Dynamic Point Fields

Sergey Prokudin[1] Qianli Ma[1,2] Maxime Raafat[1] Julien Valentin[3] Siyu Tang[1]
[1]ETH Zürich [2]Max Planck Institute for Intelligent Systems [3]Microsoft
[1] firstname.lastname@inf.ethz.ch

Abstract

Recent years have witnessed significant progress in the field of neural surface reconstruction. While extensive focus was put on volumetric and implicit approaches, a number of works have shown that explicit graphics primitives, such as point clouds, can significantly reduce computational complexity without sacrificing the reconstructed surface quality. However, less emphasis has been put on modeling dynamic surfaces with point primitives. In this work, we present a dynamic point field model that combines the representational benefits of explicit point-based graphics with implicit deformation networks to allow efficient modeling of non-rigid 3D surfaces. Using explicit surface primitives also allows us to easily incorporate well-established constraints such as isometric-as-possible regularization. While learning this deformation model is prone to local optima when trained in a fully unsupervised manner, we propose to also leverage semantic information, such as keypoint correspondence, to guide the deformation learning. We demonstrate how this approach can be used for creating an expressive animatable human avatar from a collection of 3D scans. Here, previous methods mostly rely on variants of the linear blend skinning paradigm, which fundamentally limits the expressivity of such models when dealing with complex cloth appearances, such as long skirts. We show the advantages of our dynamic point field framework in terms of its representational power, learning efficiency, and robustness to out-of-distribution novel poses. The code for the project is publicly available [1].

1. Introduction

Neural surface reconstruction and rendering have been subject to giant leaps in the last couple of years [81]. While original works have suffered from the representational limitations stemming from using a single global neural network to represent a surface as implicit signed distance [4, 52, 73] or occupancy field [43], recent models pushed the level of reconstruction quality by introducing auxiliary data structures such as octrees [80] or hashmaps [46]. While providing high level of accuracy, these lines of work rely on the ability to perform large amounts of queries to the underlying multi-layer perceptrons (MLPs) to reconstruct the surface.

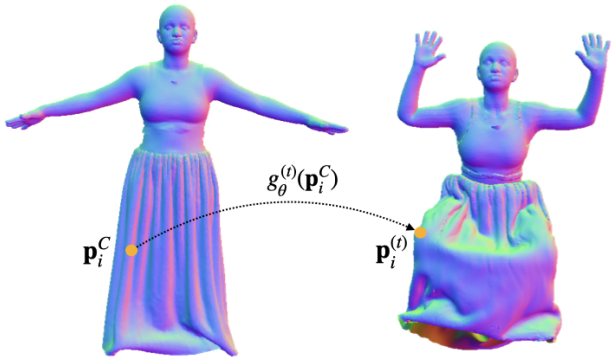

Figure 1. *Dynamic point field.* We propose to model dynamic surfaces with a point-based model, where the motion of a point $\mathbf{p}_i$ over time is represented by an implicit deformation field $g_\theta^{(t)}$. Working directly with points rather than SDFs allows us to easily incorporate various well-known deformation constraints, e.g. as-isometric-as-possible [23]. We showcase the usefulness of this approach for creating animatable avatars in complex clothing.

In the field of volumetric neural rendering, this restriction has motivated development of compact and computationally attractive models that can train quickly and render at interactive rates [6, 11, 36, 67, 91]. Among them, point-based methods for neural rendering [1, 26, 66, 68, 86, 87, 93] proved to be scalable alternatives to purely implicit or volumetric approaches. In the context of differentiable neural surface modeling, points have also been explored for static scenes [57, 90] and dynamic human modeling [42].

This work aims to achieve three goals. First, we showcase the advantage of a point-based surface representation compared to the most recent implicit models [46]. We do this by carefully analyzing the behavior of the existing frameworks when representing complex 3D meshes with

[1] sergeyprokudin.github.io/dpf

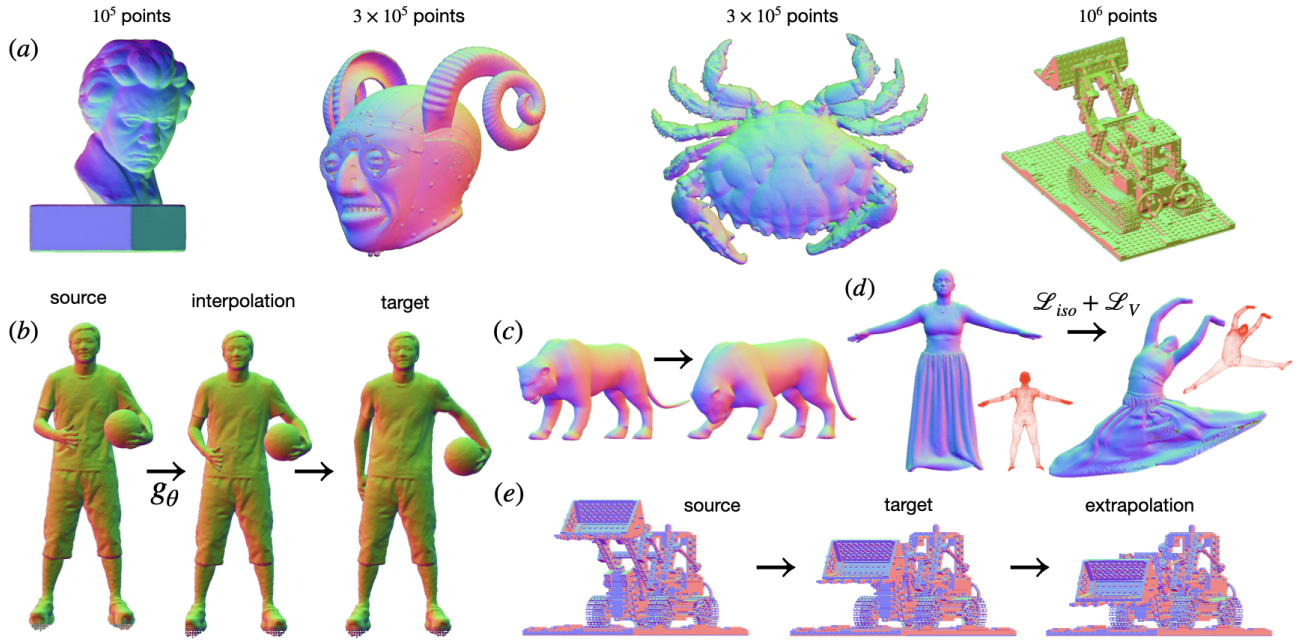

Figure 2. *Overview of the main results. (a)*: static surface reconstruction with optimised point clouds. We propose a point cloud optimisation scheme that efficiently utilises available off-the-shelf point renderers [20, 86] and surpasses the performance of more sophisticated methods [46, 80], with an explicit surface representation which requires zero inference time. *(b-e)*: we extend the approach to allow modeling dynamic surfaces of varying topology and complexity, and propose a guided learning for the case of complex deformations. Our formulation also allows us to interpolate easily between canonical and target surfaces.

perfect ground truth available. Here, point clouds as primitives demonstrate direct benefits over their implicit alternatives in all the comparison dimensions: reconstructed geometry quality, training time, runtime, and model size.

