Motion2VecSets: 4D Latent Vector Set Diffusion for Non-rigid Shape Reconstruction and Tracking

Wei Cao[1*‡] Chang Luo[1*] Biao Zhang[2] Matthias Nießner[1] Jiapeng Tang[1†]

[1]Technical University of Munich [2]King Abdullah University of Science and Technology

https://vveicao.github.io/projects/Motion2VecSets

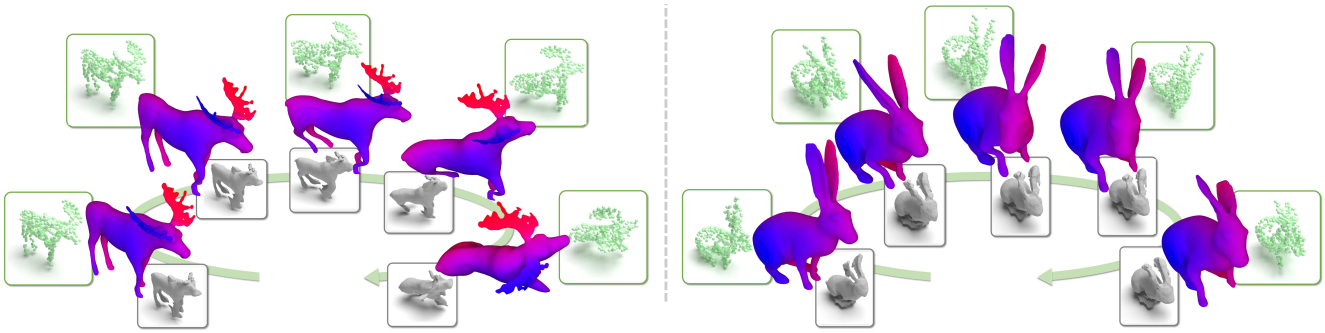

Figure 1. We present **Motion2VecSets**, a 4D diffusion model for dynamic surface reconstruction from sparse, noisy, or partial point cloud sequences. Compared to the existing state-of-the-art method CaDeX [25], our method can reconstruct more plausible non-rigid object surfaces with complicated structures and achieve more robust motion tracking.

Abstract

We introduce Motion2VecSets, a 4D diffusion model for dynamic surface reconstruction from point cloud sequences. While existing state-of-the-art methods have demonstrated success in reconstructing non-rigid objects using neural field representations, conventional feed-forward networks encounter challenges with ambiguous observations from noisy, partial, or sparse point clouds. To address these challenges, 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 enable more plausible and probabilistic reconstructions when handling ambiguous inputs. We parameterize 4D dynamics with latent sets instead of using global latent codes. This novel 4D representation allows us to learn local shape and deformation patterns, leading to more accurate non-linear motion capture and significantly improving generalizability to unseen motions and identities. For more temporally-coherent object tracking, we synchronously denoise deformation latent sets and exchange information across multiple frames. To avoid computational overhead, we designed an interleaved space and time attention block to alternately aggregate deformation latents along spatial and temporal domains. Extensive comparisons against state-of-the-art methods demonstrate the superiority of our Motion2VecSets in 4D reconstruction from various imperfect observations.

1. Introduction

Our world, dynamic in its 4D nature, demands an increasingly sophisticated understanding and simulation of our living environment. This offers significant potential for practical applications, including Virtual Reality (VR), Augmented Reality (AR), and robotic simulations. There have been notable advances in 3D object modeling, particularly in representations through parametric models [27, 30, 36, 45, 67]. Unfortunately, these template-based models are not effectively suited to capture the 4D dynamics of general non-rigid objects, due to the assumption of a fixed template mesh. Model-free approaches [25, 32, 52] represent a significant advance by using coordinate-MLP representations for deformable object reconstruction with arbitrary topologies and non-unified structures. However, these state-of-the-art methods still encounter challenges when facing ambiguous observations of noisy, sparse, or partial point clouds since it is an ill-posed problem where multiple pos-

*Equal Contribution.
†Corresponding author.
‡Work done during master's thesis.

sible reconstructions can match the input. In addition, 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 generalizability of global latent representation.

To address the above-mentioned challenges, we propose Motion2VecSets, a diffusion model designed for 4D dynamic surface reconstruction from sparse, noisy, or partial point clouds. It explicitly learns the joint distribution of non-rigid object surfaces and temporal dynamics through an iterative denoising process of compressed latent representations. This enables more realistic and varied reconstructions, particularly when dealing with ambiguous inputs. Inspired by the observation that objects with varying topologies often share similar local geometry and deformation patterns, we represent dynamic surfaces as a sequence of latent sets to preserve local shape and deformation details: one for shape modeling of the initial frame and others for describing the temporal evolution from the initial frame. This latent set representation naturally enables the learning of more accurate shape and motion priors, enhancing the model's generalization capacity to unseen identities and motions. Specifically, we introduce the Synchronized Deformation Vector Set Diffusion, which simultaneously denoises the deformation latent sets across different time frames to enforce spatio-temporal consistency over dynamic surfaces. To manage the memory consumption associated with multiple deformation latent set diffusions, we design an interleaved space and time attention block as the basic unit for the denoiser. These blocks aggregate deformation latent sets along spatial and temporal domains alternately. As illustrated in Figure 1, our Motion2VecSets can reconstruct more plausible non-rigid object surfaces with complicated structures and achieve more robust motion tracking than the state-of-the-art method. Our contributions can be summarized as follows:

- We present a 4D latent diffusion model designed for dynamic surface reconstruction, adept at handling sparse, partial, and noisy point clouds.
- We introduce a 4D neural representation with latent sets, significantly enhancing the capacity to represent complicated shapes as well as motions and improving generalizability to unseen identities and motions.
- We design an Interleaved Spatio-Temporal Attention mechanism for synchronized diffusion of deformation latent sets, achieving robust spatio-temporal consistency and advanced computational efficiency.

