Inferring Dense Human Representation from Sparse or Incomplete Point Clouds

Boyao Zhou

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Informatique
Unité de recherche : Laboratoire Jean Kuntzmann

Inférer une représentation dense de l'humain avec un nuage de points épars ou incomplet

Inferring Dense Human Representation from Sparse or Incomplete Point Clouds

Présentée par :

Boyao ZHOU

Direction de thèse :

Edmond BOYER INRIA	Directeur de thèse
Jean-Sébastien FRANCO Université Grenoble Alpes	Co-directeur de thèse

Rapporteurs :

MATHIEU AUBRY
Ingénieur HDR, ENPC PARIS
TONY TUNG
Ingénieur de recherche, META Reality Labs

Thèse soutenue publiquement le **22 novembre 2022**, devant le jury composé de :

EDMOND BOYER Directeur de recherche, INRIA CENTRE GRENOBLE-RHONE- ALPES	Directeur de thèse
MATHIEU AUBRY Ingénieur HDR, ENPC PARIS	Rapporteur
TONY TUNG Ingénieur de recherche, META Reality Labs	Rapporteur
REMI RONFARD Directeur de recherche, INRIA CENTRE GRENOBLE-RHONE- ALPES	Président
GUILLAUME LAVOUE Professeur, EC NAT INGENIEURS ST- ETIENNE	Examinateur
SERGI PUJADES ROCAMORA Maître de conférences, UNIVERSITE GRENOBLE ALPES	Examinateur

Invités :

JEAN-SEBASTIEN FRANCO
MAITRE DE CONFERENCE HDR, Grenoble INP

Abstract

With the development of 3D vision techniques, in particular neural network based methods, the 3D neural avatar representation has gained growing interest both in the academia and in the industry. Such digital representations have been applied to movie, video game, fashion and virtual reality environments to enrich user experiences. In terms of 3D representation and reconstruction, classic methods rely on heavy setups and costly computation processes while neural networks open up the possibility to handle this problem from partial observations thanks to tolerance to insufficient information. In particular, neural networks achieve promising results for reconstruction tasks with a high speed inference process. However, at the time of the thesis few neural network designs used spatial constraint on the 3D human shape and temporal coherence of motion for dense reconstruction and completion. This thesis proposes to build 3D and 4D models for dense human shape estimation/reconstruction from sparse or incomplete point clouds and investigates how the proposed network and training strategy contribute. To evaluate the effectiveness of the proposed methods, we collect data from synthetic and real datasets, with dressed humans and undressed humans. We first examine a static intermediate task, in which we deform the key points of a reference template to fit to input sparse point clouds and densify the deformed points with our proposed Gaussian Process layer. Our Gaussian Process layer enforces the smoothness of 3D geometry and adversarial training can further improve robustness across the datasets, which allow us to reconstruct 3D human shapes from sparse unstructured point clouds and avoid local optima during inference. Instead of static frame-by-frame poses, humans perform dynamic motion in daily life. We thus examine temporal continuity in dense shape inference. We develop a continuous representation of human motion sequences from partial observations with neural implicit modeling, which enables to complete spatial information and to enhance temporal frames. Our proposed method outperforms the static methods which lack temporal coherence, by correcting artefacts due to holes or noises. However we are still missing some high-frequency details in our results when using a naive training strategy. Therefore, we investigate how to represent fine details of humans with a coarse-to-fine strategy and temporal feature aggregation from an input sequence of depth images. This allows us to pyramidally learn a signed distance field in both spatial and temporal directions in order to recover fine details on cloth wrinkles and facial expressions.

Keywords. 3D human body modeling • implicit representation • temporal modeling • coarse-to-fine