Second, we extend the point-based surface model to support *non-rigidly deforming surfaces*. A number of approaches has been proposed to tackle this problem, both for generic 3D scenes [19, 49] and human-specific models [42, 92]. However, these models often tend to constrain the space of deformations via restrictions to specific algebraic operations, commonly linear blend skinning functions [42, 89] or a mixture of affine warps [19, 37]. Instead, we rely on a deformation field represented by a neural network [51], similar to the approach taken in dynamic neural radiance fields [53, 64]. Again, we show that learning such deformation networks on point sets is easier and more robust compared to SDFs. More importantly, it also allows us to easily incorporate various well-known regularisation techniques, such as as-isometric-as-possible [17, 23], enforcing the preservation of distances between points in the canonical and deformed space.

Further, we propose a guided learning regime to improve the robustness of the dynamic point fields and allow 3D shape manipulation with our framework. Our key idea is the following: undesirable local optima can be avoided by exploiting sparse correspondences in space-time as additional constraints, further improving optimisation speed and stability. For instance, to reconstruct a highly dynamic and complex deformation of a long skirt (Figure 5), we supervise the learning of the deformation field with the body vertices correspondences between frames from the readily available underlying unclothed body model registrations [42]. As demonstrated in the experiments section, the correspondence constraints on the minimally-clothed bodies provide strong generalization capabilities on the deformation of a long skirt largely deviating from the body surface. Based on this observation, we propose a method for zero-shot avatar reposing based on the introduced deformation framework. Our method compares favorably to other point-based [40, 42] and implicit methods [70] which are fundamentally based on the linear blend skinning paradigm and fail to correctly represent challenging cloth types.

To summarise, our main contributions are as follows. We first introduce *dynamic point fields*, a simple and computationally attractive model combining the compactness and efficiency of point primitives with the flexibility and accuracy of neural networks to model deformations. Advantages of the proposed approach are demonstrated by comparing it with the state-of-the-art in various surface reconstruction and deformation learning tasks. Second, we propose a

guided deformation field learning and show how additional constraints on available keypoint correspondences can be used to efficiently guide the learning of surface dynamics, without sacrificing the expressivity of a model.

2. Related Work

Point-based representations. Point clouds have a long history of applications in computer vision, both for scene rendering and surface modeling [13, 24, 28, 61, 69]. With the rise of differentiable rendering techniques [81], they gained a renowned interest and in many use cases, have shown to be efficient alternatives to both volumetric and mesh-based techniques [1, 22, 25, 26, 68, 86, 87, 90]. Our paper mainly uses the differentiable point rendering technique proposed in [86] and [26] for point cloud optimisation, which is in turn similar in spirit to differential surface splatting [90].

Differentiable surface representations. The early age of visual computing was mostly dominated by explicit representations such as meshes and volumes. However, recent advances have shown the benefits of implicit representations, where a signed distance field or an occupancy is parameterised via a deep neural network [43, 52]. Follow-up works have further improved expressivity and efficiency of the deep representations by changing the network structure and activations [59, 73] or incorporating regularisations [4, 12]. Further augmenting neural networks with auxiliary data structures such as octrees [80] or hashmaps [46] has also demonstrated to significantly boost model performance. Compared to prior works, these models have shown the capacity to represent highly detailed surfaces with many intricate details. We mostly focus on comparing to these baselines, as they represent the state-of-the-art in the surface modeling field. Similar to [58], we show that a simple point cloud, which naturally handles 3D space sparsity, can efficiently rival the aforementioned recent approaches while offering numerous benefits such as model simplicity, explicitness and compactness.

Dynamic surface modeling, non-rigid registration. Modeling deformations has in turn a long history in computer vision [3, 10, 16, 39, 45, 47, 50, 65, 79, 83]. Among the most intensively studied are the models that are built on top of 3D templates for the learning of dynamic shapes [2, 8, 14, 18, 38, 56, 62]. Several point-based models have recently been introduced to facilitate neural rendering and surface modeling of dynamic humans [40, 42, 63]. However, these models are still limited when it comes to modeling challenging clothes and dynamics such as skirts, due to heavy reliance on linear blend skinning functions. In contrast, our model is comprised of a simple and expressive point set and a flexible neural network to model non-rigid deformations of varying complexity. The formulation is also more general than recently proposed point-based methods for human draping [92]. Modeling generic point cloud deformations has been addressed in several works [29, 71, 78, 84], with many works escaping local optima by restricting the set of possible deformations [19] or building priors on the cloud motion [9, 47].

In contrast, we utilise recent developments in implicit dynamic scene and surface modeling [27, 30, 48, 53, 54, 75, 82]. Similar to [64], we employ a deep neural network that learns to transfer points from a canonical frame to a target frame. Our differentiating factor here is the use of point clouds as our canonical space graphics primitive instead of SDFs or radiance fields. Second, in the case of challenging human surface modeling, using explicit 3D primitives also enables the direct supervision of our deformation network with motion of minimally clothed body model, which significantly boosts the training time and helps avoiding local optima.

Clothed human modeling and animation. Given a set of 3D clothed human scans, how do we animate them into new poses with visually plausible clothing shapes? A majority of recent clothed human models address this problem by learning a regression model from the body pose to the clothing geometry, using a variety of shape representations such as meshes [15, 41, 55], implicit surfaces [7, 70, 82, 85] and point sets [35, 40, 42, 92]. These methods typically rely on linear blend skinning (LBS) to handle body articulation. While efficient, using LBS limits the models' capability in representing loose clothing and certain common garment types, such as skirts and dresses. We refer the reader to [40] for a detailed discussion on the limitation of LBS-based clothed human models. In this paper, we demonstrate that our guided deformation learning serves as an effective alternative in animating clothed humans. Being free from the drawbacks of LBS, it can generate coherent and plausible clothing shape even under unseen extreme poses.

3. Method

3.1. Dynamic Point Field Model

Our framework comprises of two elementary components: a point cloud with learnable spatial locations and features, and a set of compact neural networks (one for each time step) that warp every point into a new location.

Dynamic point set. More formally, we define a point cloud X as a set of tuples $\mathbf{p}_i$:

$$\mathbf{X} = \{\mathbf{p}_i = (\mathbf{x}_i, \mathbf{n}_i), \mathbf{x}_i \in \mathbb{R}^3, \mathbf{n}_i \in \mathbb{R}^3\}_{i=1,\dots,N_p}, \quad (1)$$

where $\mathbf{x}_i$ are the 3D point locations and $\mathbf{n}_i$ the corresponding point normals. The point cloud consists of N_p points.

We then represent a dynamic scene $\mathbf{S}$ as sequences of point sets:

$$\mathbf{S} = [\mathbf{X}^{(0)}, \dots, \mathbf{X}^{(t)}, \dots, \mathbf{X}^{(T)}], \quad (2)$$

where $\mathbf{X}^{(t)}$ is the point cloud representing the scene at time step t, and T is the total number of steps.

