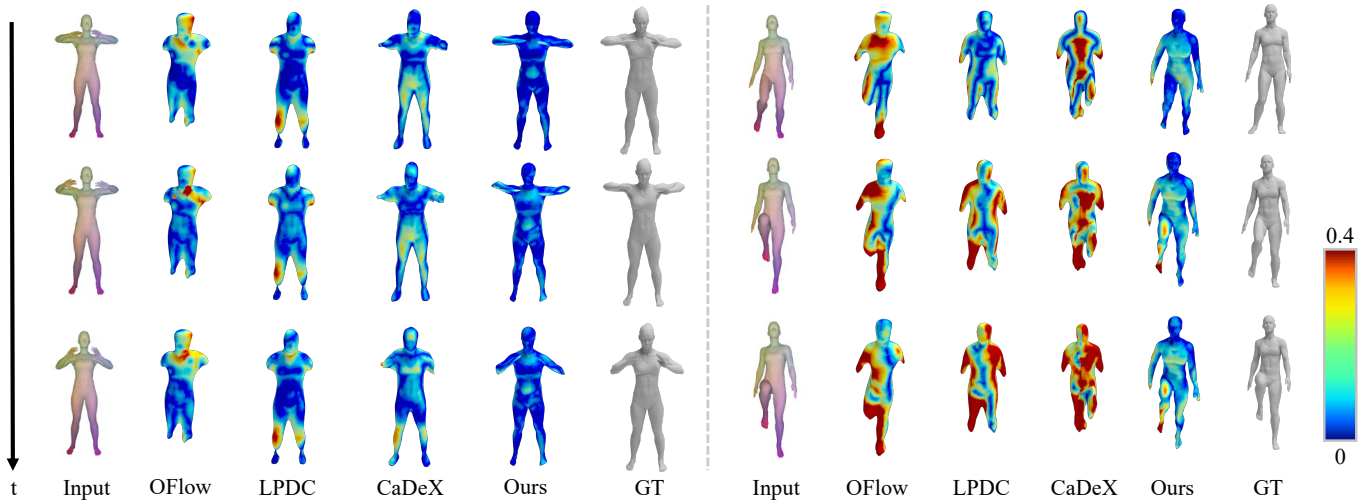


Fig. 10: Qualitative comparisons of 4D shape completion from **RGB image sequences** on the D-FAUST [25] dataset. Our method exhibits lower reconstruction errors and achieves more plausible tracking.

smoother temporal evolution and more complete structures. Remarkably, it can even hallucinate plausible details, such as continuous antlers or smooth limb connections, which are absent in the coarse voxel inputs, highlighting the strength of our diffusion-based 4D surface priors.

4.3 4D Shape Reconstruction from RGB Images

We also evaluate our method on the task of 4D shape reconstruction from RGB image sequences rendered from the D-FAUST [25] dataset. This setting poses greater challenges than voxel or point cloud inputs, as it lacks explicit 3D geometry and relies solely on 2D visual cues. Quantitative results in Table 5 show that on unseen motions, our method improves IoU by over 16% compared to the strongest baseline, CaDex, while also significantly reducing Chamfer Distance and ℓ_2-correspondence errors. As shown in Fig. 10, existing methods often generate incomplete surfaces and exhibit temporal instability, especially around articulated limbs, failing to track motion consistently. However, our model produces coherent shape sequences with improved structural integrity and smooth temporal evolution. Overall, these comparisons demonstrate that our approach enables accurate and temporally consistent 4D reconstruction directly from RGB image sequences.

TABLE 6: **Quantitative ablation studies** of 4D shape completion from **monocular noisy depth scans** on the D-FAUST [25] dataset. M denotes the number of latent codes and C_d represents the number of latent code channels.

Method	Unseen Motion			Unseen Individual		
	IoU↑	CD↓	Corr↓	IoU↑	CD↓	Corr↓
W/o. Diffusion	71.1%	0.097	0.173	64.2%	0.107	0.194
$M = 1$	68.5%	0.120	0.301	57.7%	0.149	0.327
$C_d = 8$	78.9%	0.078	0.180	68.0%	0.105	0.225
$C_d = 16$	78.0%	0.080	0.189	66.8%	0.109	0.254
W/o. TimeAttn.	81.2%	0.061	0.127	70.8%	0.086	0.158
Full ($M = 512$, $C_d = 32$)	**83.8%**	**0.054**	**0.111**	**74.4%**	**0.075**	**0.140**

4.4 Ablation Study

We conducted ablation studies to validate the effectiveness of each component (see Table 6, Fig. 11) under the setting of 4D shape completion from **monocular noisy depth scans** on the D-FAUST [25] dataset.

What is the effect of diffusion model? 4D surface reconstruction from ambiguous observations of noisy, sparse, or partial point clouds is an ill-posed problem. Deterministic models often yield suboptimal results. We compare against a feedforward baseline that uses the same autoencoder architecture but without diffusion. Specifically, it takes as input a sequence of partial point clouds (300 points per frame) and directly predicts the full surfaces in a single forward pass. As shown in Fig. 11 and Table 6, the diffusion model adopts a probabilistic way to deal with highly ambiguous inputs and generates plausible predictions. Moreover, diffusion models can handle "one-to-many" problems and generate diverse and creative outputs as shown in Fig. 16.

What is the effect of 4D latent set representation? Instead of using a single global latent code, our approach employs 4D latent sets. Both variants share the same encoder-decoder backbone to ensure a fair comparison. As indicated in Table 6, our method significantly outperforms the global latent codes (with $M = 1$) and captures more accurate 4D motions. This advantage becomes more apparent for unseen identities, demonstrating better generalization ability.

