

Learning Parallel Dense Correspondence from Spatio-Temporal Descriptors for Efficient and Robust 4D Reconstruction

Jiapeng Tang^{1,4}, Dan Xu², Kui Jia^{*1,5,6}, and Lei Zhang^{3,4}

¹School of Electronic and Information Engineering, South China University of Technology

²Department of Computer Science and Engineering, HKUST, HK

³Department of Computing, The Hong Kong Polytechnic University, HK

⁴DAMO Academy, Alibaba Group

⁵Pazhou Lab, Guangzhou, China

⁶Peng Cheng Laboratory, Shenzhen, China

msjptang@mail.scut.edu.cn, danxu@cse.ust.hk, kuijia@scut.edu.cn, cslzhang@comp.polyu.edu.hk

Abstract

This paper focuses on the task of 4D shape reconstruction from a sequence of point clouds. Despite the recent success achieved by extending deep implicit representations into 4D space [29], it is still a great challenge in two respects, i.e. how to design a flexible framework for learning robust spatio-temporal shape representations from 4D point clouds, and develop an efficient mechanism for capturing shape dynamics. In this work, we present a novel pipeline to learn a temporal evolution of the 3D human shape through spatially continuous transformation functions among cross-frame occupancy fields. The key idea is to parallelly establish the dense correspondence between predicted occupancy fields at different time steps via explicitly learning continuous displacement vector fields from robust spatio-temporal shape representations. Extensive comparisons against previous state-of-the-arts show the superior accuracy of our approach for 4D human reconstruction in the problems of 4D shape auto-encoding and completion, and a much faster network inference with about 8 times speedup demonstrates the significant efficiency of our approach. The trained models and implementation code are available at <https://github.com/tangjiapeng/LPDC-Net>.

1. Introduction

We are surrounded by spatio-temporally changing environments that consist of various dynamics, such as observer

*Corresponding author

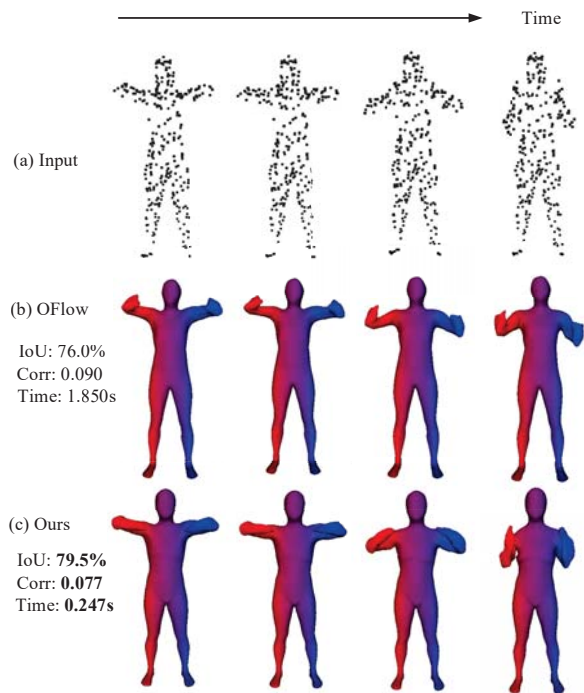


Figure 1: Given a sequence of 3D point clouds sampled in space and time, our goal is to reconstruct time-varying surfaces with dense correspondences. Compared to the state-of-the-art, i.e. OFlow [29], our approach can obtain more accurate geometries (higher IoU), better coherence (lower correspondence error) while supporting 8x faster inference.

movements, object motions, and human articulations. Reconstructing the human bodies evolving over time is vital

for various application scenarios such as robot perception, autonomous driving, and virtual/augmented reality.

Traditional works have achieved varying degrees of success in learning 4D reconstruction (*i.e.* 3D reconstruction along time) from a temporal sequence of point clouds, they are faced with various restrictions including the requirement of an expensive template mesh [1, 15, 41, 16, 18] or the dependence on smooth and clean inputs in space and time [43]. To overcome these issues, OccFlow [29] proposes a learning-based 4D reconstruction framework that establishes dense correspondences between occupancy fields by calculating the integral of a motion vector field defined in space and time to implicitly describe the trajectory of a 3D point. Although impressive results have been achieved, there are still several inherent limitations in this framework. Firstly, its spatial encoder does not take into account the aggregation of shape properties from multiple frames, which degrades the capability to recover accurate surface geometries. In addition, its temporal encoder ignores the time information which is of great importance to capture the temporal dynamics. Secondly, the integral of estimated immediate results leads to accumulated prediction errors in the temporal continuity and the reconstructed geometries. Lastly, it demonstrates low computational efficiency during training and inference because of the demanding computations of solving complex neural ordinary differential equations [6] to sequentially calculate the trajectories of points over time.

To tackle the above-mentioned problems, we aim to design a novel framework for 4D shape reconstruction from spacetime-sampled point clouds, to advance the 4D reconstruction from computational efficiency, accurate geometry, and temporal continuity. Our key idea is a mechanism which *parallelly* establishes the dense correspondence among different time-step occupancy fields predicted from the learned robust spatio-temporal shape representations. A high-level design of our proposed approach is a combination of static implicit field learning and dynamic cross-frame correspondence predicting. The former one focuses on occupancy field predictions from a novel spatio-temporal encoder that can effectively aggregate the shape properties with the temporal evolution to improve the robustness of geometry reconstructions. The latter one is utilized to identify the accurate correspondences within cross-frame occupancy fields, which are produced from representative embeddings describing the spatio-temporal changes in an efficient manner. The key to achieving this goal is a strategy of simultaneously learning occupancy field transformation from a first time step to others. It can help to remarkably reduce the convergence time in the network training, because of the bypassing of the expensive computation caused by solving ordinary differential equations. Moreover, benefiting from the advantages of parallel isosurface deformations

for the different time steps, our method provides a significant speed-up of the inference time. As shown in Fig. 1, we can achieve more robust surface reconstructions and more accurate correspondence prediction while allowing for considerably faster inference.