Résumé

Avec le développement de techniques de vision 3D, en particulier des méthodes basées sur les réseaux neuronaux, la représentation de l'avatar neuronal 3D a suscité un intérêt croissant à la fois dans l'académie et dans l'industrie. Une telle représentation numérique a été appliquée aux environnements du cinéma, du jeu vidéo, de la mode et de la réalité virtuelle pour enrichir l'expérience utilisateur. En termes de représentation et de reconstruction 3D, les méthodes classiques reposent sur une mise en place lourde et des processus de calcul coûteux tandis que les réseaux neuronaux ouvrent la possibilité de traiter ce problème à partir d'observations partielles grâce à une meilleure tolérance aux informations insuffisantes. En particulier, les réseaux neuronaux obtiennent des résultats prometteurs pour les tâches de reconstruction avec un processus d'inférence à grande vitesse. Cependant, au moment de la thèse la contrainte spatiale sur la forme humaine et la cohérence temporelle du mouvement est peu ou pas prise en compte dans la conception des réseaux neuronaux pour la reconstruction dense et la complétion. Cette thèse propose de construire des mèthodes 3D et 4D pour l'estimation/reconstruction de formes humaines à partir de nuages de points épars ou incomplets et étudie comment le réseau proposé et la stratégie d'apprentissage contribuent. Pour évaluer l'efficacité des méthodes proposées, nous collectons des données à partir d'un jeu de données synthétiques et réelles, avec des humains habillés et des humains sans vêtements. Nous examinons d'abord une tâche intermédiaire statique, dans laquelle nous déformons les points clés de la référence(template) pour épouser la forme des nuages de points épars d'entrée et densifions les points déformés avec notre couche de processus gaussien proposée. Notre couche de processus gaussien renforce le lissage de la géométrie 3D et l'apprentissage adversarial peut encore améliorer la robustesse sur les ensembles de données, qui nous permettent de reconstruire des formes humaines 3D à partir de nuages de points épars non structurés et éviter les optimums locaux pendant l'inférence. Au lieu de poses statiques image par image, les humains effectuent des mouvements dynamiques dans la vie quotidienne. Donc nous examinons la continuité temporelle dans l'inférence de forme dense. Nous développons une représentation continue des séquences de mouvements humains à partir d'observations partielles avec modélisation neuronale implicite, qui permet de compléter l'information spatiale et d'augmenter la frèquence des sèquences d'entrèe. Notre méthode proposée surpasse les méthodes statiques qui manquent de cohérence temporelle

en corrigeant les artefacts provoqués par des données manquantes ou bruitées. Mais il nous manque encore des détails à haute fréquence dans nos résultats avec une stratégie d'entraînement naïve. Par conséquent, nous étudions comment représenter les détails fins de l'humain avec une stratégie hiérarchique et l'agrégation des caractéristiques temporelles à partir d'une séquence d'entrée des images de profondeur. Cela nous permet d'apprendre de manière pyramidale le champ de distances signées dans les directions spatiale et temporelle afin de récupérer des détails fins sur les plis des vêtements et les expressions faciales.

Mots Clés. modélisation d'humain 3D • représentation implicite • modélisation temporelle • stratégie hiérarchique

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Digital Humans have become a trending topic in the last few years with booming applications such as movie, video game, telepresence, virtual/augmented reality and computer-assisted coaching. A digital human is an avatar which has similar image to human and can be manipulated and animated in the digital world. Our general goal is to understand and represent humans in the digital world with novel methods in computer vision and graphics.

Capture and measurement of real humans from sensors are essential, in particular with customer commodity devices, in quantifying the variation of the population and in preparing a data-driven model. Some 2D contents for human body have been well studied from the RGB image for the past decade in the research community, *e.g.* keypoints Cao et al. [2019], Raaj et al. [2019], silhouette Chen et al. [2018a], Liu et al. [2020a] and style editing Sarkar et al. [2021]. However, 3D human raw shape capture relies still on costly devices Bartol et al. [2021], *e.g.* passive stereo, structured light and time-of-flight camera. The static 3D scanning Lanman and Taubin [2009], Yamaguchi et al. [2014] achieves high quality reconstruction of millimetric precision, with detailed wrinkles on the cloth, at a high speed computation which gets the success in fashion and movie industry, however it is restricted to small scale capture. Moreover, missing parts in the acquisition occur frequently due to partial observations and occlusion, so heavy manual postprocessing is required by artists. Recently, with the development of processing and storage techniques, dynamic modeling has gained more attention which maintains temporal continuity with respect to static frame-by-frame scanning. We can also take the benefit of temporal information to overcome the issue of partial observation, especially with the data-driven methods Li et al. [2021]. In this thesis, we take interest in building dynamic model for reconstructing dense human shape from monocular depth-like input, which helps

us to avoid the fundamental depth ambiguity and is encountered in many capture setups. We come up with learning-based methods which allow us to handle missing observations. The captured dynamic raw data generally exhibits noise which is not clean enough to serve as ground truth for training. So we also study the problem of reconstructing template-aligned mesh from raw captured data, unstructured point cloud.