Deformation field. We aim to model dynamics of a scene (Eq. 2) via a set of compact neural networks $\{g_\theta^{(t)}, t = 1, \dots, T\}$, which update locations of every point $\mathbf{p}_i^{\mathcal{C}}$ from a canonical set $\mathbf{X}^{\mathcal{C}}$:

$$g_\theta^{(t)} : \mathbb{R}^3 \rightarrow \mathbb{R}^3, \tag{3}$$

$$\mathbf{x}_i^{(t)} = \mathbf{x}_i^{\mathcal{C}} + g_\theta^{(t)}(\mathbf{x}_i^{\mathcal{C}}). \tag{4}$$

The corresponding warped normals $\mathbf{n}_i^{(t)}$ can be estimated in a fully differentiable manner using several approaches discussed in the appendix. In general, thanks to the isometric loss described in the next section and a bias of the network towards smooth predictions, the predicted deformations in our experiments allow us to directly transfer and reuse the canonical space mesh connectivity for estimation.

Applying the deformation network to a set of points is simply equal to applying it to every point in the set:

$$\mathbf{X}^{(t)} = g_\theta^{(t)}(\mathbf{X}^{\mathcal{C}}) = \{\mathbf{x}_i^{\mathcal{C}} + g_\theta^{(t)}(\mathbf{x}_i^{\mathcal{C}})\}_{i=1,\dots,N_p}. \tag{5}$$

We set $\mathbf{X}^{\mathcal{C}} = \mathbf{X}^{(0)}$, if not stated otherwise. θ are the parameters of the neural networks.

Network architecture. We use a multi-layer perceptron with periodic activations as our deformation network for its performance and flexibility in representing various input domains [73]. If not stated otherwise, we use a small network with three hidden layers, each of size 128. This compact architecture makes running the deformation module on large point clouds extremely efficient. As an avenue for future work, we also consider using a meta-learning approach for learning these deformation networks based on a time-dependent latent variable, similar to [72, 85]. In this work, we focus on a simple scenario where every target deformation is modeled with a dedicated small module.

We will now discuss the ways to efficiently optimise dynamic point fields.

3.2. Training

Canonical surface reconstruction. Given a ground truth mesh $\mathcal{M}_{gt}$, for each optimisation step we obtain a point cloud $\mathbf{X}_{gt}$ by sampling points from a mesh with the associated normal directions. We then optimise a combination of the following losses:

$$\mathcal{L}_S(\mathbf{X}, \mathcal{M}_{gt}) = \lambda_{CD}\mathcal{L}_{CD} + \lambda_n\mathcal{L}_n + \lambda_{n_I}\mathcal{L}_{n_I}, \tag{6}$$

where $\mathcal{L}_{CD}$ and $\mathcal{L}_n$ are the standard Chamfer-based discrepancy losses between a ground truth cloud $\mathbf{X}_{gt}$ from the mesh $\mathcal{M}_{gt}$, and the optimised cloud $\mathbf{X}$, and their corresponding normals [20]. $\mathcal{L}_{n_I}$ is the image-space normal discrepancy between rendered mesh normal image $\mathcal{I}_{gt}^n$ and rendered point normals $\mathcal{I}_X^n$:

$$\mathcal{L}_{n_I}(\mathbf{X}, \mathcal{M}_{gt}) = ||\mathcal{I}_{gt}^n - \mathcal{I}_X^n||_2. \tag{7}$$

Here, images $\mathcal{I}_{gt}^n$ and $\mathcal{I}_X^n$ are obtained by rendering ground truth mesh and optimised point cloud normals from the same random camera position C, sampled from a unit sphere:

$$\mathcal{I}_{gt}^n = \mathcal{R}_m(\mathcal{M}_{gt}, C), \tag{8}$$

$$\mathcal{I}_X^n = \mathcal{R}_p(\mathbf{X}, C), \tag{9}$$

where $\mathcal{R}_m$ is a mesh renderer, and $\mathcal{R}_p$ is the differentiable point-based renderer [86] available as a part of [20]. Compared to solely optimizing for Chamfer-based normal loss, enforcing image-based normal loss produces more visually appealing and consistent point normals.

We use $\lambda_{CD} = 10^4, \lambda_n = 1$ and $\lambda_{n_I} = 10^1$ for balancing our loss functions and bringing them to the same scale. For the image-space loss $\mathcal{L}_{n_I}$, we render point clouds with a point radius of $25 \times 10^{-3}, 9 \times 10^{-3}, 5 \times 10^{-3}$ when using $10^4, 10^5$ and 10^6 points respectively, and an image resolution of 512×512.

Dynamic surface reconstruction. Given a sequence of ground truth meshes

$$\mathbf{M}_{gt}^{\sim} = [\mathcal{M}_{gt}^{(0)}, \dots, \mathcal{M}_{gt}^{(t)}, \dots, \mathcal{M}_{gt}^{(T)}], \tag{10}$$

the above described formulation allows direct extension for the case of dynamic surface reconstruction with a deformation network:

$$\mathbf{X}^{\mathcal{C}*}, \theta^* = \underset{\mathbf{X}^{\mathcal{C}}, \theta}{\operatorname{argmin}} \sum_{t=1}^{T} \mathcal{L}_S\big(\mathbf{X}^{(t)}, \mathcal{M}_{gt}^{(t)}\big), \tag{11}$$

$$\mathbf{X}^{(t)} = g_\theta^{(t)}(\mathbf{X}^{\mathcal{C}}). \tag{12}$$

However, the optimisation fails to find a plausible deformation field in the case of rapid and complex motion. To alleviate the problem, we propose two techniques for learning plausible deformations.

As-isometric-as-possible constraint. Our explicit 3D surface formulation allows us to use some classical constraints for learning deformations in 3D space. In this work, we use the as-isometric-as-possible constraint to guide our learning [23], which enforces the preservation of distances between points in the canonical and deformed space:

$$\mathcal{L}_{iso} = \sum_{i=1}^{N_p} \sum_{j \in \mathcal{N}_k(i)} ||d(\mathbf{x}_i^{\mathcal{C}}, \mathbf{x}_j^{\mathcal{C}}) - d(\mathbf{x}_i^{(t)}, \mathbf{x}_j^{(t)})||_1, \tag{13}$$

where $\mathcal{N}_k(i)$ is a k-neighborhood of the point $\mathbf{x}_i$ in $\mathbf{X}^{\mathcal{C}}$. We set $k = 5$, if not stated otherwise.

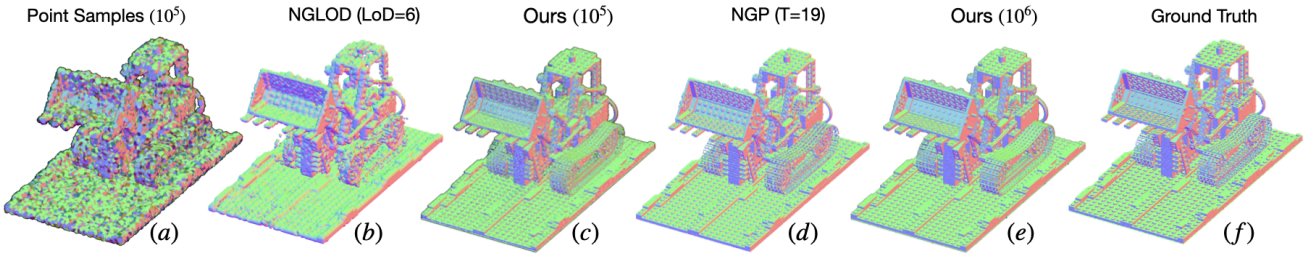

Figure 3. *Representing 3D surfaces with optimisable point sets.* Compared to the state-of-the-art implicit models for 3D surface representation [46, 80], optimised point cloud models offer better reconstruction quality on all metrics, while taking zero inference time thanks to its explicit nature. See Section 4.1 for details. Best viewed zoomed-in on a color screen.