Extensive comparisons against state-of-the-art methods demonstrate the superiority of our Motion2VecSets in dynamic surface reconstruction on the Dynamic FAUST [3] and the DeformingThings4D-Animals [28] datasets.

2. Related works

3D Shapes Traditional methods in 3D representation have primarily used meshes [18, 29, 38, 49, 50, 59], point clouds [1, 15, 65], and voxels [12, 16, 19, 43, 48] to represent geometry. Complementing these are parametric models, which have effectively modeled specific shape categories, such as human bodies (e.g., SMPL [30], STAR [36]), faces (e.g., FLAME [27]), hands (e.g., MANO [45]), and animals (e.g., SMAL [67]). However, these parametric approaches often rely on fixed templates, which can result in difficulties accurately modeling general non-rigid objects without consistent topological structures. Meanwhile, recent advancements in 3D representation are increasingly using implicit methods [5, 6, 9, 10, 17, 32, 39, 51, 58, 62, 63], known for their greater flexibility to represent objects with arbitrary topologies. In particular, Occupancy Networks [32] and DeepSDF [39] employ a continuous implicit framework, enabling the representation of volumetric grids offering potentially infinite resolution.

4D Dynamics Recent advancements have successfully extended 3D representations to 4D, which more effectively captures object dynamics [4, 15, 23, 25, 26, 35, 37, 47, 52, 53, 66]. For example, OFlow [35] incorporates Neural-ODE [56] for simulating deformations. LPDC [52] replaces Neural-ODE with an MLP and learns local spatio-temporal codes, capturing both shape and deformations. CaDeX [25] employs a learnable deformation embedding between each frame and its canonical shape. However, methods relying on either ODE [35] solvers or a single global latent vector [25, 37, 52] coupled with an MLP network face challenges in capturing complex real-world 4D dynamics, particularly in non-rigid objects. Drawing inspiration from 3DShape2Vecset [63], which uses a set of latent codes to represent similar local geometry patterns across objects, our proposed method leveragesa similar approach for characterizing 4D dynamics. Different objects share similar local deformation patterns, our framework uniquely assigns a distinct learnable latent code to each local region, significantly enhancing their ability to precisely model and generalize to unseen identities and motions.

Diffusion Models Diffusion models [20], known for their Markov chain-based denoising capability, have made impressive progress in multiple tasks, including image and video processing [7, 14, 21, 33, 34], 3D vision [2, 11, 22, 31, 44, 54, 55, 57, 61]. These models are adept at capturing complex data distributions. In the field of 3D vision, their applications are varied: Luo et al. [31] have successfully applied diffusion models to point cloud generation, and Rombach *et al*. [2, 44] have adapted them for latent space representations. Additionally, integration with PointNet [41] and triplane features [40], as seen in DiffusionSDF [11], has further enhanced their training capabilities. Con-

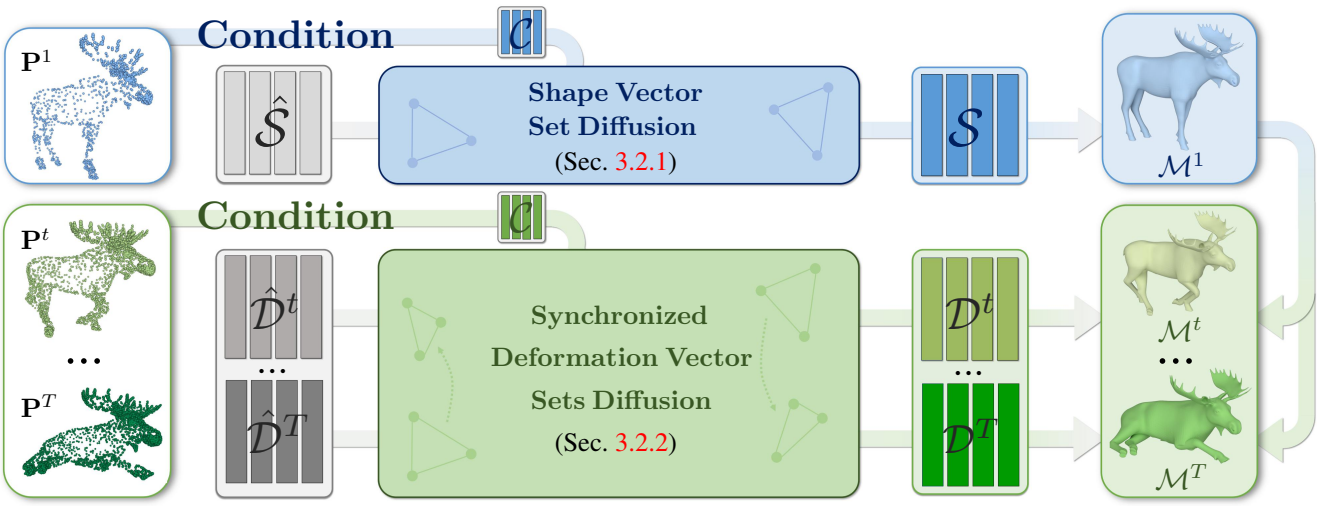

Figure 2. **Overview Pipeline of Motion2VecSets.** Given a sequence of sparse and noisy point clouds as inputs $\{\mathbf{P}^t\}_{t=1}^T$, Motion2VecSets outputs a continuous mesh sequence $\{\mathcal{M}^t\}_{t=1}^T$. The initial input frame $\mathbf{P}^1$ (top left) is used as a condition in the Shape Vector Set Diffusion, yielding denoised shape codes $\mathcal{S}$ for reconstructing the geometry of the reference frame $\mathcal{M}^1$ (top right). Concurrently, the subsequent input frames $\{\mathbf{P}^t\}_{t=2}^T$ (bottom left) are utilized in the Synchronized Deformation Vector Sets Diffusion to produce denoised deformation codes $\{\mathcal{D}^t\}_{t=2}^T$, where each latent set $\mathcal{D}^t$ encodes the deformation from the reference frame $\mathcal{M}^1$ to subsequent frames $\mathcal{M}^t$.