The raw data is processed or "labeled" in order to have a clear insight of the variation of the human identities. The alignment and the correspondence are considered as the "label" in this task. To this end, classic approaches rely on parameter model fitting. For example, SCAPE Anguelov et al. [2005] parametrizes the identity dependent deformation with principal component analysis PCA. Another famous family of parameter model, SMPL Loper et al. [2015], models shape identity and pose and represents the pose dependent deformation on top of dual quaternion blend skinning Kavan et al. [2007, 2008]. The difficulty of fitting parameter model to raw data is how to initialize the parameter during the optimization process as the classic optimization approach is susceptible to local optima. In recent research, data-driven methods Li et al. [2019], Groueix et al. [2018a] in general infer the start parameters after the training phase, furthermore the optimization of latent parameters could still be conducted to reach the global optima.

Although the parameter model handles the problem of modeling shape identity and pose simultaneously, clothing is not included into the parameter model. These models do not model soft-tissue deformation and clothing and thus can not represent detailed surface phenomena. Moreover, the deformation of cloth is non-rigid and should be heavily correlated with the human shape and motion. Pons-Moll et al. [2017], Alldieck et al. [2019a] propose to parametrize cloth as the displacement layer from undressed character. However the pre-defined topology in parameter model restricts the representable capability of cloth styles, in particular the cloth fitting to different identities. Recently, implicit representation along with neural networks Saito et al. [2019, 2020] gained attention in research community because such representation can model shape and cloth at the same time.

As mentioned previously, scalability of current scanning system is small due to the limitation of optical intensity. Moreover, sensor array in such system is generally expensive and a hand-crafted process is typically required to figure

out noisy raw data by experienced artists. In contrast, commodity depth sensors, such as Microsoft Kinect and Intel RealSense, are widely used Newcombe et al. [2011, 2015], Yu et al. [2017], Dai et al. [2017a], Yu et al. [2018] to avoid the high cost of equipment. Thus, shape reconstruction/completion from a single depth camera is an attractive solution for the reason of easy set-up and of low cost. However such setup suffers inherent incompleteness Newcombe et al. [2011] On angle is to process streams of a single moving sensor and fuse them into a canonical model Newcombe et al. [2015], but this applies only to specific scenarios and requires sensor calibration and localization.

Data-driven methods deal with this ill-posed problem and give promising results thanks to the tolerance to missing information. The goal is to learn the mapping from partial data to completed shape which could recover details from visible region and complete surface in invisible parts. To this end, data collection is important, in which the input data might contain the imperfect such as noisy surface and holes and completed ground truth should be provided for the learning purpose. Synthetic data with proper ground truth contributes a lot in this research direction Saito et al. [2019, 2020], however the perfect surface normally does not contain any noise to challenge robustness of the proposed approach. Some open source data is also available to train with, however some processing such as parameter model fitting is required on top of raw data in order to prepare ground truth. Given the interest received from research community, body shape data such as SCAPE Anguelov et al. [2005], FAUST Bogo et al. [2014], DFAUST Bogo et al. [2017] and AMASS Mahmood et al. [2019] and clothed person von Marcard et al. [2018], Ma et al. [2020] are presented in the last decade. Some aforementioned datasets provide not only raw scan data but also template fitted ground truth. Since ground truth data share the same topology as template, the correspondence can be built. The correspondence plays an important role in dynamic modeling. On one hand, we can overcome occlusion issue with correspondence between visible and invisible parts in different frames. On the other hand, fine detail recovery is priority and difficult to achieve, which relies on the corresponding feature aggregation. Aforementioned correspondence is established through template-aligned mesh. We call template deformation technique as registration. Classic registration methods rely on manually designed markers which initialize the optimization process. Deep neural network is another direction. For example, Groueix et al. [2018a], Deprelle et al.

[2019] give a good initialization with data-driven inference and provide a further optimization process from such initialization. They can still benefit from inclusion of a human topological prior, which we explore. We thus examine in this thesis how to improve template deformation technique by considering spatial constraint with our Gaussian kernel. We take point clouds from scan data as input, especially sparse point clouds, to deform the pre-defined template in order to fit the input points. Once watertight mesh is built from raw scan data and the correspondence in sequence is done, we also take interest in modeling human shape continuity in both space and time dimensions from a sequence of single-view depth-like input in data-driven techniques. First, single-view reconstruction is challenging with respect to multi-view stereo setup. Second, dynamic component, temporal continuity, is little or not concerned in existing works Saito et al. [2019, 2020], Chibane et al. [2020a]. Third, we would like to take into account all spatial components such as human shape, cloth style and human pose. So we explore using spatio-temporal implicit function to represent input sequence. Furthermore, we examine how to combine spatial and temporal aspects with a novel pyramidal learning strategy in order to preserve fine details and temporal consistency.