To promote meaningful interpolations (see Figure 2*b*), we can also enforce the same constraint on the intermediate deformations train order to enforce continuous deformation between discrete timesteps:

$$\mathcal{L}_{iso}^{\gamma} = \sum_{i=1}^{N_p} \sum_{j \in \mathcal{N}_k(i)} ||d(\mathbf{x}_i^{\mathcal{C}}, \mathbf{x}_j^{\mathcal{C}}) - d(\mathbf{x}_i^{(\gamma t)}, \mathbf{x}_j^{(\gamma t)})||_1, \quad (14)$$

$$\mathbf{x}_i^{(\gamma t)} = \mathbf{x}_i^{\mathcal{C}} + \gamma \cdot g_\theta^{(t)}(\mathbf{x}_i^{\mathcal{C}}), \quad (15)$$

where $\gamma \in [0, 1]$. Enforcing this loss for $\gamma \in [1, \text{inf})$ allows us to also generate plausible extrapolations outside of the motion interval. See Figure 2*e* for an example of extrapolation of excavator bucket motion.

Guided deformation field learning. In many scenarios, paired information on the transformation of certain 3D points and their features is available:

$$\mathbf{V}^{(t)} = \{(\mathbf{v}_i^{\mathcal{C}}, \mathbf{v}_i^{(t)})\}_{i=1,\dots,N_v^{(t)}}, \quad (16)$$

where $(\mathbf{v}_i^{\mathcal{C}}, \mathbf{v}_i^{(t)})$ are the locations and features of the i^{th} keypoint in the canonical and target frames respectively, and $N_v^{(t)}$ is the total number of available keypoints per frame t. We denote the full set of keypoints for a scene (Eq. 2) as:

$$\mathbf{V}_S = [\mathbf{V}^{(0)}, \dots, \mathbf{V}^{(t)}, \dots, \mathbf{V}^{(T)}]. \quad (17)$$

This information can come from either 3D feature matching between canonical and target frames [32], 3D keypoints [5], or via registration or regression of a certain parametric model to 3D surfaces [38]. In the latter case, $(\mathbf{v}_i^{\mathcal{C}}, \mathbf{v}_i^{(t)})$ are simply 3D model mesh vertices in the canonical and posed spaces. We can utilise this information as a direct supervision signal for our deformation field model:

$$\mathcal{L}_V = \sum_{i=1}^{N_v^{(t)}} ||\mathbf{v}_i^{(t)} - \left(\mathbf{v}_i^{\mathcal{C}} + g_\theta^{(t)}(\mathbf{v}_i^{\mathcal{C}})\right)||_1. \quad (18)$$

We then combine it with our surface loss and isometric losses to form the following final loss:

Method	$\mathcal{L}_{CD} \downarrow$	$\mathcal{L}_n \downarrow$	Size (Mb)	Params (10^6)
Sampled points (10^4)	5.522	0.680	0.24	0.06
Sampled points (10^5)	1.101	0.554	2.4	0.6
Sampled points ($3 \cdot 10^5$)	0.536	0.473	7.2	1.8
Sampled points (10^6)	0.274	0.378	24	6
DSS [90] (10^5 points)	3.060	0.909	2.4	0.6
SAP [58] (10^4)	0.856	0.764	0.24	0.06
SAP [58] (10^5)	0.659	0.678	2.4	0.6
SAP [58] (10^6)	0.536	0.640	24	6
NGLOD [80] (LoD=4)	2.862	0.755	5.61	1.35
NGLOD [80](LoD=5)	1.832	0.748	38.7	10.14
NGLOD [80] (LoD=6)	1.776	0.696	300	78.18
NGP [46] (T=11)	3.429	0.887	0.29	0.07
NGP [46](T=16)	0.422	0.681	7.21	1.8
NGP [46] (T=19)	0.361	0.468	48.8	12.2
Ours (10^4 points)	3.361	0.593	0.24	0.06
Ours (10^5)	0.765	0.434	2.4	0.6
Ours ($3 \cdot 10^5$)	0.418	0.377	7.2	1.8
Ours (10^6)	**0.243**	**0.298**	24	6

Table 1. *Static surface reconstruction with point sets.* Results reported on the Lego excavator 3D scene. $\mathcal{L}_{CD}$: Chamfer distance ($\times 10^{-4}$). $\mathcal{L}_n$: normal consistency. See Section 4.1 for details.

$$\mathcal{L}(\mathbf{X}^{\mathcal{C}}, \theta, \mathbf{V}_S, \mathbf{M}_{gt}^{\sim}) = \lambda_S \mathcal{L}_S + \lambda_{iso} \mathcal{L}_{iso} + \lambda_V \mathcal{L}_V. \quad (19)$$

We set $\lambda_S = 0, \lambda_V = 1, \lambda_{iso} = 0.1$ in the early epochs to allow accelerated learning of large-scale deformations, and then gradually increase the weight of a surface loss to allow high-frequency deformation modeling. More details and discussion on the optimisation parameters are provided in the supplementary.

Avatar animation. Setting $\lambda_S = 0$ also allows us to perform a single scan animation, as we will demonstrate in Section 4.3. In this setup, the optimisation will deform a mesh to the desired target pose while respecting the introduced geometric constraints, similar in spirit to classic algorithms in shape manipulation [23, 76, 77].

4. Experiments

4.1. Canonical Surface Reconstruction

For the static surface reconstruction example, we consider three types of baselines. First, a straightforward candidate for comparison is a point cloud acquired directly by sampling points and respective normals from the ground

truth surface, with no training involved. This simple method of storing surfaces gives good Chamfer distance reconstruction results by its very nature, since the surface is represented by its samples. However, it suffers from visual artifacts when rendering due to the overlapping points and noisy normal directions (see Figure 3a). Optimising these initial point sets with our method not only produces better normal directions but also further improves Chamfer distance metrics, since the optimisation drives the points to be better distributed along the surface. This is especially noticeable in the case of a limited sample size (Table 1).

Next, we compare our optimised clouds to the two state-of-the-art implicit methods for static surface overfitting: neural geometric level of detail (NGLOD [80]) and instant neural graphics primitives (NGP [46]). Contrary to earlier works [12, 48, 52, 73, 74], these methods are able to represent extremely complex surfaces, and hence they appear in the focus of the present work. For NGLOD, we explore the quality-memory trade-off by varying the level of detail parameter of the method (LOD). For the NGP method, same can be achieved by varying the size T of a hash table that stores spatial features. Our explicit point-based method clearly outperforms both implicit methods, while using a similar or smaller number of parameters (Table 1). This motivates our usage of points for representing canonical space when modeling deformations. We also compare with two alternative point optimisation techniques [58, 90], and provide a discussion on the difference between approaches in the appendix. The Shape-As-Points technique [58] can be complementary to our basic pipeline in the case of reconstruction from unoriented, noisy point clouds.