The main contribution can be summarized as follow:

- We propose a learning framework of modeling the temporal evolution of the occupancy field for 4D shape reconstruction, which is capable of capturing accurate geometry recoveries and coherent shape dynamics.
- We develop a novel strategy of establishing cross-frame shape correspondences by paralleling modeling occupancy field transformations from the first frame to others, which significantly improves the network computation efficiency, especially in the inference stage.
- We propose a novel 4D point cloud encoder design that performs efficient spatio-temporal shape properties aggregation from 4D point cloud sequences, which improves the robustness of reconstructed geometries.

Extensive ablation studies are conducted to validate the effectiveness of our proposed module designs. Comparisons against previous state-of-the-arts on the challenging D-FAUST dataset demonstrate the superior accuracy and efficiency of our approach in the problems of 4D shape auto-encoding and completion.

2. Related Work

In this section, we review the closely related works from three aspects as follows.

3D Shape Reconstruction The commonly used shape representations include voxel [8, 12], octree [34, 40, 45, 13], point cloud [9], mesh [11, 44, 17, 30, 38], implicit field [7, 24, 31, 47, 32, 48, 50], and hybrid representations [38, 39, 33]. Especially, the implicit representations [7, 24, 31, 47, 32], which implicitly represent a 3D surface by a continuous function, have attracted much attention. The continuity enables more accurate surface recoveries and volumetric reconstruction with infinite resolution. To inherit the advantages, we propose to extend the implicit representation for 4D reconstruction. Instead of independently extracting a surface mesh from the implicit field of each frame in the input sequence, which typically leads to slow inference and non-consistent topologies, we avoid these issues by the dense correspondence modeling which propagates the extracted surface mesh from the initial state to others.

Dynamic 4D Reconstruction Traditional works [20, 25, 26, 28, 36] utilize multi-view geometry to tackle the dynamic scene reconstruction problem from videos captured by multiple cameras. In contrast to them, we aim at

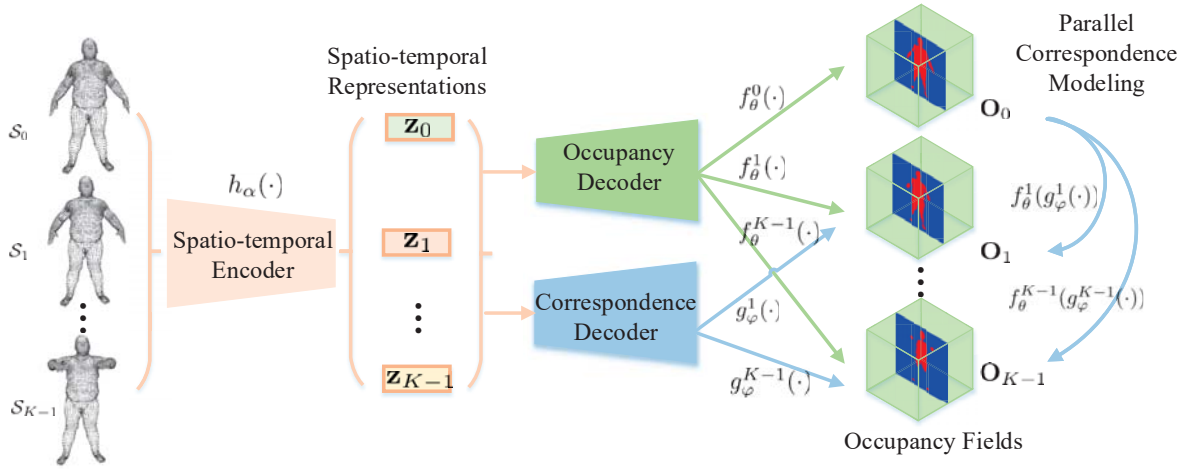


Figure 2: **Model Overview.** The proposed model first inputs a 3D point cloud sequence $\{S_0, S_1, \dots, S_{K-1}\}$ into a designed spatio-temporal encoder which extracts latent representations $\{z_0, z_1, \dots, z_{K-1}\}$. And then, the representations go through two separate decoders, *i.e.* the occupancy and the correspondence decoder. The occupancy decoder targets predicting the occupancy fields O_0, O_1, \dots, O_{K-1} in each frame. Finally, the correspondence decoder is used to parallelly model the correspondence between O_0 of the first frame and O_1, \dots, O_{K-1} of others.

4D shape reconstruction from a sequence of dynamically scanned point clouds, while the existing works with similar settings are faced with various limitations, including a heavy dependence on spatio-temporally smooth inputs [43] or the requirement of expensive template meshes [1, 15, 41, 16, 18]. Compared to OccFlow [29], instead of predicting the motion vectors for points in space and time and relying on the solver of Neural ODE [6] to calculate their 3D trajectories, we directly model the movements of points, which decreases the computation overhead during training. And the supporting of parallel surface deformation at different time steps remarkably accelerates the inference speed. Besides, we design a unified spatio-temporal encoder to effectively capture temporal dynamics and utilize the important time information in learning spatio-temporal descriptors.