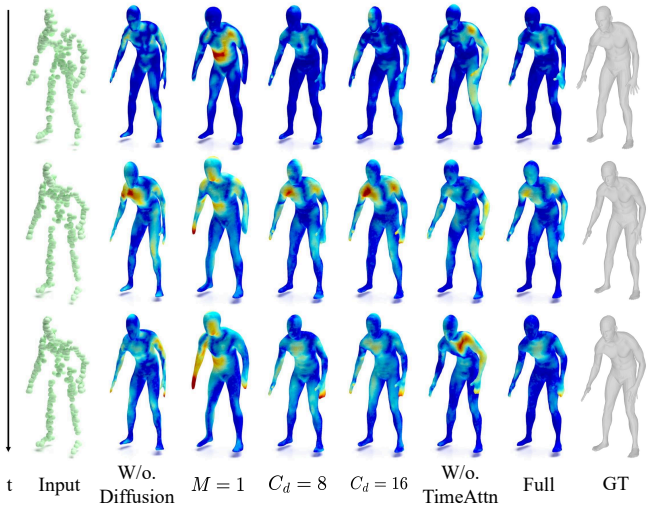


Fig. 11: **Qualitative ablation studies** of 4D shape completion from **monocular noisy depth scans** on the D-FAUST [25] dataset. Without diffusion, reconstructions suffer from incomplete geometry due to the ill-posed nature of sparse inputs. Using a single global latent ($M = 1$) or fewer channels ($C_d = 8/16$) limits the ability to capture local deformations. Removing temporal attention introduces motion discontinuities. The full model with diffusion, 4D latent sets ($M = 512$), and temporal attention achieves the most coherent and accurate 4D reconstructions.

TABLE 7: **Quantitative ablation studies of local latent codes** on the DT4D-A [26] dataset. Given a sequence of **noisy and sparse point clouds**, we compare the reconstruction results between three method variants: (a) global shape & deformation latents; (b) local shape & global deformation latents; (c) local shape & deformation latents (Our final).

Method	Unseen Motion			Unseen Individual		
	IoU ↑	L1 ↓	L2 ↓	IoU ↑	L1 ↓	L2 ↓
(a) Glo. Sha. & Glo. Def.	66.8%	0.336	0.378	57.0%	0.395	0.491
(b) Loc. Sha. & Glo. Def.	83.0%	0.116	0.228	73.1%	0.179	0.397
(c) Loc. Sha. & Loc. Def.	**88.9%**	**0.050**	**0.061**	**83.7%**	**0.058**	**0.074**

What is the effect of time attention layers? For the synchronized deformation latent set diffusion, we integrated the time self-attention layer in our interleaved spatio-temporal attention mechanism. Removing this layer decreased all metrics, highlighting its effectiveness in maintaining temporal coherence.

What is the effect of the number of channels of latent set? For the shape latent set, we follow 3DShape2VecSet [22] and set the number of shape channels to $C_s = 8$. To determine the optimal number of deformation latent channels C_d for learning deformation priors on time-varying surfaces, we conduct an ablation study while keeping C_s fixed. As shown in Table 6, setting $C_d = 32$ provides more favorable performance for 4D latent set diffusion.

What is the effect of local latent codes? To highlight the advantage of local latent codes in our 4D latent set representation for modeling complex geometries and motions, we conduct additional comparisons on the DT4D-A

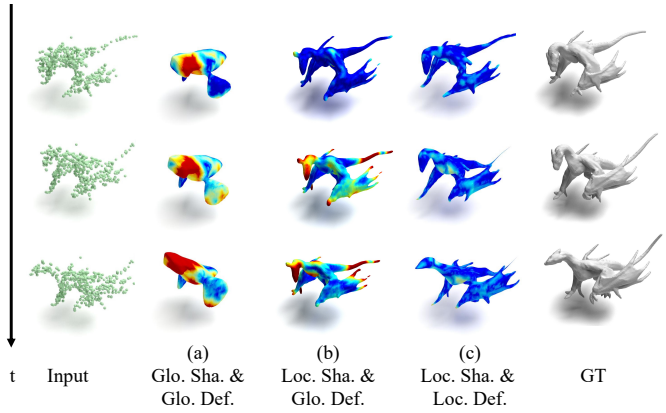


Fig. 12: **Qualitative ablation studies of local latent codes** on the DT4D-A [26] dataset. Given a sequence of noisy and sparse point clouds, we compare the reconstruction results between three method variants: (a) global shape & deformation latents; (b) local shape & global deformation latents; (c) local shape & deformation latents (Our final).

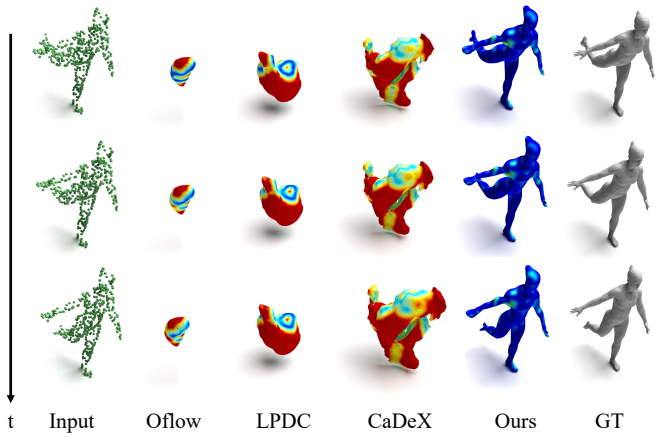


Fig. 13: Qualitative comparisons of **Cross-Dataset Generalization** on the task of 4D shape reconstruction from noisy and sparse point clouds. All methods are trained on the DT4D dataset and directly evaluated on the D-FAUST dataset.

dataset. We compare (c) our full model (local shape and deformation latents) against two variants: (a) a global shape latent with global deformation latents, and (b) local shape latents with global deformation latents, on the task of 4D shape reconstruction from noisy and sparse point clouds. As shown in the first row of Fig. 12, local shape latents reconstruct complex shapes of unseen individuals even at the first frame, whereas a single global shape latent fails to produce plausible shapes. In the 2nd and 3rd rows, local deformation latents track more accurate non-linear motions than global deformation latents. These observations are also supported by the quantitative results in Table 7.