current work NAP [26] advances 3D object generation by effectively modeling articulated objects with a novel parameterization and diffusion-denoising approach. A key challenge in representing 4D dynamics with existing diffusion models is their tendency to adapt 3D models directly to 4D and process each frame independently, which can result in discontinuities in temporal and spatial relationships. To bridge this gap, our approach implements synchronous denoising processes for sets of codes. This innovation ensures not only a reduction in spatial complexity but also consistent deformations in latent space. Moreover, recent works [13, 22, 42, 46, 57, 60, 64] in the field of 3D pose estimation and generation also indicate the power of diffusion models. DiffPose [22] utilizes the diffusion model to handle very ambiguous poses and can even predict an infinite number of poses. PhysDiff [60] produces physically plausible motions by incorporating a physics-based motion projection within its diffusion process. However, these methods are still in the realm of pose, our method expands the application of diffusion models to a broader range of deformable surfaces of general non-rigid objects. Similar to the concurrent work DPHMs [54], our approach utilizes diffusion priors to facilitate robust 4D reconstruction from imperfect observations.

3. Approach

The inputs are T frames of sparse, partial, or noisy point clouds, represented by $\mathcal{P} = \{\mathbf{P}^t\}_{t=1}^T$, where $\mathbf{P}^t = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^L$, L represents the number of points. The goal is to reconstruct continuous 3D meshes with high fidelity, denoted as $\{\mathcal{M}^t\}_{t=1}^T = \{\mathcal{V}^t, \mathcal{F}^t\}_{t=1}^T$, where $\mathcal{V}^t$ and $\mathcal{F}^t$ refer to the set of vertices faces of the reconstructed mesh at time frame t. Conventional feed-forward models face challenges in handling ambiguous inputs within an ill-posed problem setting. Particularly, when observations are sparse, partial, and noisy, generating meaningful reconstructions becomes highly challenging without robust prior knowledge. To reconstruct high-fidelity dynamic shapes accurately, we proposed 4D latent set diffusion to learn shape and motion priors, explicitly learning the distribution of deformable object surface sequences via compressed latent vector sets. While the diffusion model enhances realistic surface reconstruction and deformation tracking, generating multi-modal outputs, the latent set representation and transformer architecture provide the capability to capture more accurate geometry and deformation priors.

3.1. 4D Neural Representation with Latent Sets

Previous works often utilize single global codes [25, 35, 52] to represent 4D sequences, potentially losing significant surface geometry and temporal evolution information. To retain as much detail as possible, we advocate the use of two distinct sets of latent vectors. Specifically, the *shape latent set* is responsible for reconstructing the initial frame, serving as the reference frame, while the deformation correspondences between the reference and subsequent frames are encoded by the *deformation latent set*. Compared with previous methods [25, 35, 52] relying on a single global code, we assign local latent codes to individual local regions, which significantly improves the network's capability to accurately model non-linear motions and generalize

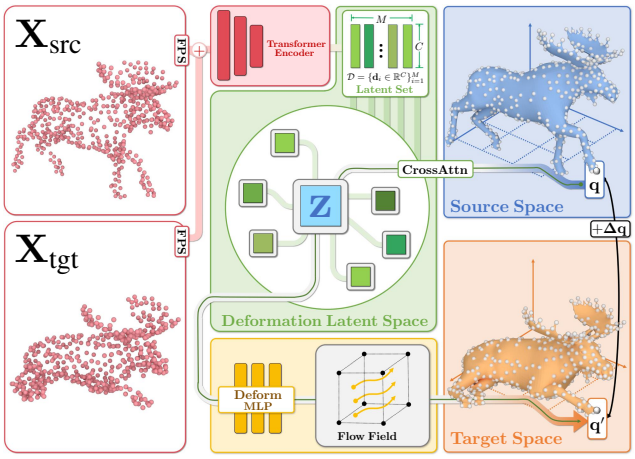

Figure 3. **Deformation Autoencoder.** Given a pair of point clouds $\mathbf{X}_{\text{src}}$ and $\mathbf{X}_{\text{tgt}}$ from two frames of a dynamic mesh sequence, we initially downsample them using farthest point sampling (FPS). Subsequently, the concatenated points are passed into transformer encoder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

to unseen identities and motions. Given that different non-rigid objects share similar local geometry and deformation patterns, the latent sets can also increase the generalization ability to handle unseen motions and identities.

Shape Latent Set Similar to 3DShape2VecSet [63], we utilize a shape autoencoder to compress the surface of the initial frame into a set of latent codes. Concretely, we leverage a transformer encoder that condenses the 3D surface of the initial frame into a set of latent vectors denoted as $\mathcal{S} = \{\mathbf{s}_i \in \mathbb{R}^C\}_{i=1}^M$. Here, M represents the overall count of codes and C denotes their dimensionality. Following that, a cross-attention layer is used to fuse the latent codes for occupancy field prediction through an MLP. Training involves minimizing the binary cross-entropy (BCE) loss, which aligns the predicted occupancy $\hat{\mathcal{O}}(\mathcal{Q})$ with the actual occupancy $\mathcal{O}(\mathcal{Q})$, $\mathcal{Q}$ refers to the query points:

$$\mathcal{L}_{\text{recon}}\left(\mathcal{S}, \mathcal{Q}\right) = \mathbb{E}_{\mathcal{Q}\in\mathbb{R}^3}[\text{BCE}(\hat{\mathcal{O}}, \mathcal{O})] \tag{1}$$