Shape Correspondence Modeling Modeling point-to-point correspondence between two 3D shapes [3, 42, 37] is a well-studied area in computer vision and graphics. Our goal of modeling time-varying occupancy fields is closely related to deformation field-based methods [23, 27]. However, most of these works only define vector fields on the surfaces rather than in the whole 3D space as us. Eisenberger *et al.* [35] choose to model the evolution of the signed distance field to implicitly yield correspondences. They optimize an energy function in the evolution equation to impose similarity relationships of the Laplacian eigenfunction representations between the input and target shapes. However, we learn the dense correspondences between time-varying occupancy fields based on an intuitive observation, namely that the occupancy values of points are always invariant along the temporal evolution.

3. Approach

In this section, we first formulate the dynamic 4D surface reconstruction problem as the following. We consider as input a sequence of potentially incomplete, noisy 3D point clouds (of human bodies, easily captured by depth sensors). The observation of each frame can be represented as a point set $\mathcal{P} = \{\mathbf{p}_i \in \mathbb{R}^3 = \{x_i, y_i, z_i\} | i = 0, 1, \dots\}$. For a sequence of K frames with possibly non-uniform time intervals, we consider it as a 4D spatio-temporal point cloud donated by $\mathcal{S} = \{S_k\}_{k=0}^{K-1}$, where $S_k = \{s_i \in \mathbb{R}^4 = \{x_i, y_i, z_i, t_k\}_{i=0}^{N_k-1}\}$. N_k is the number of points at frame $k \in [0, K-1]$ and at time $t_k \in [t_0, t_{K-1}] \subset \mathbb{R}$ with $N = \sum_{k=1}^K N_k$. Our goal is to reconstruct time-varying 3D surfaces with accurate geometry, temporal coherence and fast inference. We achieve this by (1) developing our model based on the implicit surface representation which has been demonstrated the impressive capacity of capturing complex object geometries [31, 47, 2, 46, 10] (2) capturing the robust spatio-temporal shape properties by efficiently fusing information from each frame and taking the time information into consideration to obtain accurate temporal dynamics (3) parallelly modeling the dense correspondences between cross-time occupancy fields, which facilitates parallel surface deformations from the first to other time steps to accelerate the inference speed.

Overview The overall pipeline is shown in Fig. 2. It is composed of three key components which are respectively responsible for spatio-temporal representation learning, occupancy fields predicting, and dense correspondences modeling. We firstly process the input (*i.e.* a point cloud se-

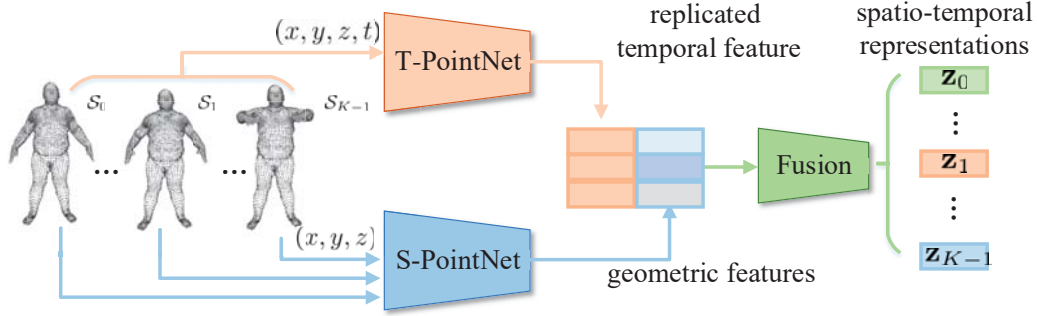


Figure 3: **Spatio-Temporal Encoder**. It contains a T-PointNet branch that utilizes the entire 4D point cloud $\mathcal{S} = \{\mathcal{S}_k\}_{k=0}^{K-1}$ to extract a temporal representation by treating time explicitly and equally to each spatial dimension. It also uses a S-PointNet branch to extract a sequence of geometric representations by individually applying it for each frame \mathcal{S}_k without considering timestamps. Finally, the geometric and temporal representations are fused into a sequence of spatio-temporal descriptors $\{\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_{K-1}\}$, to aggregate shape properties and explore dynamics variations.

sequence \mathcal{S}) through a spatio-temporal encoder $h_\alpha(\cdot)$ to get a sequence of latent embeddings $\{\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_{K-1}\}$ encoding the geometric shape properties and the temporal changes. Then we learn the occupancy field $f_\theta^k(\cdot)$ at $t = t_k$ using a shared decoder $f_\theta : \mathbb{R}^3 \rightarrow [0, 1]$ conditioned on a time-specific latent embedding \mathbf{z}_k . And a correspondence decoder $g_\varphi : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is utilized to simultaneously estimate continuous displacement vector fields to model occupancy field evolutions from initial to future time steps. More specifically, it establishes dense correspondences between $f_\theta^0(\cdot)$ and $f_\theta^k(\cdot)$ by learning a function $g_\varphi^k(\cdot)$ that transforms 3D spatial points at time t_0 into coordinate system at time t_k conditioned on the associate embeddings \mathbf{z}_0 and \mathbf{z}_k . So the occupancy field at time t_k can be predicted through $f_\theta^k(g_\varphi^k(\cdot))$. The α , θ and φ respectively denote learnable network parameters of $h(\cdot)$, $f^k(\cdot)$ and $g^k(\cdot)$. In the following, we explain more details about the spatio-temporal encoder (Section 3.1), the occupancy field decoder (Section 3.2), the correspondence decoder (Section 3.3), the training paradigm (Section 3.4), and the inference (Section 3.5).