4.5 Cross-Dataset Generalization

To evaluate cross-dataset generalization, we train all methods on DT4D-Animals and test them on the D-FAUST human dataset for 4D shape reconstruction from noisy and sparse point clouds. As shown in Fig. 13, our method remains stable in both geometry reconstruction and motion tracking, while the baseline methods fail to generalize across datasets. As reported in Table 8, our approach outperforms all baselines across all listed metrics on both test subsets in the cross-dataset setting."Test Set 1" and "Test Set 2" correspond to "Unseen Motion" and "Unseen Individual" respectively in the in-dataset evaluation in the Sec. 4.1. Both subsets are unseen individuals in the cross-dataset evaluation. Notably, the performance gaps are larger than those observed on D-FAUST in Table 2, which further demonstrates the superior generalization ability of the proposed 4D latent set representation compared to existing methods.

TABLE 8: Quantitative comparisons of **Cross-Dataset Generalization** on the task of 4D shape reconstruction from noisy and sparse point clouds. All methods are trained on the DT4D dataset and directly evaluated on the D-FAUST dataset.

Dataset	Method	Test Set 1			Test Set 2		
		IoU ↑	L1 ↓	L2 ↓	IoU ↑	L1 ↓	L2 ↓
D-FAUST [25]	OFlow [15]	29.8%	0.427	0.499	24.5%	0.466	0.533
	LPDC [16]	27.1%	0.437	0.527	23.1%	0.430	0.509
	CaDeX [17]	37.6%	0.312	0.405	28.8%	0.337	0.474
	Ours	**82.0%**	**0.059**	**0.078**	**79.9%**	**0.055**	**0.073**

4.6 Additional Results

4.6.1 *Shape Manipulation from Sparse Handle Movements*

We further showcase the applicability of our method to user-driven shape editing. Given a source mesh, we randomly select ten sparse vertices as control handles, each assigned a target position to specify the desired deformation. The goal is to propagate these sparse constraints to generate a sequence of globally coherent deforming meshes.

This task only requires motion diffusion. We first obtain the shape latent set by encoding the source mesh using the pre-trained shape auto-encoder. The trajectories of the control handles are then encoded into conditioning embeddings for motion diffusion, following the same architecture described in Eq. (4). As illustrated in Fig. 15, our model generates realistic global deformations that faithfully respect sparse user edits, even on unseen identities. The results preserve fine geometric details and exhibit coherent surface displacements, highlighting the effectiveness and potential of our learned deformation priors for interactive non-rigid motion synthesis.

4.6.2 *Highly Partial Scan Sequences*

To assess the robustness of our method to extremely ambiguous data, we set up a challenging experiment on the D-FAUST [25] dataset. This involved reconstructing whole body motions based on partial point clouds of the upper bodies. This setup creates a highly ambiguous scenario, as the same upper body motion can correspond to many different lower-body motions. We adopt the same configuration as Sec. 4.1, with a frame length ($T = 17$) and input point cloud size ($L = 300$). As shown in Fig. 16, OFlow [15], LPDC [16], and CaDex [17] face challenges in reconstructing the complete shape, often producing distorted shapes such as broken feet. In contrast, our method excels in reconstructing more complete geometries while achieving temporally

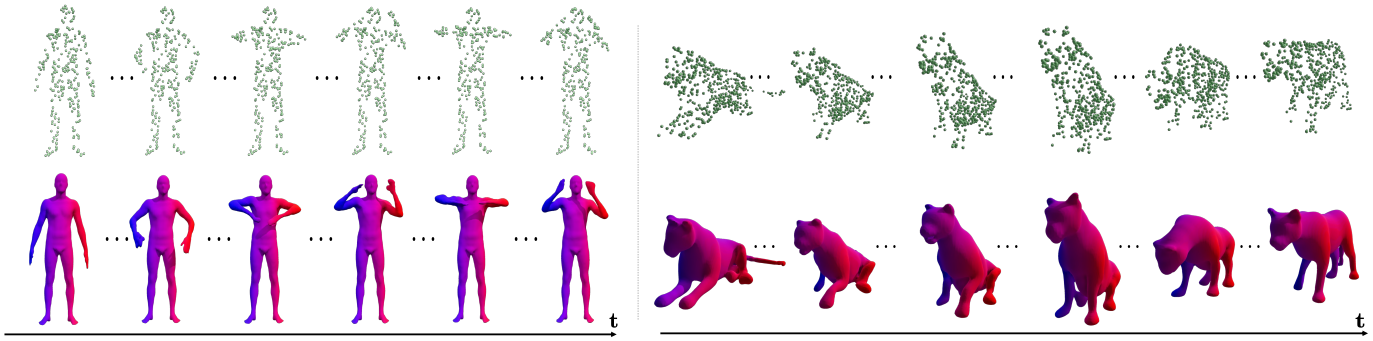


Fig. 14: **4D shape reconstruction on 100-frame sequences from D-FAUST [25] and DT4D-A [26] using sparse and noisy inputs**. Although trained only on 17-frame clips, our method can be extended to much longer sequences by performing 17-frame diffusion in an autoregressive manner, maintaining temporal coherence and preserving fine geometric details over extended motions.

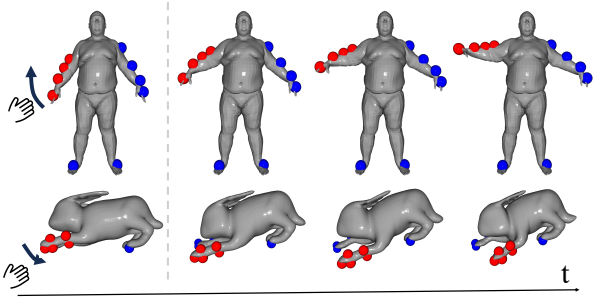


Fig. 15: **Shape manipulation by sparse handle movements.** Given a 3D shape as input, we edit the shape by dragging a few handles for regions of interest. Red spheres indicate displaced handles, while blue spheres represent fixed constraints. Our model propagates local edits into globally consistent deformations, producing realistic surface motions even for unseen shapes.