Deformation Latent Set As shown in Figure 3, to represent the deformation between different non-rigid poses, we first sample a pair of point clouds $\mathbf{X}_{\text{src}}$ and $\mathbf{X}_{\text{tgt}}$ of size N with same sampling indices from a mesh sequence. Then, we employ a uniform farthest point sampling (FPS) strategy to eliminate spatial redundancy while preserving point correspondence. This process facilitates a concatenation step, where we combine the original and downsampled pairs of point clouds $\{\mathbf{X}_{\text{src}}, \mathbf{X}_{\text{tgt}}\}$, respectively. The subsequent transformer encoder is applied to extract deformation details in the local regions around subsampled points, resulting in the deformation latent set denoted as $\mathcal{D} = \{\mathbf{d}_i \in \mathbb{R}^C\}_{i=1}^M$. Query point $\mathbf{q} \in \mathcal{Q}_{src}$ from the source space is utilized as the query for cross-attention, extracting the most relevant fused feature $\mathbf{z}$ in the deformation latent space. A linear deformation layer then maps these features to the predicted target points $\mathbf{q}'$ through a flow field. The correspondence loss calculates the ℓ_2-norm distance between the predicted and true target point clouds:

$$\mathcal{L}_{\text{corr}}\left(\mathcal{D}, \mathcal{Q}_{\text{src}}\right) = \mathbb{E}_{\mathcal{Q}_{\text{src}}\in\mathbb{R}^3}[\text{MSE}(\hat{\mathcal{Q}}_{\text{tgt}}, \mathcal{Q}_{\text{tgt}})] \tag{2}$$

KL Regularization Consistent with the latent diffusion framework [44], our model incorporates KL-regularization in the latent space to modulate feature diversity. This ensures the preservation of high-level features and keeps coherent global geometric and deformation patterns, which promotes the learning of diffusion models. In summary, we characterize a sequence through the shape latent set $\mathcal{S}^1$ of the initial reference frame, which describes the implicit surface, and deformation latent sets $\mathcal{D}^2, \mathcal{D}^3, ..., \mathcal{D}^T$ that depict the dense correspondences between the initial reference frame and the subsequent frames.

3.2. 4D Latent Set Diffusion

3.2.1 Shape Diffusion

Following the diffusion paradigm in EDM by Karras et al. [24], 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 $\mathbb{R}^{N\times C}$). The whole process is denoted as:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}\left\|\text{Denoiser}\left(\hat{\mathcal{S}}, \sigma, \mathcal{C}\right) - \mathcal{S}\right\|_2^2 \tag{3}$$

Here, σ represents the noise level. Conditional latent set $\mathcal{C}$ is defined as $\mathcal{C}(\mathbf{P}^1) = \{c_i \in \mathbb{R}^C\}_{i=1}^M$, which is generated by sending the first input frame $\mathbf{P}^1$ to the conditional encoder.

3.2.2 Synchronized Deformation Diffusion

To adapt these 3D models [27, 36, 45, 67] 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 point clouds, which would significantly increases the time complexity to $O(T^2N^2)$ for a sequence of T frames and N points each. However, our 4D latent set representation allows us to bypass the need for brute-force attention across spatial and temporal domains. As discussed in Sec. 3.1, the deformation latents at identical spatial positions across different frames correspond to the deformation behaviors of the same local surface region. Leveraging this property, we implement an alternating aggregation approach for the latent features, systematically switching

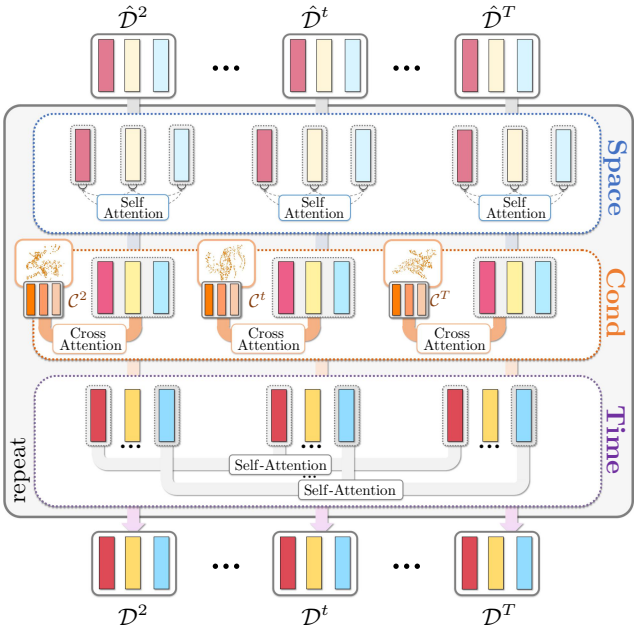

Figure 4. **Synchronized Deformation Vector Set Diffusion.** Given noised deformation vector sets $\{\hat{\mathcal{D}}^t\}_{t=2}^T$ (top) from a sequence, each set denoted as $\hat{\mathcal{D}}^t = \{\hat{\mathbf{d}}_1^t, ..., \hat{\mathbf{d}}_M^t\}$ of timestep $t \in [2, T]$, we use repeated Interleaved 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 sparse or partial point clouds via cross-attention (Condition Attention) between conditional codes $\mathcal{C}^t$ and noised deformation codes $\hat{\mathcal{D}}^t$ at each frame. 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.* $\{\hat{\mathbf{d}}_i^t\}_{t=2}^T$. 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.

between the spatial and temporal domains. This method not only enhances efficiency but also preserves the accuracy of our model, leading to a reduction in time complexity to $O(TN^2)$ in the spatial domain and $O(NT^2)$ in the temporal domain. The details of synchronized deformation diffusion are described as follows. Given sparse input point clouds $\mathcal{P} = \{\mathbf{P}^t\}_{t=1}^T$, we pair subsequent frames with the first reference frame $\mathbf{P}^1$, *i.e.*, $\{\mathbf{P}^1, \mathbf{P}^t\}_{t=2}^T$. These pairs are encoded into a series of conditional latents $\mathcal{C}^t(\mathbf{P}^1, \mathbf{P}^t) = \{\mathbf{c}_i \in \mathbb{R}^C\}_{i=1}^M$ via a transformer encoder. Then these conditional latents, together with the diffused shape latent set $\mathcal{S}^1$ in Sec. 3.2.1, are injected into the denoising network as the condition providing guidance for the network to handle ambiguous scenarios, like partial observation.