3.1. Spatio-temporal Representations Learning

To understand 3D shape motions in a sequence of consecutive observed frames, it is crucial to learn both geometric features for shape recovery and the temporal correlation for continuity maintenance. A straightforward solution is to follow the previous method that [29] uses two parallel PointNet-based [5] encoders to extract shape and motion embeddings respectively. However, its proposed shape encoder only uses the first point cloud \mathcal{S}_0 to acquire the shape feature. Thus the reconstructed surfaces are always sub-optimal if \mathcal{S}_0 is seriously incomplete, as the shape properties in other frames can not be incorporated. And its proposed temporal encoder strictly assumes that the point-wise correspondence between different input frames is known in the learning, which restricts its flexibility of processing real

scans without this relationship. Moreover, the aggregation of temporal information does not explicitly take into account the time information. Our motivation is to aggregate the shape features from different frames to capture robust embeddings for the implicit surface generation, and to treat time as important as the spatial coordinates to capture expressive embeddings for describing the dynamic shape evolution. Although it is possible to utilize the technique of 4D point cloud processing proposed by [21], it would be insufficient due to time-consuming spatio-temporal neighborhood queries. Thus we introduce a novel spatio-temporal encoder shown in Fig. 3. Specifically, it contains a T-PointNet branch that accepts and transforms the entire 4D point cloud $\mathcal{S} = \{\mathcal{S}_k\}_{k=0}^{K-1}$ to extract a temporal representation by treating time explicitly and equally to each spatial dimension. It also uses the S-PointNet branch to produce a sequence of geometric representations by individually applying it for each frame \mathcal{S}_k without considering timestamps. Finally, the geometric and temporal features are fused to obtain a sequence of descriptors $\{\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_{K-1}\}$ aggregating shape properties and exploring dynamics variations.

3.2. Occupancy Fields Predicting

In this section, we present the details of the learning occupancy field at each time step for 4D shape reconstruction. The occupancy field represents the 3D shape as a continuous boundary classifier where each 3D point is classified as 0 or 1, depending on whether the point lies inside or outside the surface. Although the signed distance field (SDF) can be an alternative choice, we observed convergence issues by employing encoder-decoder architecture for SDF learning from sparse point clouds.

According to the universal approximation theorem [14], we implement the occupancy field learning as a multi-layer perceptron (MLP) to predict occupancy states for the points sampled in space and time. Given a point \mathbf{p}_k sampled

at time t_k , the probability of locating outside the 3D human body is predicted by $f_{\theta}^k(\mathbf{p}_k) := f(\mathbf{p}_k; \mathbf{z}_k)$ that feed-forwards the feature constructed by concatenating the point coordinates \mathbf{p}_k and its associate spatio-temporal feature \mathbf{z}_k .

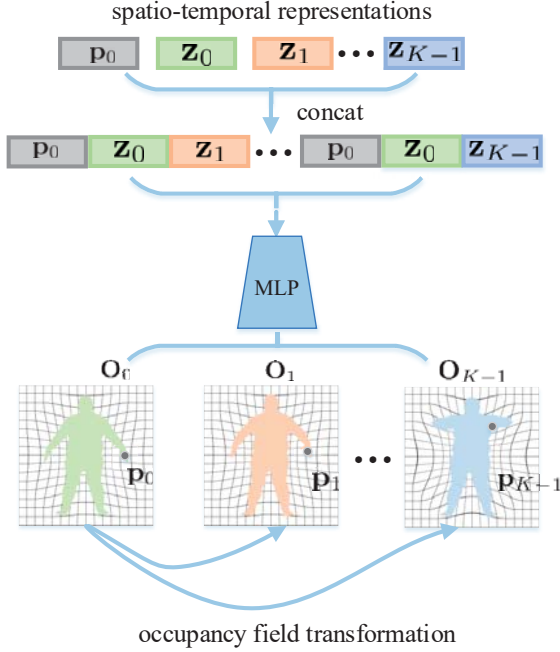


Figure 4: **Dense Correspondence Decoder.** Based on the learned spatio-temporal representations $\{\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_{K-1}\}$, the correspondence decoder utilizes a shared MLP conditioned on the concatenation of the associate representations \mathbf{z}_0 and \mathbf{z}_k , to predict the displacements from \mathbf{p}_0 at time t_0 to \mathbf{p}_k located in the coordinate system at time t_k

3.3. Cross-time Correspondence Modeling

In this section, we describe the details about dense correspondence modeling between the initial occupancy field \mathbf{O}_0 and others ($\mathbf{O}_1, \dots, \mathbf{O}_{K-1}$). Although it is feasible to model the occupancy field transformation between consecutive two frames in order to omit the complex computations of solving neural ODE [6], the sequential manner would lead to accumulated prediction errors and slow inference. Thus we choose to implement this by predicting displacement vector fields to future time steps in parallel paths.

Each displacement vector function is responsible for describing the occupancy field deformation from initial frame to subsequent frames. More specifically, the transformation process from time t_0 to time t_k can be formulated as:

$$\mathbf{p}_k = \mathbf{p}_0 + g_{\varphi}^k(\mathbf{p}_0) \quad (1)$$

where $g_{\varphi}^k : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is used to predict the displacement of each point \mathbf{p}_0 at time t_0 to the associate position \mathbf{p}_k located in the coordinate system at time t_k . According to Mo-

tion Coherent Theory [49], it is significant to ensure the deformation vector function g_{φ}^k are continuous. To meet this goal, we implement g_{φ}^k with a shared multi-layer perceptron (MLP) conditioned on \mathbf{z}_0 and \mathbf{z}_k that capture geometric properties and temporal dynamics at time t_0 and t_k . The dense correspondence decoder is shown in Fig. 4. Specifically, for each point \mathbf{p}_0 , we concatenate its coordinates with the spatiotemporal features $\mathbf{z}_0, \mathbf{z}_i$ that are associated with time t_0 and t_k . Then the displacement vector \mathbf{p}_0 to \mathbf{p}_k can be obtained by

$$\mathbf{p}_k - \mathbf{p}_0 = g_{\varphi}(\mathbf{p}_0 \oplus \mathbf{z}_0 \oplus \mathbf{z}_k) \quad (2)$$

where the symbol \oplus denotes a concatenation operation along the feature channel direction.