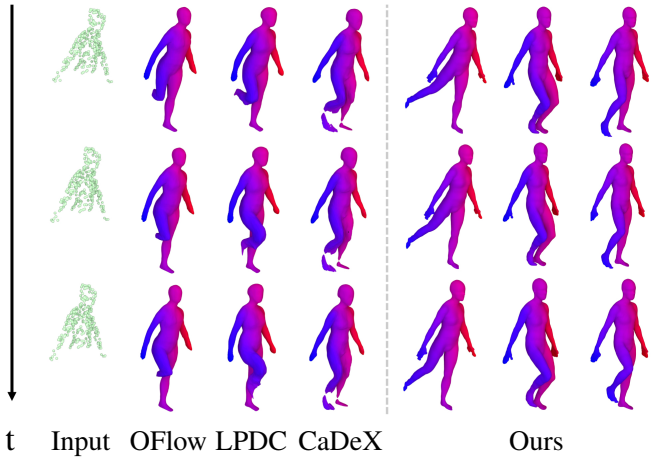


Fig. 16: Qualitative comparisons of 4D shape reconstruction from **highly partial** point cloud sequences, such as half-body scans obtained from the D-FAUST [25] dataset. The colors of the meshes encode the correspondence. Our diffusion-based method produces highly complete human shapes with more favorable motions, offering multiple possible outputs that match the input observations.

coherent tracking. Additionally, our approach presents a diverse range of plausible full-body reconstructions that align with the given upper-body scans. The superior performance is primarily attributed to the 4D latent set diffusion. Our diffusion-based method is more capable of tackling the 'one-to-many' complexities from extremely partial data.

4.6.3 Long Sequence Reconstruction

Although our method is trained on 4D sequences of fixed length $T = 17$, it can be naturally extended to longer sequences without retraining. We achieve this by splitting long videos into overlapping 17-frame clips and applying our 4D diffusion model autoregressively across time. As shown in Fig. 14, our method generates temporally consistent and structurally coherent deformations over 100-frame sequences from the D-FAUST [25] and DT4D-A [26] datasets, demonstrating strong scalability and robustness for long-range 4D shape generation.

4.6.4 Real-world Data Generalization

We validate our model on real-world human scans from the BEHAVE dataset [108], which captures RGB-D data using four Kinect sensors. To align with our partial scan setting, we use a single fixed-view RGB-D stream and back-project the depth map into a partial 3D point cloud as input. Fig. 18 illustrates our reconstruction results alongside the corresponding RGB input. Even in the presence of severe occlusions, such as limbs obscured by objects, our model successfully infers plausible and complete surface reconstructions. This robustness stems from the learned latent distribution and the generative power of the diffusion model, which enables it to reason about missing geometry. Remarkably, our approach maintains high-quality results without any fine-tuning on real data. It also handles sensor noise effectively, demonstrating strong generalization to real-world, imperfect observations.

To demonstrate the generalization and robustness of our model on real-world monocular RGB videos, we use TripoSG [109] to obtain initial reconstructions from each RGB frame, and then sample dynamic point clouds as inputs for our model. The results are visualized in Fig. 19. Our method can still reconstruct plausible geometries and reliable motions with temporal correspondence for animals and articulated objects, such as a bear and a laptop.

4.6.5 Failure Cases

While our method shows strong performance for 4D reconstruction, it still has several limitations. We present failure

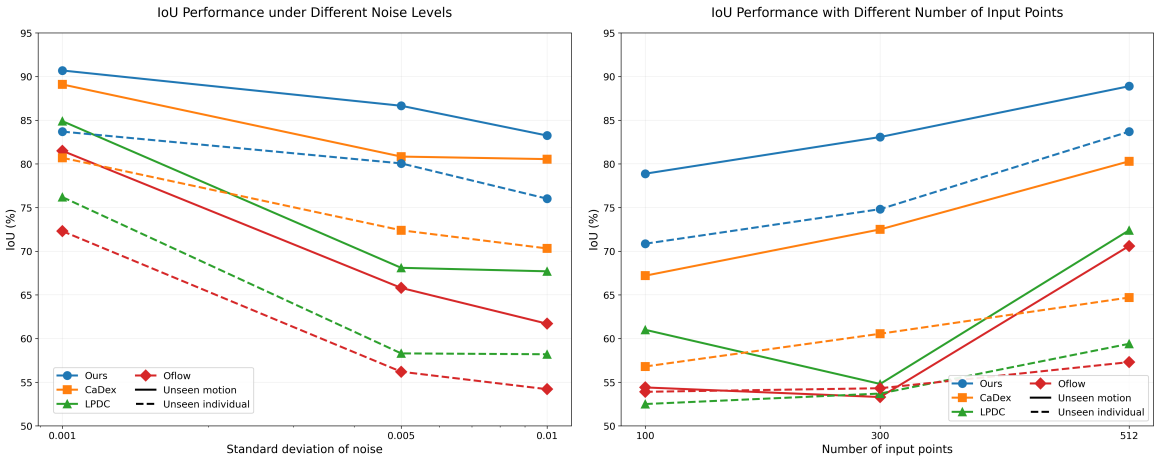


Fig. 17: **Robustness analysis** of Motion2VecSets and previous methods. Left: Results on the D-FAUST [25] dataset under varying noise levels. Right: Results on the DT4D-A [26] dataset with different numbers of input points. Colors and marker shapes denote different methods, while line styles indicate evaluation settings (solid: unseen motions; dashed: unseen individuals).

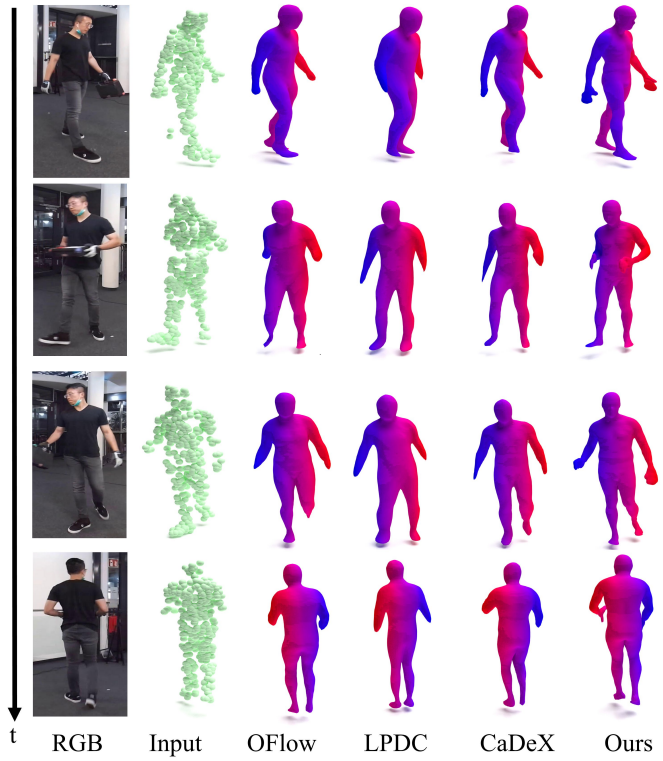


Fig. 18: 4D Shape Completion on the BEHAVE [108] dataset.