4.2. Learning Deformation Fields

Our deformation model is based on a neural network that maps every point in the canonical space to a point in the deformed space. The idea was introduced before in the context of deformable radiance fields [53, 64]. For implicit signed distance fields, a similar deformation framework has been introduced in the context of deformable human modeling [51]. In this section, we will show that learning deformation networks directly on explicit graphics primitives bring a number of benefits, including faster training, higher reconstruction fidelity, and lower memory requirements. Another additional benefit of working with explicit surfaces is the ability to directly enforce well-known constraints, such as as-isometric-as-possible deformation.

Datasets and metrics. To achieve this goal, we select target and source deformation surfaces from several datasets of varying complexity: 10 pairs for the DeformingThings4D dataset (Figure 2*c*), 5 pairs from the Owlii dynamic scans [88] (Figure 2*b*), 5 pairs from the synthetic Resynth dataset [42] and a single source-target pair of a Lego scene (Figure 2*e*). These datasets imply different combinations of surface complexity and non-rigidity at different scales (skeletal motion, high-frequency cloth wrinkles), making it a hard task for deformation learning. We use two types of metrics to compare various deformation architectures and supervision regimes. First, for all datasets, we measure how well the deformation can match the target shape. We use the same Chamfer distance and normal metrics for this as in the previous section. Second, for the DeformingThings4D dataset and Lego, we additionally measure the point registration quality by comparing the endpoint error (EPE) and strict and relaxed 3D accuracies Acc_S, Acc_R, metrics utilized in the current state-of-the-art approach for non-rigid point cloud registration [33].

Baselines. For the SDF baseline, we consider the SDF deformation framework introduced in [51]. The main focus of this section is to compare the deformation learning process on SDFs and explicit primitives. We, therefore, are leaving aside the latent shape and pose space models also introduced by the approach, and simply compare the capability to learn a single non-trivial deformation of a surface. We also compare to the recent state-of-the-art non-rigid shape registration method (NDP [33]), which has been shown to outperform a large number of previously introduced baselines [3, 10, 16, 47, 50, 65]. The method is generally similar in spirit to our proposed point deformation network, varying on the utilized network architectures and deformation constraints. While this approach has been shown to perform robustly on the data with relatively smooth and small deformations [32], it is interesting to consider its applicability in the case of rapid body motion and high-frequency surface deformations. We have also experimented with the SE-3 variant of deformation parametrization introduced in [53], as well as neural scene flow prior approach [31].

We consider both unsupervised and weakly supervised scenarios for learning deformations. In the case of ReSynth, this supervision comes in the form of correspondence between minimally clothed human bodies in the target and source scans. For all the pairs, we can also obtain guidance supervision by running Lepard [32] keypoint matcher; however, the keypoint matching approach often fails on the parts such as loose clothing (see appendix), motivating the need for a more robust weakly supervised method.

Results. The results are presented in Table 2 and Figures 4, 5. Without the full supervision of scene flow originally implied by the framework and rarely available in practice, the SDF baseline NPMs [51] fails to find a plausible deformation in case of complex rapid motion, resulting in high reconstruction error (Table 2, row 1, results for DeformingThings4D, ReSynth) and severe visual artifacts (Figures 4*c*, 5*b*). State-of-the-art point cloud registration approach NDP [33] provides the best results among other baselines; however, the constraints baked into the architecture do not allow it fully deform into the desired shape and model high-

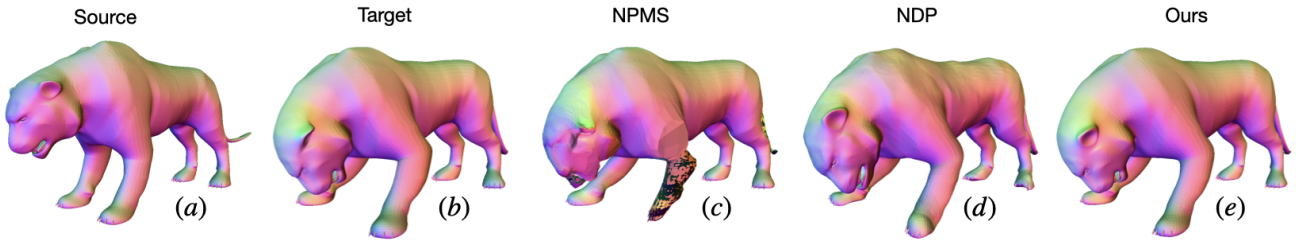

Figure 4. *Unsupervised learning of non-rigid deformations.* We compare our deformation model to both SDF-based model [51] and state-of-the-art non-rigid point cloud registration techniques [33] on diverse types of surfaces, and show qualitative and quantitative advantages of our model, both in terms of reconstruction and registration quality. Please pay attention to the reconstruction of the paws and ear locations of the presented feline model. See Section 4.2 for details.

	DeformingThings4D [34]					Lego [44]					Owlii [88]		ReSynth [42]	
Method	$\mathcal{L}_{CD}\downarrow$	$\mathcal{L}_n\downarrow$	EPE $\downarrow$	$Acc_S\uparrow$	$Acc_R\uparrow$	$\mathcal{L}_{CD}\downarrow$	$\mathcal{L}_n\downarrow$	EPE $\downarrow$	$Acc_S\uparrow$	$Acc_R\uparrow$	$\mathcal{L}_{CD}\downarrow$	$\mathcal{L}_n\downarrow$	$\mathcal{L}_{CD}\downarrow$	$\mathcal{L}_n\downarrow$
1 NPMs [51]	198.8	0.305	0.233	27.2	34.6	9.351	0.432	0.012	91.1	93.0	2.168	0.072	312.2	0.401
2 NSFP [31]	16.3	0.38	0.162	30.6	47.6	0.644	0.823	0.021	68.3	90.1	6.789	0.309	5.33	0.449
3 Nerfies [53]	12.1	0.351	0.211	41.7	53.2	1.458	0.883	0.039	37.2	71.6	1.631	0.305	3.345	0.409
4 NDP [33]	2.819	0.178	0.122	39.4	63.1	0.759	0.896	0.021	71.1	97.8	1.365	0.221	3.889	0.372
5 Ours (w/o $\mathcal{L}_{iso}$)	1.045	0.11	0.222	27.2	34.4	0.376	0.713	0.008	96.1	**99.2**	**0.111**	**0.068**	0.613	0.112
6 Ours (w. $\mathcal{L}_{iso}$)	**0.786**	**0.076**	**0.103**	**59.8**	**69.1**	**0.166**	**0.416**	**0.003**	**97.1**	98.3	0.123	0.079	**0.309**	**0.096**
Supervised w. [32]:														
7 LNDP [33]	3.769	0.192	0.115	46.7	69.1	1.051	0.896	0.026	54.2	94.2	3.559	0.27	9.959	0.489
8 Ours (w. $\mathcal{L}_{iso}$)	**0.644**	**0.066**	**0.099**	**72.6**	**80.4**	**0.204**	**0.495**	**0.004**	**96.8**	**99.3**	**0.116**	**0.081**	0.188	0.097
Supervised w. SMPL-X:														
9 NPMs [51]	-	-	-	-	-	-	-	-	-	-	-	-	7.258	0.164
10 Ours (w. $\mathcal{L}_{iso}$)	-	-	-	-	-	-	-	-	-	-	-	-	**0.165**	**0.071**

Table 2. *Dynamic surface modeling with point fields.* See Section 4.2 for the detailed discussion.