3.4. Training Objective

Our network learning is supervised by two types of optimization losses. For the occupancy field generation and transformation, we employ the standard binary cross-entropy loss for measuring the discrepancy between the predicted probabilities and the ground truths. It is defined as:

$$\mathcal{L}_{occ} = \sum_k \sum_{\mathbf{p}_k \in \mathcal{P}_k} \mathcal{L}_{bce}(f_{\theta}^k(\mathbf{p}_k), \mathbf{O}^k(\mathbf{p}_k)) + \mathcal{L}_{bce}(f_{\theta}^k(g_{\varphi}^k(\mathbf{p}_0)), \mathbf{O}^k(g_{\varphi}^k(\mathbf{p}_0))), \quad (3)$$

where $\mathbf{O}^k(\mathbf{p}_k)$ denotes the ground truth occupancy value of \mathbf{p}_k at time t_k . The first term is used to constrain the implicit surface generation at each time. And the second term is used to constrain the occupancy states changing for non-surface points.

The dense correspondence decoder is also trained by constraining the temporal evolution of 3D points sampled from the dynamic surfaces. The temporal correspondence loss can then be defined as:

$$\mathcal{L}_{corr} = \sum_k \sum_{\mathcal{Q}} |g_{\varphi}^k(\mathcal{Q}(t_0)) - \mathcal{Q}(t_k)| \quad (4)$$

where \mathcal{Q} denotes the a trajectory $\mathcal{Q}(t_0), \mathcal{Q}(t_1), \dots, \mathcal{Q}(t_{K-1})$ sampled from the dynamic surfaces at different time steps.

Then the overall optimization objective of our proposed approach \mathcal{L}_{total} can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{occ} + \lambda * \mathcal{L}_{corr} \quad (5)$$

where λ is a hyper-parameter weighting the importance of the temporal correspondence loss \mathcal{L}_{corr} .

3.5. Inference

During the inference stage, we predict the dynamic 3D shapes for a new observation \mathcal{S} by first reconstructing the shape at starting time $t = t_0$, followed by propagating

the reconstruction into the future $t \in [t_1, \dots, t_{K-1}]$ using the trained correspondence decoder. Thus we do not need to predict the occupancy field at each time step. We use the Multiresolution IsoSurface Extraction (MISE) [24] and marching cubes algorithms [22, 19] to extract the triangular mesh $\mathcal{M}_0 = \{\mathcal{V}_0, \mathcal{E}_0, \mathcal{F}_0\}$ where $\mathcal{V}_0, \mathcal{E}_0, \mathcal{F}_0$ represent the vertices, edges, and faces of mesh \mathcal{M}_0 from the predicted occupancy field at initial time $t = t_0$. For other time steps in the future, we use the learned deformation vector fields to calculate the displacement $g_\varphi^k(\mathbf{v}_i)$ for the each vertice \mathbf{v}_i in \mathcal{V}_0 while fixing the topology connectivity relationships $\mathcal{E}_0, \mathcal{F}_0$. So the mesh at time t_k is obtained through:

$$\mathcal{M}_k = \{\{g_\varphi^k(\mathbf{v}_i) | \mathbf{v}_i \in \mathcal{V}_0\}, \mathcal{E}_0, \mathcal{F}_0\} \quad (6)$$

Notably, compared with previous methods [24, 29], our approach can provide a faster network inference as it parallelly estimates the vertex displacements of \mathcal{M}_0 for different time steps and avoids the expensive computation of solving ordinary differential equations.

4. Experiment

Datasets Our experiments are performed on the challenging Dynamic FAUST (D-FAUST) [4] dataset which contains raw-scanned and registered meshes for 129 sequences of 10 humans (5 females and 5 males) with various motions such as “shake hips”, “running on spot”, or “one leg jump”. Same as the train/val/test split of OFlow [29], we divide all sequences into training (105), validation (6), and test (21) sequences. All models are evaluated on unseen actions or individuals during training. The test set consists of two subsets. One (S1) contains 9 sequences of seen individuals with unseen motions in the train set. The other (S2) contains 12 sequences of an unseen individual. To increase the size of the training samples, we subdivide each sequence into short segments of 17 time steps or long segments of 50 time steps according to different experiment settings.

Baselines We compare our approach with three state-of-the-arts for 4D reconstruction from point cloud sequences, including PSGN 4D, ONet 4D, and OFlow. The PSGN 4D extends the PSGN [9] to predict a 4D point cloud, *i.e.* the point cloud trajectory instead of a single point set. The ONet 4D is a natural extension of ONet [24] to define the occupancy field in the spatio-temporal domain by predicting occupancy values for points sample in space and time. The OFlow [29] assigns each 4D point an occupancy value and a motion velocity vector and relies on the differential equation to calculate the trajectory. For a fair comparison, we train all models of baselines with the paradigms in [29].