1.1 Outline & Contributions

Our first contribution is a detailed review of building a human body model in Chapter 2, including the raw data acquisition, the data registration, static modeling and temporal modeling.

In Chapter 3, we present the Gaussian Process network which integrates shape prior with Gaussian kernel into network architecture for the purpose of template matching. We leverage Gaussian Process layer to encode surface smoothness and local coherence. In addition, an adversarial training strategy is applied in order to improve the generalization capability. The proposed model is validated in both synthetic dataset SURREAL Varol et al. [2017] and real dataset FAUST Bogo et al. [2014]. Our first approach achieves reconstructing template-aligned mesh from sparse point clouds, but only in the context of undressed human.

In Chapter 4, we tackle the problem of inferring the complete human shape in a sequence from single view. We leverage implicit neural modeling to give a continuous representation in both spatial and temporal directions. The proposed

approach could not only complete the shape from partial observations but also enhance the temporal frames. We evaluate it in both undressed person dataset DFAUST Bogo et al. [2017] and dressed human dataset CAPE Ma et al. [2020]. Our approach provides plausible results but fine details such as cloth wrinkles are still missing and correspondence across sequence frames is not concerned.

In Chapter 5, we study the problem of aggregating the features in a coarse-to-fine manner in order to recover the high-frequency details from input depth frame itself and neighboring frames. We leverage temporal redundancy with spatio-temporal features and carefully design the corresponding training strategy to allow us to learn a rich latent space and to achieve promising results.

Finally, we conclude and discuss future directions in Chapter 6.

In summary, we propose the following contributions:

- We introduce the shape prior and adversarial loss into neural network to better constrain the network during training.
- We build an implicit neural network model to continuously represent the 3D human body in both spatial and temporal directions.
- We efficiently recover the high-frequency details from a sequence of partial input in the coarse-to-fine manner.

1.2 Publications

The contributions are summarized and the material is conducted by the publications:

Chapter 3:

- *Reconstructing Human Body Mesh from Point Clouds by Adversarial GP Network*
  ACCV2020 - Asian Conference on Computer Vision.
  Boyao Zhou, Jean-Sébastien Franco, Federica Bogo, Bugra Tekin, and Edmond Boyer

Chapter 4:

- *Spatio-Temporal Human Shape Completion with Implicit Function Networks*
  3DV2021 - International Conference on 3D Vision.
  Boyao Zhou, Jean-Sébastien Franco, Federica Bogo, and Edmond Boyer

Chapter 5:

- *Pyramidal Signed Distance Learning for Spatio-Temporal Human Shape Completion*
  ACCV2022 - Asian Conference on Computer Vision.
  Boyao Zhou, Jean-Sébastien Franco, Martin delaGorce, and Edmond Boyer

Chapter 2

Background

Through this thesis, we study the problem of modeling human body in both spatial and temporal manners. In this chapter, we would like to give a brief presentation which could serve as the background to better understand the following material. We will review the related works in six categories, human data acquisition, registration, parametric body model, single-view shape completion, temporal modeling and coarse-to-fine detail recovery.

2.1 Human Data Acquisition

With the development of 3D shape acquisition hardware and reconstruction algorithms, the high-quality human dataset contributes a lot in the various research topics for the incoming demand of data-driven methods. In addition, datasets Bogo et al. [2014, 2017], Ma et al. [2020] along with the ground truth allow the state-of-the art approaches, including our methods, to train and to evaluate. Synthetic data, e.g. RenderPeople ren [2018], provides the high-resolution and watertight models which can be easily manipulated. However, such synthetic data might not contain the realistic deformation of human or self contacts. So real data involving the variety of shape and pose information, different clothing styles, high-frequency details and sometimes the non-rigid deformation mapping is typically used by data-driven model. We provide in this section an analysis of acquisition techniques. On one hand, we identify that acquired data is fed into our networks to model the static or dynamic human space. On the other hand, we would like to simplify the acquisition system with less constraints. Current acquisition system is looking forward the high-resolution, accurate and robust results with the requirement of low-latency and rapid processing. To this end, two possibilities in general exist. One is to reconstruct 3D