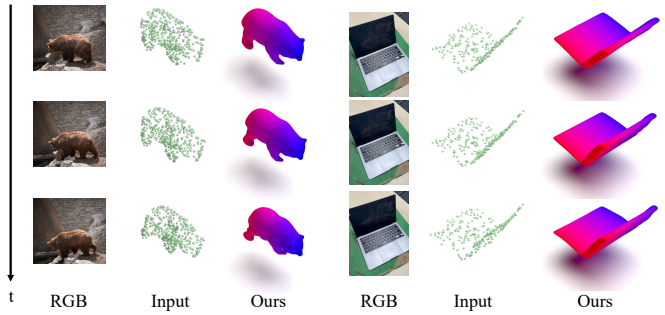


Fig. 19: Real-world generalization results on the sparse point clouds estimated by TripoSG [109] from monocular RGB videos.

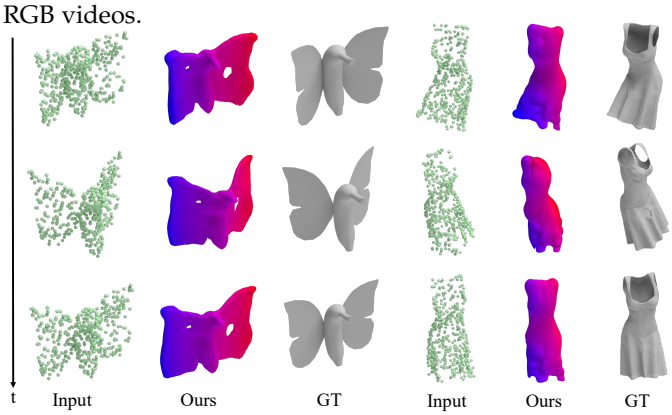


Fig. 20: Failure cases. Our method shows limitations when reconstructing identities that are far from the training data distribution, such as the butterfly and the cloth. It also has difficulty recovering realistic cloth deformations that require physical properties and simulation.

cases in Fig. 20. As shown, our method is inferior at reconstructing geometries of unseen identities that are far from the training data distribution, such as the butterfly and the cloth. It also has difficulty recovering realistic surface deformations that depend on physical properties and simulation, such as the cloth deformations on the right. We expect these issues to be reduced by training on a larger dataset, using stronger data augmentation, and incorporating physical attributes as additional inputs in future work.

4.7 Robustness Analysis

We evaluate the robustness of our model on the D-FAUST [25] and DT4D-A [26] datasets under various input degradations. For D-FAUST, we inject Gaussian noise with standard deviations of 0.001, 0.005, and 0.01. For DT4D-A, we vary the number of input points (100, 300, 500). As shown in Fig. 17, our method consistently outperforms state-of-the-art baselines and generalizes better to unseen subjects under various ambiguous observations.

4.8 Runtime and Memory

In Table 9, we report the computational cost of each method, including inference time, training memory, and the

TABLE 9: Computational efficiency analysis. **Top:** Comparison with state-of-the-art methods. **Bottom:** Runtime breakdown of our M2V pipeline.

Method	Avg. Time (s) ↓	Avg. GPU (GB) ↓	Params (M) ↓
OFlow [15]	1.800	0.265	1.2
LPDC [16]	1.769	0.453	8.5
CaDeX [17]	4.357	5.814	2.1
Ours	9.479	4.358	5.3

Stage (Ours)	Time (s) ↓	Percentage (%)
Shape Diffusion	3.595	37.9%
Deformation Diffusion	5.384	56.8%
Mesh Generation	0.352	3.7%
Total	**9.479**	**100.0%**

total number of trainable parameters. The runtime evaluation is conducted on the D-FAUST [25] dataset with a sparse setting of $L = 300$ points per frame. We also provide a detailed runtime breakdown for each stage of our pipeline under this configuration. Notably, the inference time remains consistent across different input modalities (e.g., point cloud, voxel grid, image), as the conditional encoders are lightweight relative to the core backbone networks. Due to the multiple denoising steps, our method runs slower than conventional feed-forward methods. The inference speed can be improved by applying distribution matching distillation [110].

5 Conclusion, Limitations, and Future Work

We presented Motion2VecSets, a 4D diffusion model for dynamic surface reconstruction and generation from various imperfect inputs, including sparse point clouds, coarse voxel grids, and RGB images. To enable high-quality 4D diffusion, we introduced a 4D latent set representation that compactly encodes initial geometry and pairwise deformations using transformer-based encoders and decoders. Our model captures the probabilistic distribution of non-rigid shape and motion through iterative denoising, enabling plausible and diverse outputs. To ensure temporally smooth motion, we jointly diffuse stacked deformation latent sets across frames. For computational efficiency, we design an interleaved space-time attention block that synchronizes deformation latents over time while significantly reducing memory usage. Unlike global latent representations, our 4D latent set offers more accurate modeling of complex nonlinear motions and improves generalization to unseen identities and motion patterns. Extensive experiments across multiple datasets, including D-FAUST (humans), DT4D-A (animals), and S2M (articulated objects), demonstrate the effectiveness of our approach in reconstructing and tracking non-rigid objects from ambiguous inputs.