Implementation details The model implementation is based on OFlow [29]. For all experiments, our model is trained in an end-to-end manner using a batch size of 16 with a learning rate of $1e-4$ for 400k iterations. For the loss

	Method	IoU	Chamfer	Correspond.
S1	PSGN 4D [9]	-	0.101	0.102
	ONet 4D [24]	77.9%	0.084	-
	OFlow [29]	81.5%	0.065	0.094
	Ours	84.9%	0.055	0.080
S2	PSGN 4D [9]	-	0.119	0.131
	ONet 4D [24]	66.6%	0.140	-
	OFlow [29]	72.3%	0.084	0.117
	Ours	76.2%	0.071	0.098

Table 1: Quantitative comparisons on the task of **4D Shape Reconstruction** from **time-evenly** sampled point cloud sequences. We evaluate the performance on the *test* set of (S1) **unseen motions** (but seen individuals) and (S2) **unseen individuals**. The metrics of Chamfer distance, correspondence, and IoU are reported.

	Method	IoU	Chamfer	Correspond.
S1	PSGN 4D [9]	-	0.148	0.121
	ONet 4D [24]	71.9%	0.114	-
	OFlow [29]	76.9%	0.090	0.134
	Ours	83.8%	0.059	0.090
S2	PSGN 4D [9]	-	0.155	0.140
	ONet 4D [24]	62.8%	0.130	-
	OFlow [29]	67.2%	0.112	0.178
	Ours	74.8%	0.076	0.118

Table 2: Quantitative comparisons on the task of **4D Shape Reconstruction** from **time-unevenly** sampled point cloud sequence with large variations between adjacent frames.

calculation in each training iteration, we randomly sample a fixed number of points in space and time. More specifically, for the occupancy prediction loss L_{occ} , we sample 512 points that are uniformly distributed in the bounding box of the 3D shapes at each respective time. For the correspondence loss L_{corr} , we uniformly sample the trajectories of 100 points from the sequence of ground truth surfaces. And the hyperparameters used in Equation 5 are $\lambda = 1$.

Evaluation Metrics We use Chamfer distance (lower is better), Intersection over Union (higher is better), and correspondence distance (lower is better) as primary metrics to evaluate the reconstructed surface mesh sequences. We follow OFlow [29] to compute these evaluation metrics.

4.1. 4D Shape Reconstruction

We first compare the performance of our approach with previous methods for reconstructing time-varying surfaces from the sparse point cloud sequences in two kinds of input, including time-evenly and time-unevenly sampled se-

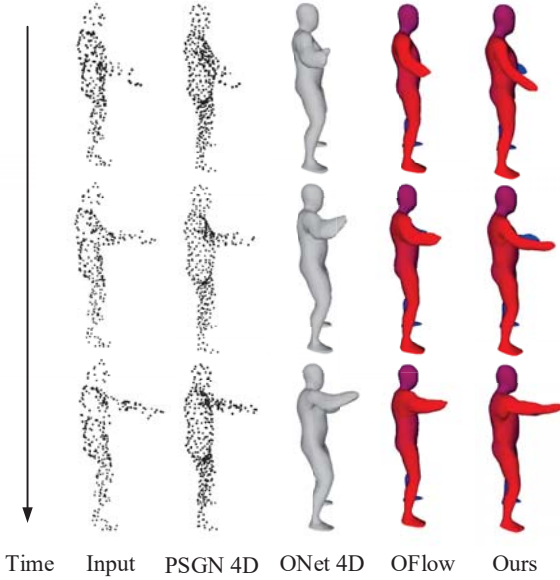


Figure 5: **Qualitative Results on 4D Shape Reconstruction.** We qualitatively show the input of three unequally space and time steps with large variations, and the output of PSGN 4D [9], ONet 4D [24], and OFlow [29]. Colors ranging from red to blue index mesh faces to better illustrate the surface correspondence across time.

quences. The network input is 300 discrete point trajectories randomly sampled from dynamic groundtruth surfaces. In order to simulate the noises in the real world, we add gaussian noise with standard deviation 0.05 to perturb the point clouds. For the former one, each trajectory consists of $K = 17$ time steps with uniform intervals. For the latter one, we randomly select 6 frames from a long segment of 50 time steps as input. Each trajectory experiences non-uniform intervals and large variations.

The quantitative and qualitative comparisons are respectively shown in Table 1 Table 2 and Fig. 5. As shown in the quantitative results, our approach achieves superior performance over all previous methods. From the Fig. 5, we can observe that our method can capture plausible motions and correspondences over time but ONet 4D can not. PSGN 4D predicts sparse and noisy point cloud sequences, causing the challenge to get clean dynamic surface meshes. Besides, compared to OFlow [29], our method can achieve more robust geometry recoveries benefiting from the proposed multi-frame shape information aggregation. Moreover, large performance improvements shown in Table 2 demonstrate that our multi-frame bundled correspondence modeling can achieve higher robustness on non-uniform sequences with large variations.

4.2. 4D Shape Completion

In addition to 4D shape reconstruction experiments, we also compare the performance on incomplete observations.

	Method	IoU	Chamfer	Correspond.
S1	OFlow	75.9%	0.094	0.142
	Ours (C1)	81.0%	0.070	0.112
	Ours (C2)	58.6%	0.124	0.254
	Ours (Full)	82.4%	0.064	0.105
S2	OFlow	67.0%	0.113	0.183
	Ours (C1)	71.7%	0.087	0.139
	Ours (C2)	56.8%	0.164	0.344
	Ours (Full)	72.9%	0.082	0.134

Table 3: **Ablation studies & 4D Shape Completion:** Ours (C1) indicates our method without using the proposed spatio-temporal encoder, and Ours (C2) denotes our method without using the parallel correspondence modeling.