Interleaved Spatio-Temporal Attention Figure 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). Each ISTA block contains three attention layers: *Space Self-Attention Layer*, *Conditional Cross-Attention Layer* and *Time Self-Attention Layer*. The *Space Self-Attention Layer* initiates spatial information exchange within each set of noised deformation codes $\hat{\mathcal{D}}^t = \{\hat{\mathbf{d}}_i^t\}_{i=1}^M$:

$$\mathbf{SpaceAttn} = \text{SelfAttn}(\{\hat{\mathbf{d}}_i^t\}_{i=1}^M) \tag{4}$$

This is then followed by the *Conditional Cross-Attention Layer*. Conditional codes $\mathcal{C}^t(\mathbf{P}^1, \mathbf{P}^t) = \{\mathbf{c}_i \in \mathbb{R}^C\}_{i=1}^M$ from a partial or sparse point cloud are subjected to cross-attention with $\mathbf{CondAttn}$:

$$\mathbf{CondAttn} = \text{CrossAttn}(\{\hat{\mathbf{d}}_i^t\}_{i=1}^M, \mathcal{C}^t) \tag{5}$$

Finally, to improve coherence in the time dimension, a *Time Self-Attention Layer* is implemented among deformation codes from different timesteps but from the same position (same index i but different t). Consequently, through this setup, the $\mathbf{TimeAttn}$ is effectively obtained:

$$\mathbf{TimeAttn} = \text{SelfAttn}(\{\hat{\mathbf{d}}_i^t\}_{t=2}^T) \tag{6}$$

In the denoising phase, we regard the entire sequence of deformation codes as a unified entity and apply a uniform noise reduction strategy across all codes, which preserves the consistency of local deformation patterns. Contrary to assigning individual noise to each set of shape codes, we add a consistent uniform noise ϵ to the deformation codes of the entire sequence $\{\hat{\mathcal{D}}^t\}_{t=2}^T = \{\mathcal{D}^t\}_{t=2}^T + \epsilon$. The denoising objective is thus formulated as:

$$\mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \left\| \text{Denoiser}\left(\{\hat{\mathcal{D}}^t\}_{t=2}^T, \sigma, \mathcal{C}\right) - \{\mathcal{D}^t\}_{t=2}^T \right\|_2^2 \tag{7}$$

Here, $\mathcal{C}$ represents conditional codes derived from observations, $\{\mathcal{D}^t\}_{t=2}^T$ can also be represented in its 3D matrix form as $\mathbb{R}^{M \times (T-1) \times C}$. This approach not only ensures uniformity in the denoising process but also significantly reduces computational overhead.

4. Experiments

Datasets: We conducted experiments on two 4D datasets. The first, Dynamic FAUST (D-FAUST) [3], focuses on human body dynamics, including 10 subjects and 129 sequences. It is split into training (70%), validation (10%), and test (20%) subsets, following [35]. The second, DeformingThings4D-Animals (DT4D-A) [28], includes 38 identities with a total of 1227 animations, divided into training (75%), validation (7.5%), and test (17.5%) subsets as [25]. The training and validation sets use motion sequences of seen individuals. The test set is divided into two parts: unseen motions and unseen individuals.

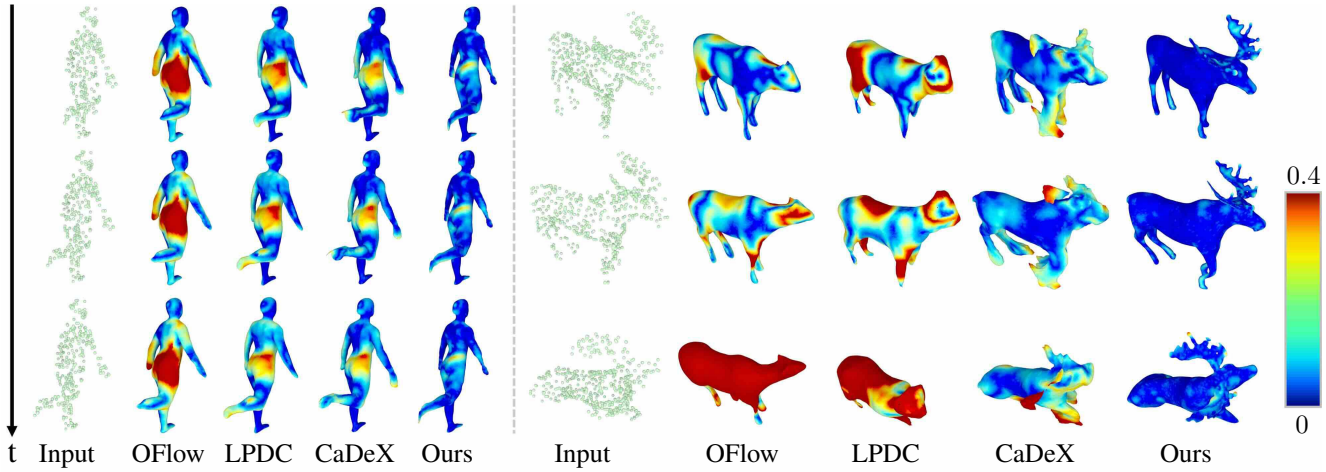

Figure 5. Comparisons of 4D Shape Reconstruction from **sparse and noisy** point clouds on the D-FAUST [3] (left) and the DT4D-A [28] (right) datasets. We visualize the Chamfer Distance between reconstruction and ground-truths as error maps. Our method can reconstruct more accurate surface geometries and motion dynamics.