data in multi-view stereo setup of multi-view RGB images. Another choice is to scan whole body directly with 3D measurement sensor, *e.g.* structured light or time-of-flight senor. The former one is called passive stereo, and latter one active scanning Bartol et al. [2021]. The RGB camera system is less expensive than the high-fidelity depth sensor in general case, and we can capture fast motion in large-scale space covering several meters as an advantage. However, they both rely on heavy system which contains dozens of cameras or depth sensors. Calibration and synchronization of camera system is complicated. Our proposed static and dynamic models in this thesis are intended to replace such systems with a simple setup of sparse inputs and single monocular sensor.

Multi-view stereo reconstructs the 3D scene with the concept of binocular matching, *i.e.* matching the same point on two images from two viewing directions and finding the intersection point in 3D world, which is also called triangulation Hartley and Zisserman [2003]. The corresponding point pair can be solved by estimating pair-wise disparity as Pollefeys et al. [1998, 2004] or by feature matching as Furukawa and Ponce [2009], Geiger et al. [2011], Bleyer et al. [2011], Galliani et al. [2016]. The depth is then estimated by changing camera view Pollefeys et al. [2004], by windows-marching with normalized cross-correlation Goesele et al. [2006], or by contour tracking Salman and Yvinec [2009]. Such depth is then processed with normally high latency volumetric fusion, *e.g.* Curless and Levoy [1996], in order to output the final consistent surface. To capture large-scale performance, multi-view stereo system is typically used with the benefit of high-speed computation and low cost. Moreover, deep neural network techniques, *e.g.* 3D CNN, RNN, can also contribute in this classic 3D vision topic along with the coarse-to-fine strategy, *e.g.* to match the image features with photo consistency Leroy et al. [2018], to inference the depth information Yao et al. [2018, 2019], Yang et al. [2020]. Following the idea of differentiable volumetric rendering Yariv et al. [2020], Niemeyer et al. [2020] and radiance field rendering Yariv et al. [2021], Wang et al. [2021], this topic regains attentions combining the idea of neural network.

3D scanning techniques capture human body based on structured light or time-of-flight sensors. In terms of structured light approach, we have laser-based scanner D'Apuzzo [2007] which sweeps human body from top to bottom and projector-based Wang et al. [2007] one which can project directly complex patterns and scan human body from one view direction at a time. Structured light approach Wang et al. [2012] searches for the correspondence between light

pattern and camera to measure human body which achieves a high quality acquisition but with a slow speed. Alternatively, time-to-flight camera emits and receives light signal and estimate the measurement from object to camera with traveling time. So fast frame capture can be achieved and used in dynamic reconstruction Newcombe et al. [2015] with customer-level sensor, *e.g.* Kinect. However, 3D scanning techniques are limited with illumination source intensity for structured light approach or with light emitter intensity for time-of-flight approach, so scanning range is typically less than 5 meters and with a low resolution.

All kinds of human body data, from multi-view stereo Leroy et al. [2018], 3D scanning Bogo et al. [2014] and synthetic data Varol et al. [2017], boosts research in this topic. In particular, real data contains the realistic self-contact but missing data due to the invisibility during capture, see Figure 2.2. Moreover, the registration of static model and the correspondence/tracking of the dynamic model are also necessary for some research purpose. In this thesis, we study the problem of human reconstruction and registration with the synthetic dataset SURREAL Varol et al. [2017] and real scanning dataset FAUST Bogo et al. [2014], and of spatio-temporal shape completion with the undressed human dataset DFAUST Bogo et al. [2017] and dressed human dataset CAPE Ma et al. [2020].

2.2 Registration & Correspondence

Registration is a long-standing topic in computer vision. In the case of digital human, the goal is to find dense, point-to-point, correspondence between target and reference shape. After acquisition, we can not feed raw data to networks to learn with as ground truth without registration. The raw acquired data is "noisy" in terms of semantic, shape topology and missing data, see Figure 2.1 and 2.2(a). In other words, such data is inconsistent with each other and correspondence from one to another is missing. The registration data accompanying the correspondence has contributed to data-driven methods Groueix et al. [2018a], Deprelle et al. [2019] in the supervised manner. Some of the dynamic models we will propose rely on correspondences to represent temporal continuity and to aggregate the corresponding feature to recover fine details. For this reason, we review in this section registration methods for human body model.