Specifically, we create partial point clouds by randomly select 5 seeds on the surface and discard those regions within the radius of 0.1. The input is a sequence of $K = 6$ incomplete point clouds randomly sampled from a long segment, and each contains 300 points. The quantitative and qualitative results are respectively shown in Table 3 and Fig. 6. The better performances verify the superiority of our approach in the dynamic surface recoveries with temporal coherence from incomplete observations.

4.3. Ablation Studies

Our whole framework contains two key modules. In this section, we conduct the ablation studies by alternatively removing one of them to verify the necessity of each module. We perform experiments on the task of 4D shape completion from non-uniform sequences with large variations.

Without spatio-temporal encoder (C1) Based on our pipeline, an alternative solution to learn the spatio-temporal representations is to individually utilize Res-PointNet [5] (details described in the supplementary material) to process the 4D point cloud with time information at each frame. The comparison results are shown in Fig. 6 and Table 3. As can be seen, our method achieves more complete geometry as the designed spatio-temporal encoder can efficiently aggregate the shape properties of all frames.

Without parallel correspondence modeling (C2) Instead of parallelly modeling the dense correspondences, an alternative solution is to predict the occupancy field transformations between adjacent frames. As shown in Fig. 6, the sequential correspondence modeling accumulates the prediction errors in the legs in the second frame, resulting in wrong shape predictions in subsequent frames and non-continuous human dynamics. We also verify it by the increased correspondence distance (see Table 3). Moreover, the sequential manner adopted by OFlow and Ours (C2) easily produces distorted results such as stretched surfaces around the left knee when capturing large human variations.

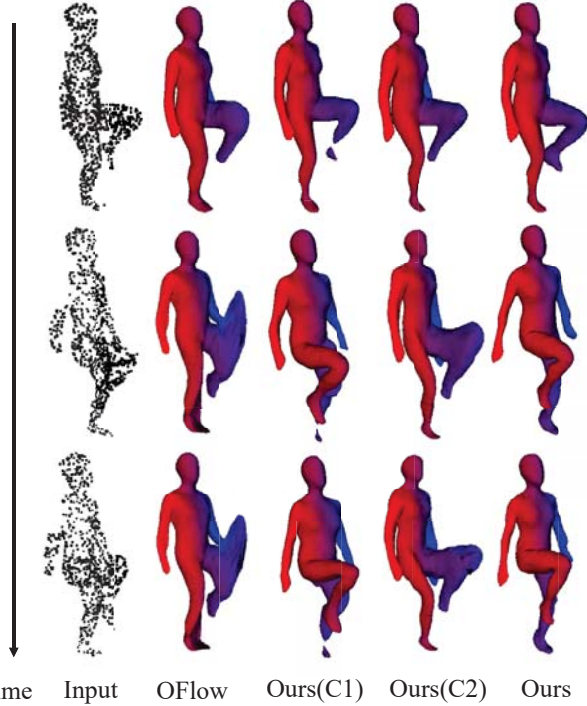


Figure 6: **4D Shape Completion & Ablation Studies:** Ours (C1) indicates our method without using the proposed spatio-temporal encoder, and Ours (C2) denotes our method without using the parallel correspondence modeling.

4.4. Space and Time Complexity

We compare our method to OFlow [29] in terms of memory footprint and computational efficiency. We train both models using a batch size of 16 for 4D shape reconstruction from a sequence of 17 time steps with uniform intervals and report the training memory footprint, total training time. And we calculate the average of batch training time, batch forward time, and batch backward time in the initial 10k iterations of training. We also report the average network inference time using a batch size of 1 for 1k test sequences. Both models were run on a single GTX 1080 Ti. We observed the slow training procedure of OFlow [29] as ODE-solver requires demanding computations and gradually increases the number of iterations to meet the error tolerance. From the results shown in Table 4, we can see that although our model has a higher training memory footprint, it is about 4 times faster in training and 8 times in inference.

4.5. Shape Matching

In this section, we investigate our pipeline for the task of shape matching. The inputs are the underlying surfaces of two randomly sampled point clouds, and the outputs are the point displacements of source surface to the target surface. Since this task does not need to recover 3D surface meshes,

Method	Mem. (GB)	Train (day)	Inference (s)
OFlow [29]	3.53	42	1.84
Ours	10.8	10	0.23
		Forward (s).	Backward (s).
OFlow [29]	0.33	4.02	4.35
Ours	0.45	0.49	0.94

Table 4: **Space and time complexity comparison** between OFlow [29] and our method.

Method	Correspond	Time(s)
Nearest Neighbor	0.374	0.004
Coherent Point Drift [27]	0.189	343.6
OFlow [29]	0.167	0.309
Ours	0.102	0.011

Table 5: **Shape Matching Experiments.** We report the correspondence distance (correspond) of two randomly sampled point clouds with the size 10k.

the model consists of only a spatio-temporal encoder and a dense correspondence decoder, and the training is only supervised by the correspondence loss in Eq. 4. From the quantitative comparisons shown in Table 5, we can conclude that although our method is primarily designed for 4D reconstruction, it can also predict accurate correspondences. Moreover, we remark that our inference speed is remarkably faster than that of CPD [27] and OFlow [29], with approximately 31,000 and 30 times faster, respectively.