While our method achieves notable progress in 4D reconstruction, it has several limitations. First, it relies on direct 4D supervision from dynamic mesh sequences, which are difficult to collect at large scale. As a result, more diverse and larger datasets are still needed to learn more general 4D shape priors. A promising direction is to explore hybrid supervision combining mesh and video data to improve generalization. Second, our current framework does not incorporate texture reconstruction or generation, which is important for high-fidelity applications. Third, our framework focuses on single-object modeling; extending it to dynamic scenes with multiple interacting objects remains an open challenge. Finally, our deformation priors are learned purely from data, without incorporating physical constraints. Integrating physics-based priors could yield more realistic and physically plausible motion patterns for objects such as hair, cloth, and fluids. We leave these challenges for future work.

Acknowledgements

This project was funded by the ERC Consolidator Grant Gen3D (101171131). Wei Cao and Yaoyao Liu are also supported by the National Artificial Intelligence Research Resource (NAIRR) Pilot under award NAIRR250199 and by Delta and DeltaAI at the National Center for Supercomputing Applications (NCSA) through ACCESS allocations CIS250012, CIS250816, and CIS251188.

References

[1] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *Seminal graphics: pioneering efforts that shaped the field*, 1998, pp. 347–353. 1

[2] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, vol. 7, no. 4, 2006. 1

[3] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE international symposium on mixed and augmented reality*. Ieee, 2011, pp. 127–136. 1

[4] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 519–528. 1

[5] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 343–352. 1, 4

[6] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger, "Volumedeform: Real-time volumetric non-rigid reconstruction," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 2016, pp. 362–379. 1, 4

[7] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, "Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7287–7296. 1, 4

[8] O. Sorkine and M. Alexa, "As-rigid-as-possible surface modeling," in *Symposium on Geometry processing*, vol. 4, 2007, pp. 109–116. 1, 4

[9] R. W. Sumner, J. Schmid, and M. Pauly, "Embedded deformation for shape manipulation," in *ACM siggraph 2007 papers*, 2007, pp. 80–es. 1, 4

[10] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015. 1, 2, 3

[11] A. A. A. Osman, T. Bolkart, and M. J. Black, "STAR: A sparse trained articulated human body regressor," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 598–613. [Online]. Available: https://star.is.tue.mpg.de 1, 3, 7

[12] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, Nov. 2017. 1, 2, 3, 7

[13] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, pp. 194:1–194:17, 2017. [Online]. Available: https://doi.org/10.1145/3130800.3130813 1, 3, 7

[14] S. Zuffi, A. Kanazawa, D. Jacobs, and M. J. Black, "3D menagerie: Modeling the 3D shape and pose of animals," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017. 1, 3, 7

[15] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Occupancy flow: 4d reconstruction by learning particle dynamics," in *International Conference on Computer Vision*, Oct. 2019. 1, 2, 4, 5, 7, 8, 9, 12, 15

[16] J. Tang, D. Xu, K. Jia, and L. Zhang, "Learning parallel dense correspondence from spatio-temporal descriptors for efficient and robust 4d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6022–6031. 1, 2, 4, 5, 8, 9, 12, 15

[17] J. Lei and K. Daniilidis, "Cadex: Learning canonical deformation coordinate space for dynamic surface representation via neural homeomorphism," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. [Online]. Available: https://cis.upenn.edu/~leijh/projects/cadex 1, 2, 4, 5, 7, 8, 9, 12, 15

[18] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3, 4

[19] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174. 2

[20] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4460–4470. 2

[21] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 523–540. 2

[22] B. Zhang, J. Tang, M. Nießner, and P. Wonka, "3DShape2VecSet: A 3d shape representation for neural fields and generative diffusion models," *ACM Trans. Graph.*, vol. 42, no. 4, jul 2023. [Online]. Available: https://doi.org/10.1145/3592442 2, 4, 5, 11

[23] X. Zhang, N. Li, and A. Dai, "Dnf: Unconditional 4d generation with dictionary-based neural fields," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 26 047–26 056. 2, 4

[24] W. Cao, C. Luo, B. Zhang, M. Nießner, and J. Tang, "Motion2vecsets: 4d latent vector set diffusion for non-rigid shape reconstruction and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 20 496–20 506. 2

[25] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black, "Dynamic FAUST: Registering human bodies in motion," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017. 2, 7, 8, 9, 10, 11, 12, 13, 14, 15

[26] Y. Li, H. Takehara, T. Taketomi, B. Zheng, and M. Nießner, "4dcomplete: Non-rigid motion estimation beyond the observable surface." *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 7, 8, 9, 10, 11, 12, 13, 14

[27] J. Mu, W. Qiu, A. Kortylewski, A. Yuille, N. Vasconcelos, and X. Wang, "A-sdf: Learning disentangled signed distance functions for articulated shape representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 001–13 011. 2, 8

[28] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 3

[29] X. Han, Z. Li, H. Huang, E. Kalogerakis, and Y. Yu, "High-resolution shape completion using deep neural networks for global structure and local geometry inference," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2017, pp. 85–93. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.19 3

[30] M. Gadelha, S. Maji, and R. Wang, "3d shape induction from 2d views of multiple objects," in *2017 International Conference on 3D Vision (3DV)*, 2017, pp. 402–411. 3

[31] D. Stutz and A. Geiger, "Learning 3d shape completion from laser scan data with weak supervision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2018. 3

[32] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613. 3

[33] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. J. Guibas, "Learning representations and generative models for 3d point clouds," *arXiv preprint arXiv:1707.02392*, 2017. 3

[34] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 259–16 268. 3

[35] G. Riegler, A. O. Ulusoy, and A. Geiger, "Octnet: Learning deep 3d representations at high resolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[36] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong, "O-cnn: Octree-based convolutional neural networks for 3d shape analysis," *ACM Transactions On Graphics (TOG)*, vol. 36, no. 4, pp. 1–11, 2017. 3

[37] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2088–2096. 3

[38] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *ECCV*, 2018. 3

[39] T. Groueix, M. Fisher, V. G. Kim, B. Russell, and M. Aubry, "AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[40] Y. Liao, S. Donné, and A. Geiger, "Deep marching cubes: Learning explicit surface representations," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[41] J. Tang, X. Han, J. Pan, K. Jia, and X. Tong, "A skeleton-bridged deep learning approach for generating meshes of complex topologies from single rgb images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[42] J. Pan, X. Han, W. Chen, J. Tang, and K. Jia, "Deep mesh reconstruction from single rgb images via topology modification networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9964–9973. 3