Given a pair of shapes, one is target shape **S** and another is the reference shape/template **T**, the goal is to learn a deformation operation $\mathcal{R}$ which could be used to minimize the distance between two point set,

$$d(\mathbf{S},\mathbf{T}) = \sum_{t_i \in \mathbf{T}} dist(\mathcal{R}(t_i),\mathbf{S}) \tag{2.1}$$

The distance between a point t_i and the point set **S** is normally defined with the closest distance,

$$dist(p,\mathbf{S}) = \min_{s_j \in \mathbf{S}} \quad d(t_i,s_j) \tag{2.2}$$

The distance between two points t_i and s_j can be measured with Euclidean distance. In the case of rigid deformation, the Iterative Closest Point, **ICP** Besl and McKay [1992], algorithm is widely used and the deformation operation $\mathcal{R}$ is represented with global rotation and translation.

To extend this method to non-rigid deformation requires a larger space of $\mathcal{R}$ which should consider point-to-point transformation. Such deformation operation sometimes requires initial correspondence between the reference **T** and the target shape, *e.g.* manually-designed marker or skeleton. **Marker-based** methods Allen et al. [2002, 2003] rely on the estimation of global kinematic and local pose with initial marker positions and the reference can be fitted to raw scans with displacement map. **Marker-less** methods Anguelov et al. [2004], Huang et al. [2008] solve point-to-point correspondence by considering geodesic consistency. **Skeleton-based** methods Shi et al. [2007], Pekelny and Gotsman [2008] build rigging model by optimizing skeleton position and vertex weight. **Template-based** method is common in registration and such template is often selected before deformation as a good initialization with principal component analysis in Kwok et al. [2014]. The deformation can be applied part by part via the piece-wise as-rigid-as possible manner Chang and Zwicker [2008] or parametric model fitting Zuffi and Black [2015]. All aforementioned methods rely on optimization process with data term similar to Eq. 2.1 and some regularization terms to achieve a global optima. Moreover, initialization is typically challenging the robustness of optimization process Bronstein et al. [2006]. In contrast, neural network Groueix et al. [2018a], Deprelle et al. [2019] deforms the template with latent codes through End-to-End learning process, and latent

codes can be optimized as the previous methods but without the risk of falling to a bad initialization.

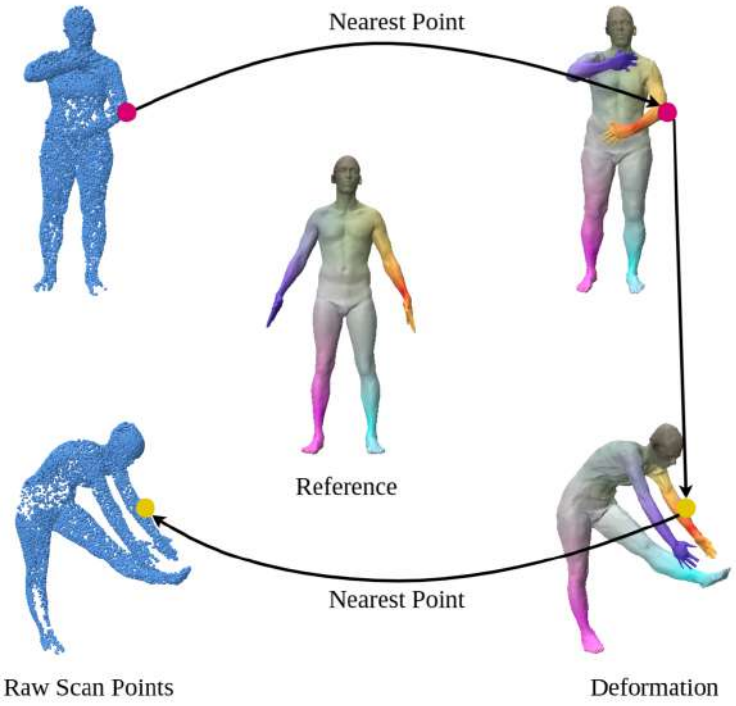


Figure 2.1: Template deformation registration. The correspondence can be built via the reference topology and shown in color. The 3D model is from FAUST Bogo et al. [2014] dataset.