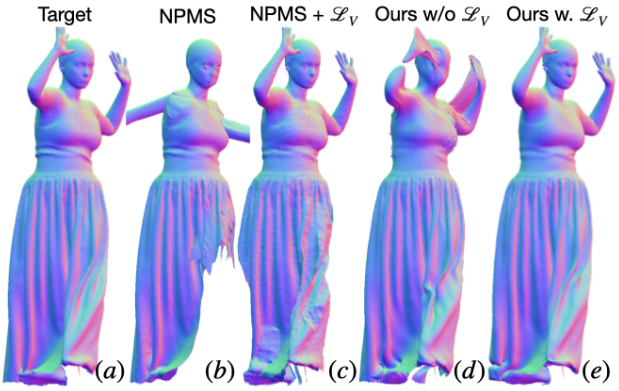

Figure 5. *Learning deformations with SMPL-X vertex guidance on SDFs and point sets.* The usage of the introduced SMPL-X guidance regime benefits both SDF baseline [51] and our reconstructions in the case of the challenging ReSynth scans [42], with our method providing better reconstruction in the areas where no supervision is available. Please pay attention to the reconstruction of the subject's dress.

frequency details, leading to higher reconstruction errors compared to our method (Table 2, r.4) and geometry distortions (Figure 4*d*, mind the animal model paws). We provide more side-by-side comparisons with the method in the appendix. Finally, the usage of the introduced SMPL-X guidance regime benefits both SDF baseline and our reconstructions (Figure 5, Table 2, r.9-10) in the case of the challenging ReSynth scans, with our method still providing significantly lower reconstruction errors.

Training and inference time. For the static reconstruction case (Table 1), the training times of our method, DSS, SAP and NGP are generally comparable and are at the order of several dozens of minutes, while taking several hours for NGLOD on the same GPU configuration. Learning deformation fields (Table 2) is, however, much faster (seconds instead of dozen minutes) on explicit surfaces, since no differentiable SDF representations for the canonical and target surfaces need to be learned and queried.

4.3. Animating Avatars in Challenging Clothing

Finally, we showcase the advantages of our approach for modeling the shapes of 3D humans in challenging clothing.

Problem setup and data. On a high level, the task is to generate plausibly-looking clothed body geometry for the test poses given the {*body pose*, *clothed body shape*} pairs in the training set. We evaluate on the ReSynth [42] dataset. It contains 3D humans in simulated clothing, featuring rich geometric details such as wrinkles and folds that vary with changing body poses. For each subject, the examples in the training- (637 examples) and test-set (347 examples) have different poses, hence different pose-dependent

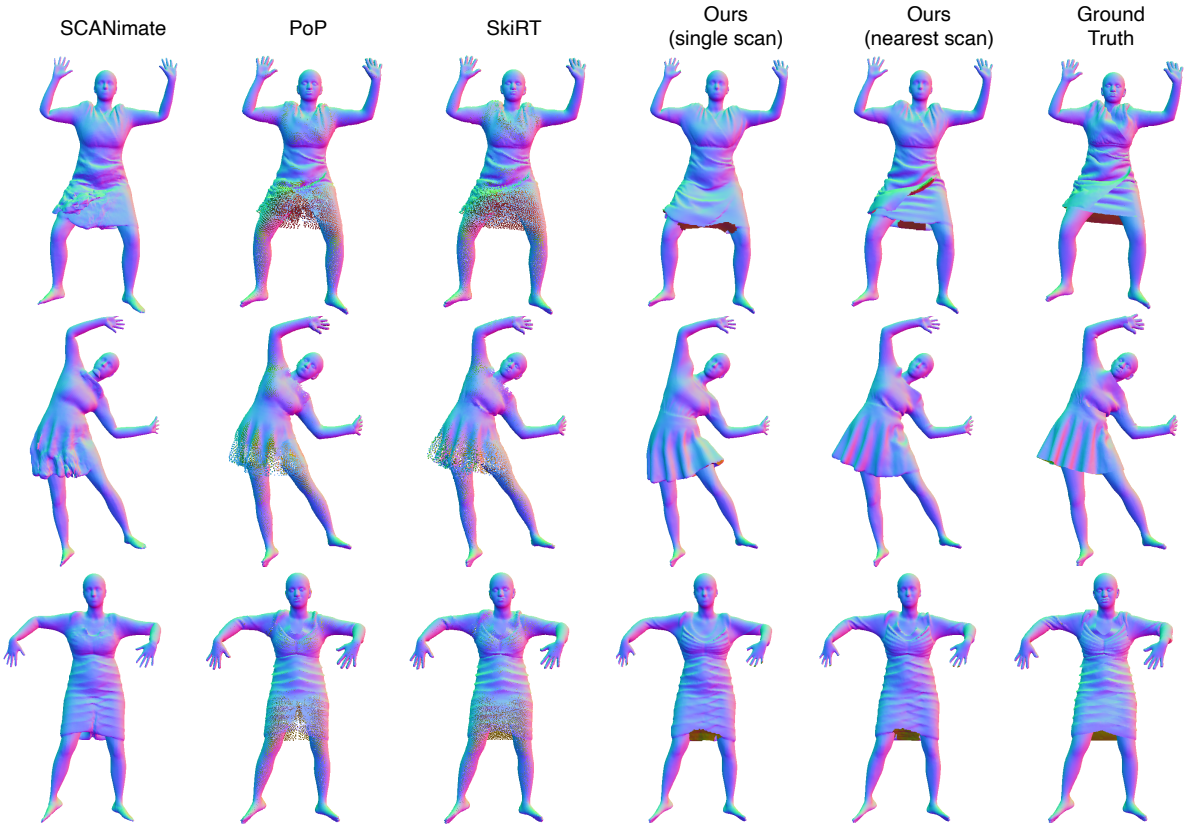

Figure 6. *Qualitative comparison of clothed body shape modeling.* SCANimate [70] results are rendered from the iso-surface extracted from the SDF, and often erroneously produce trousers-like structure for skirts. PoP [42] and SkiRT [40] are rendered with surfel-based renderer [61, 86] to achieve best visual quality, but still suffer from artifacts due to points being sparse on certain regions. The two variants of our approach both produce vivid wrinkles and a globally coherent clothing shape, with a more salient pose-dependent effect when deforming from the nearest training scan. Best viewed zoomed-in on a color screen.

clothing shapes. We specifically choose subjects wearing skirts and dresses, as these clothing types have been a long-standing challenge for learning-based 3D clothed human models.

Baselines. We compare with three state-of-the-art learning-based clothed human models. SCANimate [70] models the clothing geometry using SDFs, whereas PoP [42] and SkiRT [40] deform a point cloud sampled from the unclothed body to represent clothing. Trained in a subject-specific manner, all these methods handle body articulation with linear blend skinning (LBS), and learn a regression from body pose to the clothed body geometry, with an expectation to generalise to novel poses at test time.