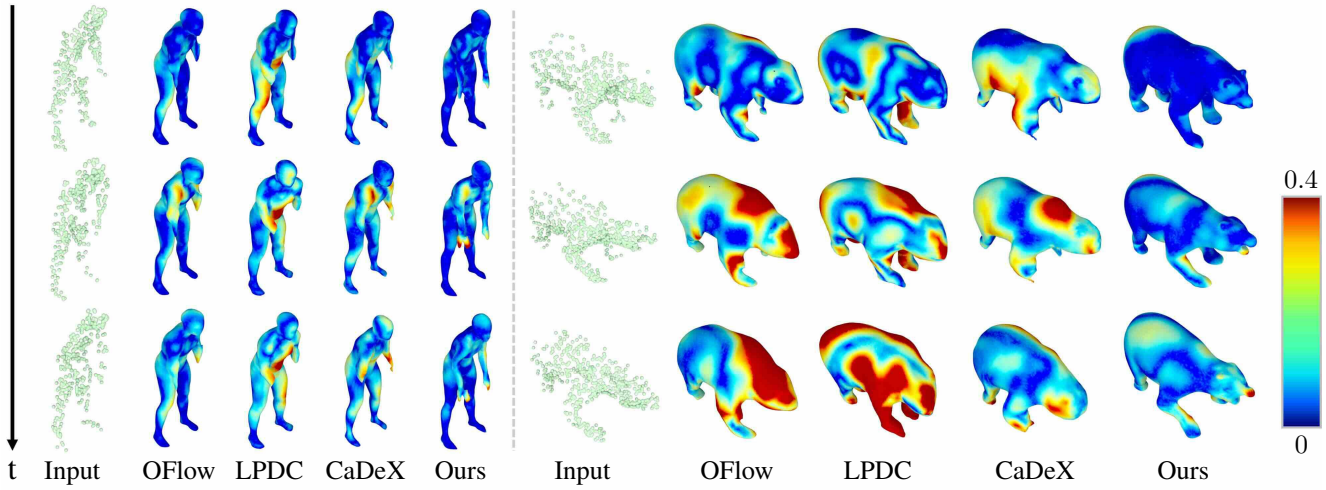

Figure 6. Comparisons of 4D Shape Completion from **monocular noisy depth scans** on D-FAUST [3] (left) and DT4D-A [28] (right) datasets. Our method exhibits lower reconstruction errors and achieves more plausible tracking.

Baselines: We compare against state-of-the-art methods in 4D reconstruction, including OFlow [35], LPDC [52], CaDex [25]. **OFlow** assigns each 4D point both an occupancy value and a motion velocity vector, utilizing a Neural-ODE framework [8] for learning deformations. **LPDC** employs a 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 overlap between predicted and ground truth meshes; The Chamfer distance calculates the average nearest-neighbor distance between two point sets; ℓ_2-distance error measures the Euclidean distance between corresponding points on the predicted and ground truth meshes.

Implementations: The training of our approach consists of two stages. The first stage involves two auto-encoders. The input point clouds ($N = 2048$) are randomly sampled from object surfaces and near-surface regions. For the shape auto-encoder, the learning rate is 10^{-4}, with KL-divergence loss weights 10^{-3}. For the deformation auto-encoder, the learning rate is 10^{-4}, with KL-divergence loss weights 10^{-6}. They are trained for 100 epochs with batch size 24. The second stage is the diffusion models, the learning rate for both shape and deformation diffusion models is 10^{-4} and they are trained for 50 epochs with a batch size of 8 for shape diffusion and 4 for deformation diffusion.

Input	Method	Unseen Motion			Unseen Individual		
		IoU ↑	CD ↓	Corr ↓	IoU ↑	CD ↓	Corr ↓
DT4D-A [28]	OFlow [35]	70.6%	0.104	0.204	57.3%	0.175	0.285
	LPDC [52]	72.4%	0.085	0.162	59.4%	0.149	0.262
	CaDex [25]	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 [3]	OFlow [35]	81.5%	0.065	0.094	72.3%	0.084	0.117
	LPDC [52]	84.9%	0.055	0.080	76.2%	0.071	0.098
	CaDex [25]	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**

Table 1. Quantitative comparisons of 4D Shape Reconstruction from **sparse and noisy** point cloud sequences on the DT4D-A [28] and the D-FAUST [3] datasets.

Input	Method	Unseen Motion			Unseen Individual		
		IoU ↑	CD ↓	Corr ↓	IoU ↑	CD ↓	Corr ↓
DT4D-A [28]	OFlow [35]	64.2%	0.305	0.423	55.1%	0.408	0.538
	LPDC [52]	62.2%	0.339	0.427	51.6%	0.467	0.488
	CaDex [25]	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 [3]	OFlow [35]	76.9%	0.084	0.165	66.4%	0.109	0.194
	LPDC [52]	68.3%	0.138	0.167	59.6%	0.156	0.204
	CaDex [25]	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**

Table 2. Quantitative comparisons of 4D Shape Completion from **monocular noisy depth scans** on the DT4D-A [28] and the D-FAUST [3] datasets.

Runtime: The training takes about 60 hours(2*RTX 4090) . The inference takes about 11s for 17 frames(1*RTX 3080).

4.1. 4D Shape Reconstruction

We initially assessed our models' ability for 4D reconstruction from sparse and noisy point cloud sequences Consistent with the setup in OFlow [35], our network processed sequences of $T = 17$ continuous frames. Each frame represented a sparse point cloud, with $L = 300$ for D-FAUST [3] or $L = 512$ for DT4D-A [28]. We also simulate noisy observations with Gaussian noise (σ = 0.05).

Quantitatively, our model demonstrates superior performance over state-of-the-art models on the D-FAUST [3] and DT4D-A [28] datasets, as detailed in Tab. 1. This superiority is particularly notable in the unseen individual category of the DT4D-A dataset, which features more diverse topologies from various animals. Additionally, both chamfer distance and ℓ_2-correspondence error are reduced to less than half of those recorded by the previous state-of-the-art methods. Qualitatively, as illustrated in Fig. 6, our model outperforms in reconstructing complete shapes and minimizing chamfer distance errors, particularly in capturing fast-moving structures like feet of humans and heads of animals.