[43] J. Tang, X. Han, M. Tan, X. Tong, and K. Jia, "Skeletonnet: A topology-preserving solution for learning mesh reconstruction of object surfaces from rgb images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 6454–6471, 2021. 3

[44] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4

[45] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3, 4

[46] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," in *Proceedings of Machine Learning and Systems 2020*, 2020, pp. 3569–3579. 4

[47] J. Chibane, T. Alldieck, and G. Pons-Moll, "Implicit functions in feature space for 3d shape reconstruction and completion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 4

[48] R. Chabra, J. E. Lenssen, E. Ilg, T. Schmidt, J. Straub, S. Lovegrove, and R. Newcombe, "Deep local shapes: Learning local sdf priors for detailed 3d reconstruction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer, 2020, pp. 608–625. 4

[49] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Stoll, and C. Theobalt, "PatchNets: Patch-Based Generalizable Deep Implicit 3D Shape Representations," *European Conference on Computer Vision (ECCV)*, 2020. 4

[50] C. Chen, Y.-S. Liu, and Z. Han, "Gridpull: Towards scalability in learning implicit representations from 3d point clouds," in

Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2023. 4

[51] J. Tang, J. Lei, D. Xu, F. Ma, K. Jia, and L. Zhang, "Sa-convonet: Sign-agnostic optimization of convolutional occupancy networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6504–6513. 4

[52] J. Tang, L. Markhasin, B. Wang, J. Thies, and M. Nießner, "Neural shape deformation priors," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 117–17 132, 2022. 4

[53] P. Palafox, A. Božič, J. Thies, M. Nießner, and A. Dai, "Npms: Neural parametric models for 3d deformable shapes," *arXiv preprint arXiv:2104.00702*, 2021. 4

[54] B. Jiang, Y. Zhang, X. Wei, X. Xue, and Y. Fu, "Learning compositional representation for 4d captures with neural ode," in *CVPR*, 2021. 4

[55] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020. 4

[56] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020. 4

[57] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021. 4

[58] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021. 4

[59] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022. 4

[60] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695. 4

[61] A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer, "Retrieval-augmented diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15 309–15 324, 2022. 4

[62] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023. 4

[63] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023. 4

[64] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022. 4

[65] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023. 4

[66] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733, 2023. 4

[67] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," *Advances in neural information processing systems*, vol. 35, pp. 4328–4343, 2022. 4

[68] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg, "Structured denoising diffusion models in discrete state-spaces," *Advances in neural information processing systems*, vol. 34, pp. 17 981–17 993, 2021. 4

[69] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, "Diffuseq: Sequence to sequence text generation with diffusion models," *arXiv preprint arXiv:2210.08933*, 2022. 4

[70] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. Salimans, "Imagen video: High definition video generation with diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2210.02303 4

[71] S. Luo and W. Hu, "Diffusion probabilistic models for 3d point cloud generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 4

[72] L. Zhou, Y. Du, and J. Wu, "3d shape generation and completion through point-voxel diffusion," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5826–5835. 4

[73] G. Chou, Y. Bahat, and F. Heide, "Diffusion-sdf: Conditional generative modeling of signed distance functions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2262–2272. 4

[74] J. Tang, Y. Nie, L. Markhasin, A. Dai, J. Thies, and M. Nießner, "Diffuscene: Denoising diffusion models for generative indoor scene synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4

[75] Y. Siddiqui, J. Thies, F. Ma, Q. Shan, M. Nießner, and A. Dai, "Texturify: Generating textures on 3d shape surfaces," in *European Conference on Computer Vision*. Springer, 2022, pp. 72–88. 4

[76] Y. Zhang, Y. Liu, Z. Xie, L. Yang, Z. Liu, M. Yang, R. Zhang, Q. Kou, C. Lin, W. Wang *et al.*, "Dreammat: High-quality pbr material generation with geometry-and light-aware diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–18, 2024. 4

[77] D. Z. Chen, Y. Siddiqui, H.-Y. Lee, S. Tulyakov, and M. Nießner, "Text2tex: Text-driven texture synthesis via diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 18 558–18 568. 4

[78] R. Jiang, C. Wang, J. Zhang, M. Chai, M. He, D. Chen, and J. Liao, "Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 371–14 382. 4

[79] N. Kolotouros, T. Alldieck, A. Zanfir, E. Bazavan, M. Fieraru, and C. Sminchisescu, "Dreamhuman: Animatable 3d avatars from text," *Advances in neural information processing systems*, vol. 36, pp. 10 516–10 529, 2023. 4

[80] J. Tang, A. Dai, Y. Nie, L. Markhasin, J. Thies, and M. Nießner, "Dphms: Diffusion parametric head models for depth-based tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 1111–1122. 4

[81] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Behave: Dataset and method for tracking human object interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 935–15 946. 4

[82] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "Grab: A dataset of whole-body human grasping of objects," in *European conference on computer vision*. Springer, 2020, pp. 581–600. 4

[83] Y. Huang, O. Taheri, M. J. Black, and D. Tzionas, "Intercap: Joint markerless 3d tracking of humans and objects in interaction," in *DAGM German Conference on Pattern Recognition*. Springer, 2022, pp. 281–299. 4

[84] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black, "Populating 3d scenes by learning human-scene interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 708–14 718. 4

[85] A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, K. Kreis *et al.*, "Lion: Latent point diffusion models for 3d shape generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 021–10 039, 2022. 4

[86] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021. 4

[87] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023. 4

[88] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022. 4

[89] J. R. Shue, E. R. Chan, R. Po, Z. Ankner, J. Wu, and G. Wetzstein, "3d neural field generation using triplane diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 875–20 886. 4

[90] B. Zhang and P. Wonka, "Lagem: A large geometry model for 3d representation learning and diffusion," *arXiv preprint arXiv:2410.01295*, 2024. 4