Our approach. Fundamentally different from the baselines, we sidestep LBS and the pose space generalisation challenge and tackle the task by optimising deformation fields from a canonical scan, using the guided deformation field learning introduced in Sec. 3.2. Regarding the choice of canonical frames, we consider two alternatives: (1) using a single scan from the training set as the canonical frame; (2) for each test frame, finding its body pose nearest neighbor frame in the training set, and using the scan from that frame as the source of deformation. We provide the details on our pipeline in the appendix.

Evaluation metric. Here we conduct a perceptual study among 30 users to characterise the shape quality and visual plausibility of the results. We randomly sample subjects and test-set poses, and render predictions from all methods under the same setting. The renderings are placed side-by-side with shuffled ordering across examples. For each example, the users are asked to choose the single result with highest quality. Further discussions on the metric and details on the user study are provided in the appendix.

Results. Figure 6 shows the shape prediction of clothed body from all the methods. While all baselines can generate pose-dependent clothing deformations, they exhibit a variety of artifacts. SCANimate [70] suffers from representing thin cloth structures such as the bottom of skirts, a typical challenge for implicit surface representations. The point-based models PoP [42] and SkiRT [40] deliver sharp wrinkles where points are dense (e.g. upper body), but lose expressiveness on low-density regions, especially for loose-

Method	SCANimate [70]	PoP [42]	SkiRT [40]	Ours
User choice%	6.1%	2.9 %	4.9%	**85.8%**

Table 3. *Perceptual study results*. Across all examples, 85.8% users prefer results from our method over others.

fitting clothing. In contrast, our method is free from such artifacts, producing a coherent global shape and sharp local details, essentially animating scans without compromising the original geometry quality. This verifies the effectiveness of our guided deformation learning on highly non-rigid data. The qualitative advantages of our method are also characterised by a clear winning margin in the user study as shown in Table 3. Overall, 85.8% users prefer the results produced by our method for higher geometry quality and visual resemblance to the ground truth. Please see the supplementary for the details on the user study protocol.

From the clothed human modeling perspective, due to the reliance on linear blend skinning, the baseline methods often struggle with loose garments and skirts, resulting in e.g. "splitting" artifacts on skirts as shown in Figure 6. In contrast, our method offers a new paradigm for clothing modeling, which directly optimises for a smooth deformation field that preserves the continuity of cloth surfaces, side-stepping drawbacks of LBS and the challenges in the pose space generalisation. In particular, it can produce plausibly-looking clothing surface even under extreme, out-of-distribution poses (Figure 2*d*), which, to our knowledge, has not been demonstrated before for learning-based clothed human models. More discussion on the clothed human modeling paradigms can be found in the appendix.

5. Conclusion

We have introduced a dynamic point field model that efficiently models non-rigid 3D surfaces by combining explicit point-based graphics with implicit deformation networks. Incorporating established constraints like as-isometric-as-possible regularization is made easy by using explicit surface primitives. Extensive experiments have shown the superiority of our dynamic point field in representational power, learning efficiency and robustness. We have also demonstrated that our framework offers an advantageous new paradigm for animating clothed humans, surpassing the limitations of linear blend skinning-based methods, especially on complex clothing types such as skirts.

While efficient, our model faces several limitations in its current form. First, our method requires learning a small MLP for every deformation, which makes the method unsuitable for real-time avatar animation. Second, it does not explicitly model the dependency of clothing shape on body poses and factors such as motion acceleration. Finally, our deformation optimisation pipeline still struggles to find a plausible deformation when large deformations are present and no guidance is available. Please see the appendix for more discussion on the topic.

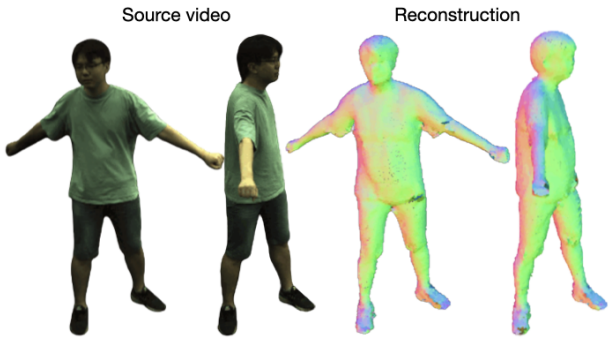

Figure 7. *Dynamic point field reconstruction from multi-view video streams*. As an avenue for future work, we consider the reconstruction of dynamic surfaces from multi-view videos. Early results on the sequence from [60].

Future work can leverage the advantages of our representation to build robust, high-quality pose-dependent human avatars, as well as general dynamic 3D models. A particularly interesting direction here is to combine modern photorealistic point-based renderers [21, 87] with our point deformation model, to allow efficient reconstruction and synthesis of dynamic scenes from multi-view imagery.

Acknowledgements. This work was supported by an ETH Zürich Postdoctoral Fellowship. Qianli Ma is partially funded by the Max Planck ETH Center for Learning Systems. We sincerely thank Marko Mihajlovic, Yan Zhang, Anpei Chen, Shaofei Wang and Dušan Svilarković for the fruitful discussions and manuscript proofreading.

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 696–712. Springer, 2020. 1, 3

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8387–8397, 2018. 3

[3] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 3, 6

[4] Matan Atzmon and Yaron Lipman. SAL: Sign agnostic learning of shapes from raw data. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019. 5

[6] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensoRF: Tensorial radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1

[7] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[8] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-aware generative model for clothed people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[9] Marvin Eisenberger, Zorah Lähner, and Daniel Cremers. Divergence-free shape interpolation and correspondence. *arXiv preprint arXiv:1806.10417*, 2018. 3

[10] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019. 3, 6

[11] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14346–14355, 2021. 1

[12] 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. 3, 6

[13] Markus Gross and Hanspeter Pfister. *Point-based graphics*. Elsevier, 2011. 3

[14] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 3d-coded: 3D correspondences by deep deformation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 230–246, 2018. 3

[15] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. GarNet: A two-stream network for fast and accurate 3D cloth draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8739–8748, 2019. 3

[16] Osamu Hirose. A bayesian formulation of coherent point drift. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 43(7):2269–2286, 2020. 3, 6

[17] Qi-Xing Huang, Bart Adams, Martin Wicke, and Leonidas J Guibas. Non-rigid registration under isometric deformations. In *Computer Graphics Forum*, volume 27, pages 1449–1457. Wiley Online Library, 2008. 2

[18] 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 *International Conference on 3D Vision (3DV)*, pages 421–430. IEEE, 2017. 3

[19] Matthias Innmann, Kihwan Kim, Jinwei Gu, Matthias Nießner, Charles Loop, Marc Stamminger, and Jan Kautz. Nrmvs: Non-rigid multi-view stereo. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2754–2763, 2020. 2, 3

[20] Justin Johnson, Nikhila Ravi, Jeremy Reizenstein, David Novotny, Shubham Tulsiani, Christoph Lassner, and Steve Branson. Accelerating 3D deep learning with PyTorch3D. In *SIGGRAPH Asia 2020 Courses*, pages 1–1. 2020. 2, 4

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 9

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 3

[23] Martin Kilian, Niloy J Mitra, and Helmut Pottmann. Geometric modeling in shape space. In *ACM SIGGRAPH*, pages 64–es. 2007. 1, 2, 4, 5

[24] Leif Kobbelt and Mario Botsch. A survey of point-based techniques in computer graphics. *Computers & Graphics*, 28(6):801–814, 2004. 3