The superiority of our model is attributed to the proposed 4D latent set diffusion, enabling a more precise capture of local geometry and deformation patterns. Methods like LPDC [52] and OFlow [35] perform well in human settings thanks to similar human topologies, while CaDex [25] benefits from canonical shape learning. However, the diverse topologies and scales in animal setup, such as dragons, present a significant challenge for models that optimize global codes. Our approach, in contrast, effectively captures these complex 4D dynamics of general non-rigid objects.

4.2. 4D Shape Completion

4.2.1 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 point cloud input and the frame length are the same as Sec. 4.1. The qualitative and quantitative comparisons are presented in Fig. 6 and Tab. 2.

As seen, our method consistently outperforms all state-of-the-art methods in all metrics and produces more complete surface geometries with more plausible motion tracking. This demonstrates the effectiveness of the motion priors learned by our proposed 4D latent set diffusion in addressing ambiguous data such as partial observations.

4.2.2 Partial Scan Sequences

To assess the robustness of our method to extremely ambiguous data, we set up a challenging experiment on the D-FAUST [3] 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. We adopt the same configuration as Sec. 4.1, with a frame length ($T = 17$) and input point cloud size ($L = 300$). As the Fig. 7 shows, OFlow [35], LPDC [52], and CaDex [25] 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 coherent tracking. Additionally, our approach present 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.

4.3. Ablation Study

We conducted ablation studies to validate the effectiveness of each component (See Tab. 3, Fig. 8) under the setting of 4D Shape Completion from **monocular noisy depth scans** on the D-FAUST [3] 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 sub-optimal results. We provide comparisons against the variant of one-step regression without diffusion models. As shown in Fig. 8 and Tab. 3, diffusion model uses a probabilistic way to deal with highly ambiguous inputs and generate plausible predictions. Also, diffu-

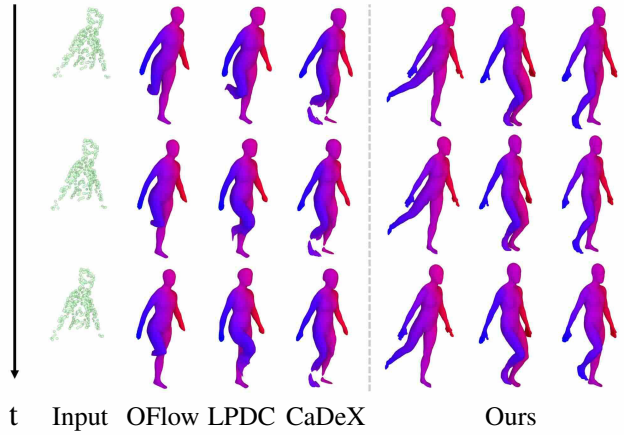

Figure 7. Comparisons of 4D Shape Reconstruction from **highly partial** point cloud sequences, such as half-body scans obtained from the D-FAUST [3] dataset. The colors of the meshes encode the correspondence. Our diffusion-based method produces highly complete human shapes with more favorable motions, offering multiple possible outputs that match the input observations.

sion models can handle "one-to-many" problems and generate diverse and creative outputs as shown in Fig. 7.

What is the effect of latent vector set representation? Instead of using single global latent code, our approach employs 4D latent vector sets. As indicated in Tab. 3, our method significantly outperforms the global latent codes (with $M = 1$) and captures more accurate 4D motions. It becomes more apparent in unseen identities, demonstrating an enhanced generalization ability.

What is the effect of time attention layers? For the synchronized deformation latent set diffusion, we have integrated the time self-attention layer in our interleaved spatio-temporal attention mechanism. We attempted to remove the layer. However, the results showed a decrease in all metrics, highlighting the effectiveness of the time self-attention layer in maintaining temporal coherence.

What is the effect of the number of channels of latent set? To find out the optimal configuration for learning shape and deformation priors within time-varying deformable surfaces, we tried the channel numbers C of the shape and deformation latent sets. The experimental results indicated that using $C = 32$ channels for 4D latent set diffusion is more suitable, yielding more favorable results.

5. Conclusion

We present Motion2VecSets, a 4D diffusion model for dynamic surface reconstruction from point cloud sequences. Our method explicitly models the shape and motion distributions of non-rigid objects through an iterative denoising process, using compressed latent sets to generate plausible and diverse outputs. The learned shape and motion diffusion priors can effectively deal with ambiguous observations, including sparse, noisy, and partial data. Compared to encoding shape and deformation with a global latent, our novel 4D latent set representation enables more accurate non-linear motion capture and improves the generalizability to unseen identities and motions. The designed interleaved space and time attention block for synchronized deformation vector sets diffusion enforces temporal-coherent object tracking while reducing computational overhead. Extensive experiments demonstrate our approach's superiority in reconstructing sparse, partial, and even half-body point clouds on the D-FAUST [3] and DT4D-A [28] datasets, underlining its robustness to various types of imperfect observations. We believe that Motion2VecSets has the potential for future extension into multi-modal domains, such as text-driven 4D generation and RGB video-based 4D reconstruction.

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
$C = 8$	78.9%	0.078	0.180	68.0%	0.105	0.225
$C = 16$	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	**83.8%**	**0.054**	**0.111**	**74.4%**	**0.075**	**0.140**

Table 3. Quantitative ablation studies on the D-FAUST [3] dataset. M denotes the number of latent codes and C represents the number of latent code channels.