[91] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang, "Structured 3d latents for scalable and

versatile 3d generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 469–21 480. 4

[92] S. Wu, Y. Lin, F. Zhang, Y. Zeng, J. Xu, P. Torr, X. Cao, and Y. Yao, "Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer," *arXiv preprint arXiv:2405.14832*, 2024. 4

[93] X. Yang, H. Shi, B. Zhang, F. Yang, J. Wang, H. Zhao, X. Liu, X. Wang, Q. Lin, J. Yu *et al.*, "Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation," *arXiv preprint arXiv:2411.02293*, 2024. 4

[94] Z. Zhao, Z. Lai, Q. Lin, Y. Zhao, H. Liu, S. Yang, Y. Feng, M. Yang, S. Zhang, X. Yang *et al.*, "Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation," *arXiv preprint arXiv:2501.12202*, 2025. 4

[95] J. Lei, C. Deng, B. Shen, L. Guibas, and K. Daniilidis, "Nap: Neural 3d articulation prior," 2023. 4

[96] I. Liu, Z. Xu, W. Yifan, H. Tan, Z. Xu, X. Wang, H. Su, and Z. Shi, "Riganything: Template-free autoregressive rigging for diverse 3d assets," *arXiv preprint arXiv:2502.09615*, 2025. 4

[97] C. Song, J. Zhang, X. Li, F. Yang, Y. Chen, Z. Xu, J. H. Liew, X. Guo, F. Liu, J. Feng *et al.*, "Magicarticulate: Make your 3d models articulation-ready," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 15 998–16 007. 4

[98] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," *arXiv preprint arXiv:2209.14916*, 2022. 4

[99] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz, "Physdiff: Physics-guided human motion diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 010–16 021. 4

[100] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 4

[101] J. Ren, C. Xie, A. Mirzaei, K. Kreis, Z. Liu, A. Torralba, S. Fidler, S. W. Kim, H. Ling *et al.*, "L4gm: Large 4d gaussian reconstruction model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 56 828–56 858, 2024. 4

[102] H. Liang, Y. Yin, D. Xu, H. Liang, Z. Wang, K. N. Plataniotis, Y. Zhao, and Y. Wei, "Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models," *arXiv preprint arXiv:2405.16645*, 2024. 4

[103] Y. Jiang, L. Zhang, J. Gao, W. Hu, and Y. Yao, "Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video," *arXiv preprint arXiv:2311.02848*, 2023. 4

[104] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. NeurIPS*, 2022. 6, 8

[105] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023. 7

[106] X. Wang, B. Zhou, Y. Shi, X. Chen, Q. Zhao, and K. Xu, "Shape2motion: Joint analysis of motion parts and attributes from 3d shapes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8876–8884. 8, 9, 10

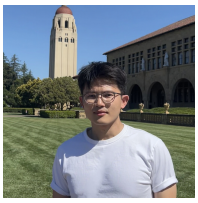
[107] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018. 8

[108] B. L. Bhatnagar, X. Xie, I. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Behave: Dataset and method for tracking human object interactions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022. 13, 14

[109] Y. Li, Z.-X. Zou, Z. Liu, D. Wang, Y. Liang, Z. Yu, X. Liu, Y.-C. Guo, D. Liang, W. Ouyang *et al.*, "Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models," *arXiv preprint arXiv:2502.06608*, 2025. 13, 14

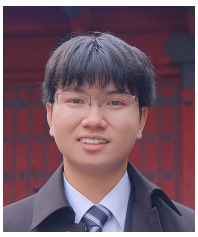
[110] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park, "One-step diffusion with distribution matching distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 6613–6623. 15

Jiapeng Tang is currently a Ph.D. student in the Department of Informatics at the Technical University of Munich. He received his B.E. and M.Sc. degrees from the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China, in 2014 and 2018, respectively. His research interests include computer vision and graphics, generative models, and deep learning, with a focus on 3D/4D reconstruction and generation.



Wei Cao is currently a Ph.D. student in the School of Information Sciences at the University of Illinois Urbana-Champaign. He received his B.Sc. and M.Sc. degrees from the School of Computation, Information and Technology, Technical University of Munich. His research interests lie in 3D/4D computer vision and autonomous driving.

Biao Zhang is a Postdoctoral researcher in KAUST. He received his B.S. and M.S. from Xi'an Jiaotong University and his Ph.D. from KAUST. His work explores connections between computer graphics and machine learning, with recent focus on generative models and representation learning. He has published in conferences including SIGGRAPH, CVPR, ICCV, ICLR, and NeurIPS.



Chang Luo is a Ph.D. student in the Department of Creative Informatics of Graduate School of Information, Science and Technology at the University of Tokyo. He received his Master degree in Informatics from the Technical University of Munich. His research interest lies in geometry processing and inverse rendering for computer graphics.

Yaoyao Liu is an assistant professor in the School of Information Sciences and the Coordinated Science Laboratory at the University of Illinois Urbana-Champaign. He is also affiliated with the Siebel School of Computing and Data Science, the National Center for Supercomputing Applications, and the Illinois Informatics. Previously, he completed his PhD in computer science at Max Planck Institute for Informatics and his BS in electronic information engineering at Tianjin University. His research interests include continual learning, few-shot learning, semi-supervised learning, generative models, 3D geometry models, and medical imaging. He is a recipient of the 2024 ECVA PhD Award.

Matthias Nießner is a Professor at the Technical University of Munich, where he leads the Visual Computing Lab. Before, he was a Visiting Assistant Professor at Stanford University. Prof. Nießner's research lies at the intersection of computer vision, graphics, and machine learning, where he is particularly interested in cutting-edge techniques for 3D reconstruction, semantic 3D scene understanding, video editing, and AI-driven video synthesis. In total, he has published over 150 academic publications, including 25 papers at the prestigious ACM Transactions on Graphics (SIGGRAPH / SIGGRAPH Asia) journal and 55 works at the leading vision conferences (CVPR, ECCV, ICCV); several of these works won best paper awards, including at SIGCHI'14, HPG'15, SPG'18, and the SIGGRAPH'16 Emerging Technologies Award for the best Live Demo.