[25] Georgios Kopanas, Thomas Leimkühler, Gilles Rainer, Clément Jambon, and George Drettakis. Neural point catacaustics for novel-view synthesis of reflections. *ACM Transactions on Graphics*, 41(6):Article–201, 2022. 3

[26] Christoph Lassner and Michael Zollhofer. Pulsar: Efficient sphere-based neural rendering. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1440–1449, 2021. 1, 3

[27] 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*, pages 6624–6634, 2022. 3

[28] Marc Levoy and Turner Whitted. *The use of points as a display primitive*. Citeseer, 1985. 3

[29] Kejie Li, Trung Pham, Huangying Zhan, and Ian Reid. Ef-

ficient dense point cloud object reconstruction using deformation vector fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 497–513, 2018. 3

[30] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3D video synthesis from multi-view video. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5521–5531, 2022. 3

[31] Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Neural scene flow prior. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:7838–7851, 2021. 6, 7

[32] Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5, 6, 7

[33] Yang Li and Tatsuya Harada. Non-rigid point cloud registration with neural deformation pyramid. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6, 7

[34] Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4DComplete: Non-rigid motion estimation beyond the observable surface. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 7

[35] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[36] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *Advances in Neural Information Processing Systems*, 35:36762–36775, 2022. 1

[37] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019. 2

[38] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, 2015. 3, 5

[39] Marcel Lüthi, Thomas Gerig, Christoph Jud, and Thomas Vetter. Gaussian process morphable models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 40(8):1860–1873, 2017. 3

[40] Qianli Ma, Jinlong Yang, Michael J. Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. In *International Conference on 3D Vision (3DV)*, 2022. 2, 3, 8, 9

[41] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[42] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021. 1, 2, 3, 6, 7, 8, 9

[43] 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)*, pages 4460–4470, 2019. 1, 3

[44] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 7

[45] Michael I Miller, Alain Trouvé, and Laurent Younes. Geodesic shooting for computational anatomy. *Journal of mathematical imaging and vision*, 24(2):209–228, 2006. 3

[46] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):102:1–102:15, July 2022. 1, 2, 3, 5, 6

[47] Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(12):2262–2275, 2010. 3, 6

[48] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4D reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5379–5389, 2019. 3, 6

[49] David Novotny, Ignacio Rocco, Samarth Sinha, Alexandre Carlier, Gael Kerchenbaum, Roman Shapovalov, Nikita Smetanin, Natalia Neverova, Benjamin Graham, and Andrea Vedaldi. Keytr: Keypoint transporter for 3D reconstruction of deformable objects in videos. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5595–5604, 2022. 2

[50] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics*, 31(4):1–11, 2012. 3, 6

[51] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. NPMs: Neural parametric models for 3D deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12695–12705, 2021. 2, 6, 7

[52] 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)*, June 2019. 1, 3, 6

[53] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 6, 7

[54] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Transactions on Graphics*, 40(6), dec 2021. 3

[55] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. TailorNet: Predicting clothing in 3D as a function of

human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 3

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

[57] Songyou Peng, Chiyu Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable poisson solver. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:13032–13044, 2021. 1

[58] Songyou Peng, Chiyu "Max" Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable poisson solver. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3, 5, 6

[59] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[60] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. NeuralBody: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 9054–9063, 2021. 9

[61] Hanspeter Pfister, Matthias Zwicker, Jeroen Van Baar, and Markus Gross. Surfels: Surface elements as rendering primitives. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 335–342, 2000. 3, 8

[62] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 34(4):120:1–120:14, Aug. 2015. 3

[63] Sergey Prokudin, Michael J Black, and Javier Romero. SMPLpix: Neural avatars from 3D human models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1810–1819, 2021. 3

[64] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 6

[65] Gilles Puy, Alexandre Boulch, and Renaud Marlet. Flot: Scene flow on point clouds guided by optimal transport. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII*, pages 527–544. Springer, 2020. 3, 6

[66] Ruslan Rakhimov, Andrei-Timotei Ardelean, Victor Lempitsky, and Evgeny Burnaev. Npbg++: Accelerating neural point-based graphics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15969–15979, June 2022. 1

[67] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. KiloNeRF: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14335–14345, 2021. 1

[68] Darius Rückert, Linus Franke, and Marc Stamminger. ADOP: Approximate differentiable one-pixel point rendering. *ACM Transactions on Graphics*, 41(4):1–14, 2022. 1, 3

[69] Szymon Rusinkiewicz and Marc Levoy. Qsplat: A multiresolution point rendering system for large meshes. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 343–352, 2000. 3

[70] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–2897, 2021. 2, 3, 8, 9

[71] Jieqi Shi, Lingyun Xu, Liang Heng, and Shaojie Shen. Graph-guided deformation for point cloud completion. *IEEE Robotics and Automation Letters*, 6(4):7081–7088, 2021. 3

[72] Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. MetaSDF: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:10136–10147, 2020. 4

[73] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 3, 4, 6

[74] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 6

[75] Liangchen Song, Xuan Gong, Benjamin Planche, Meng Zheng, David Doermann, Junsong Yuan, Terrence Chen, and Ziyan Wu. Pref: Predictability regularized neural motion fields. In *European Conference on Computer Vision*, pages 664–681. Springer, 2022. 3

[76] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116, 2007. 5

[77] Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H-P Seidel. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 175–184, 2004. 5

[78] Carsten Stoll, Zachi Karni, Christian Rössl, Hitoshi Yamauchi, and Hans-Peter Seidel. Template deformation for point cloud fitting. In *PBG@ SIGGRAPH*, pages 27–35, 2006. 3

[79] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM siggraph 2007 papers*, pages 80–es. 2007. 3

[80] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021. 1, 2, 3, 5, 6

[81] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727. Wiley Online Library, 2020. 1, 3

[82] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-GIF: Neural generalized implicit functions for animating people in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 3

[83] Wolfram Von Funck, Holger Theisel, and Hans-Peter Seidel. Vector field based shape deformations. *ACM Transactions on Graphics*, 25(3):1118–1125, 2006. 3

[84] Michael Wand, Philipp Jenke, Qixing Huang, Martin Bokeloh, Leonidas Guibas, and Andreas Schilling. Reconstruction of deforming geometry from time-varying point clouds. In *Symposium on Geometry Processing*, pages 49–58, 2007. 3

[85] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. MetaAvatar: Learning animatable clothed human models from few depth images. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. 3, 4

[86] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7467–7477, 2020. 1, 2, 3, 4, 8

[87] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-NeRF: Point-based neural radiance fields. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5438–5448, 2022. 1, 3, 9

[88] Yi Xu, Yao Lu, and Ziyu Wen. Owlii dynamic human textured mesh sequence dataset. In *ISO/IEC JTC1/SC29/WG1 1 input document m41658*, 2017. 6, 7

[89] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022. 2

[90] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (proceedings of ACM SIGGRAPH ASIA)*, 38(6), 2019. 1, 3, 5, 6

[91] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 1

[92] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14718–14727, 2021. 2, 3

[93] Qiang Zhang, Seung-Hwan Baek, Szymon Rusinkiewicz, and Felix Heide. Differentiable point-based radiance fields for efficient view synthesis. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–12, 2022. 1