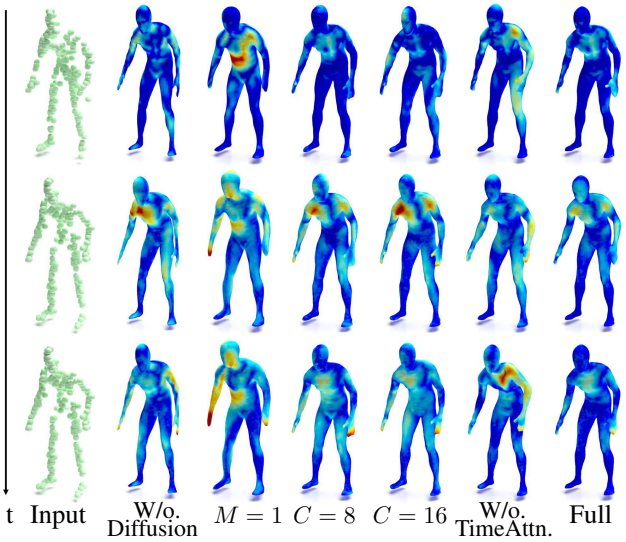

Figure 8. Qualitative ablation studies on the D-FAUST [3] dataset.

Acknowledgement. This work was supported by the ERC Starting Grant Scan2CAD (804724) as well as the German Research Foundation (DFG) Research Unit "Learning and Simulation in Visual Computing". We further thank Angela Dai for the video voice-over.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J Guibas. Learning representations and generative models for 3d point clouds. *arXiv preprint arXiv:1707.02392*, 2017. 2

[2] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022. 2

[3] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 5, 6, 7, 8

[4] Aljaž Božič, Pablo Palafox, Michael Zollhöfer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. *CVPR*, 2021. 2

[5] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard 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*, pages 608–625. Springer, 2020. 2

[6] Chao Chen, Yu-Shen Liu, and Zhizhong Han. Gridpull: Towards scalability in learning implicit representations from 3d point clouds. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[7] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020. 2

[8] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018. 6

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[10] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 2

[11] Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-sdf: Conditional generative modeling of signed distance functions. 2023. 2

[12] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio 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. 2

[13] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9760–9770, 2023. 3

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[15] Haoqiang Fan, Hao Su, and Leonidas 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*, pages 605–613, 2017. 2

[16] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pages 402–411, 2017. 2

[17] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020. 2

[18] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[19] 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)*, pages 85–93, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2

[22] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15977–15987, 2023. 2, 3

[23] Boyan Jiang, Yinda Zhang, Xingkui Wei, Xiangyang Xue, and Yanwei Fu. Learning compositional representation for 4d captures with neural ode. In *CVPR*, 2021. 2

[24] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022. 4

[25] Jiahui Lei and Kostas 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. 1, 2, 3, 5, 6, 7

[26] Jiahui Lei, Congyue Deng, Bokui Shen, Leonidas Guibas, and Kostas Daniilidis. Nap: Neural 3d articulation prior, 2023. 2, 3

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

[28] Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4dcomplete: Non-rigid motion estimation beyond the observable surface. *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 5, 6, 7, 8

[29] Yiyi Liao, Simon Donné, and Andreas Geiger. Deep marching cubes: Learning explicit surface representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

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

[31] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[32] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[33] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091*, 2021. 2

[34] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[35] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *International Conference on Computer Vision*, 2019. 2, 3, 5, 6, 7

[36] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pages 598–613, 2020. 1, 2, 4

[37] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. *arXiv preprint arXiv:2104.00702*, 2021. 2

[38] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9964–9973, 2019. 2

[39] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[40] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[41] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016. 2

[42] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[43] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4

[45] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 1, 2, 4

[46] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14725–14737, 2023. 3

[47] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 2

[48] David Stutz and Andreas 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. 2

[49] Jiapeng Tang, Xiaoguang Han, Junyi Pan, Kui Jia, and Xin 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)*, 2019. 2

[50] Jiapeng Tang, Xiaoguang Han, Mingkui Tan, Xin Tong, and Kui Jia. Skeletonnet: A topology-preserving solution for learning mesh reconstruction of object surfaces from rgb images. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6454–6471, 2021. 2

[51] Jiapeng Tang, Jiabao Lei, Dan Xu, Feiying Ma, Kui Jia, and Lei Zhang. Sa-convonet: Sign-agnostic optimization of convolutional occupancy networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6504–6513, 2021. 2

[52] Jiapeng Tang, Dan Xu, Kui Jia, and Lei 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*, pages 6022–6031, 2021. 1, 2, 3, 6, 7

[53] Jiapeng Tang, Lev Markhasin, Bi Wang, Justus Thies, and Matthias Nießner. Neural shape deformation priors. *Advances in Neural Information Processing Systems*, 35: 17117–17132, 2022. 2

[54] Jiapeng Tang, Angela Dai, Yinyu Nie, Lev Markhasin, Justus Thies, and Matthias Nießner. Dphms: Diffusion parametric head models for depth-based tracking. 2024. 2, 3

[55] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[56] Gerald Teschl. *Ordinary differential equations and dynamical systems*. American Mathematical Soc., 2012. 2

[57] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 3

[58] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Carsten Stoll, and Christian Theobalt. PatchNets: Patch-Based Generalizable Deep Implicit 3D Shape Representations. *European Conference on Computer Vision (ECCV)*, 2020. 2

[59] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 2

[60] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16010–16021, 2023. 3

[61] Biao Zhang and Peter Wonka. Functional diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[62] Biao Zhang, Matthias Niessner, and Peter Wonka. 3DILG: Irregular latent grids for 3d generative modeling. In *Advances in Neural Information Processing Systems*, 2022. 2

[63] Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3DShape2VecSet: A 3d shape representation for neural fields and generative diffusion models. *ACM Trans. Graph.*, 42(4), 2023. 2, 4

[64] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[65] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021. 2

[66] Kaifeng Zou, Sylvain Faisan, Boyang Yu, Sébastien Valette, and Hyewon Seo. 4d facial expression diffusion model. *arXiv preprint arXiv:2303.16611*, 2023. 2

[67] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 4