5. Conclusion

We have proposed a novel learning framework to reconstruct time-varying surfaces from point cloud sequences. The overall framework includes a flexible framework for learning robust spatio-temporal shape representations from 4D point clouds and an efficient cross-frame correspondence decoder that simultaneously models the occupancy field transformations from the first frame to others. Comparisons with previous works demonstrate that our approach can achieve more accurate geometries, better temporal continuity while significantly improves the computation efficiency. One limitation of our method is that to achieve superior practical efficacy and efficiency, we sacrifice theoretically temporal continuity due to the discrete field transformation, which will be further explored in the future works.

Acknowledgement. This work was partially supported by the National Natural Science Foundation of China (No.: 61771201), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (No.: 2017ZT07X183), the Guangdong R&D key project of China (No.: 2019B010155001), Alibaba DAMO Academy, and Hong Kong RGC GRF (No.: 15221618).

References

- [1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018.
- [2] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *CVPR*, 2020.
- [3] Silvia Biasotti, Andrea Cerri, Alex Bronstein, and Michael Bronstein. Recent trends, applications, and perspectives in 3d shape similarity assessment. In *Computer Graphics Forum*, volume 35. Wiley Online Library, 2016.
- [4] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic faust: Registering human bodies in motion. In *CVPR*, 2017.
- [5] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.
- [6] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.
- [7] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019.
- [8] Christopher Bongsoo Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016.
- [9] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017.
- [10] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*, 2020.
- [11] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. In *CVPR*, 2018.
- [12] Xiaoguang Han, Zhen Li, Haibin Huang, Evangelos Kalogerakis, and Yizhou Yu. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *ICCV*, 2017.
- [13] Christian Häne, Shubham Tulsiani, and Jitendra Malik. Hierarchical surface prediction for 3d object reconstruction. In *3DV*, 2017.
- [14] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 1989.
- [15] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, 2017.
- [16] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019.
- [17] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018.
- [18] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020.
- [19] Jiabao Lei and Kui Jia. Analytic marching: An analytic meshing solution from deep implicit surface networks. In *ICML*, 2020.
- [20] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Multi-view dynamic shape refinement using local temporal integration. In *ICCV*, 2017.
- [21] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. Meteor-net: Deep learning on dynamic 3d point cloud sequences. In *ICCV*, 2019.
- [22] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4), 1987.
- [23] Marcel Lüthi, Thomas Gerig, Christoph Jud, and Thomas Vetter. Gaussian process morphable models. *TPAMI*, 40(8), 2017.
- [24] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019.
- [25] Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. General dynamic scene reconstruction from multiple view video. In *ICCV*, 2015.
- [26] Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. Temporally coherent 4d reconstruction of complex dynamic scenes. In *CVPR*, 2016.
- [27] Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *TPAMI*, 32(12), 2010.
- [28] Jan Neumann and Yiannis Aloimonos. Spatio-temporal stereo using multi-resolution subdivision surfaces. *IJCV*, 47(1-3), 2002.
- [29] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *ICCV*, 2019.
- [30] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *ICCV*, 2019.
- [31] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019.
- [32] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *ECCV*, 2020.
- [33] Omid Poursaeed, Matthew Fisher, Noam Aigerman, and Vladimir G Kim. Coupling explicit and implicit surface representations for generative 3d modeling. In *ECCV*, 2020.
- [34] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, 2017.
- [35] Miroslava Slavcheva, Maximilian Baust, and Slobodan Ilic. Towards implicit correspondence in signed distance field evolution. In *ICCV*, 2017.
- [36] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE computer graphics and applications*, 27(3), 2007.
- [37] Gary KL Tam, Zhi-Quan Cheng, Yu-Kun Lai, Frank C Langbein, Yonghuai Liu, David Marshall, Ralph R Martin, Xian-Fang Sun, and Paul L Rosin. Registration of 3d point clouds

- and meshes: A survey from rigid to nonrigid. *TVCG*, 19(7), 2012.
- [38] 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 *CVPR*, 2019.
 - [39] 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. *arXiv preprint arXiv:2008.05742*, 2020.
 - [40] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *ICCV*, 2017.
 - [41] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NeurIPS*, 2017.
 - [42] Oliver Van Kaick, Hao Zhang, Ghassan Hamarneh, and Daniel Cohen-Or. A survey on shape correspondence. In *Computer Graphics Forum*, volume 30. Wiley Online Library, 2011.
 - [43] Michael Wand, Philipp Jenke, Qixing Huang, Martin Bokeloh, Leonidas Guibas, and Andreas Schilling. Reconstruction of deforming geometry from time-varying point clouds. In *SGP*, 2007.
 - [44] 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.
 - [45] Peng-Shuai Wang, Chun-Yu Sun, Yang Liu, and Xin Tong. Adaptive o-cnn: a patch-based deep representation of 3d shapes. In *SIGGRAPH Asia*, 2018.
 - [46] Dan Xu, Weidi Xie, and Andrew Zisserman. Geometry-aware video object detection for static cameras. In *BMVC*, 2019.
 - [47] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *NeurIPS*, 2019.
 - [48] Mingyue Yang, Yuxin Wen, Weikai Chen, Yongwei Chen, and Kui Jia. Deep optimized priors for 3d shape modeling and reconstruction. *arXiv preprint arXiv:2012.07241*, 2020.
 - [49] Alan L Yuille and Norberto M Grzywacz. A mathematical analysis of the motion coherence theory. *IJCV*, 3(2), 1989.
 - [50] Wenbin Zhao, Jiabao Lei, Yuxin Wen, Jianguo Zhang, and Kui Jia. Sign-agnostic implicit learning of surface self-similarities for shape modeling and reconstruction from raw point clouds. *arXiv preprint arXiv:2012.07498*, 2020.