After registration, the correspondence in different shapes can be established via the reference topology, see Figure 2.1. Correspondence is a common concept in computer vision for the purpose of building consistency between inputs. Besides, we could build correspondence between source and target shape by matching local pattern. Classic feature matching Khairy and Howard [2008], Tombari et al. [2010], Dey et al. [2010] methods provide sometimes the **sparse correspondence**. Geometry feature, such as FPFH Rusu et al. [2009] and SHOT Tombari et al. [2010], describes local surface based on measurement of the query point and its neighboring support. Some continuous signatures, such as HKS Sun et al. [2009] and WKS Aubry et al. [2011], follow physical principles to represent surface and open the possibility to do the shape analysis. Deformable shape can be represented with continuous function, *e.g.* Laplace-Beltrami operator Aubry et al. [2011], Andreux et al. [2014], functional map Ovsjanikov et al. [2012], in particular with deep networks Litany et al. [2017], Rodolà et al. [2017], Donati et al. [2020], Attaiki et al. [2021]

based feature extraction and supervision of pre-computed ground truth, has been well studied for **dense correspondence** however it is in general limited to the symmetry ambiguity, disconnected component and noisy data. Some ICP-like methods Newcombe et al. [2015], Dou et al. [2016], Yu et al. [2017] can also figure out dense correspondence by solving point-to-point or point-to-plane data term like Equation 2.1. Recent network methods solve the problem of shape matching by using the learning-based shape representation. 3D convolution is a common strategy to extract learned geometric features Zeng et al. [2017], Choy et al. [2019] along with contrastive loss for dense correspondence. Deng et al. [2018], Aoki et al. [2019] explore the representation capability of PointNet Qi et al. [2017] for point cloud registration. Attention mechanism Huang et al. [2021], Li and Harada [2022] is another line of research to match overlap regions. Bozic et al. [2021a] encodes the neural deformation graph Sumner et al. [2007] which represents the deformation based on key nodes, rotations and importance weights, so the query point is deformed with the weighted sum of all nodes, however this network relies on the costly 3D convolution and pre-computed signed distance in the grid.

As articulated human can be pre-defined by a reference shape or common parametrization, we focus on the template deformation in which the dense correspondence is established by deforming the reference/template to raw scan data. Even if the template deformation is suited to build the dense correspondence between reference and target shape, local constraint on body part is not carefully considered in network design. In Chapter 3, we propose to use Gaussian Process with radial basis function as the local constraint for template deformation. As benefit of registration, the imperfect raw data would be updated by the registration information, *e.g.* missing hole can be filled by the corresponding reference region, see Figure 2.1 and 2.2, and outlier signal can be filtered.

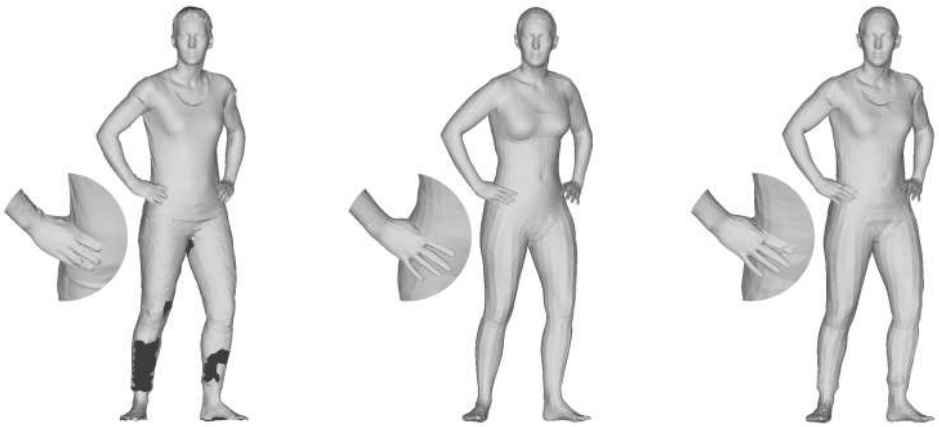
2.3 Parametric Body Model

Parametric body models facilitate downstream applications Bogo et al. [2016], Kanazawa et al. [2018], Ugrinovic et al. [2021], Sun et al. [2021a]. In particular, they solve also 3D human recovery from sparse or incomplete inputs Litany et al. [2018], Palafox et al. [2021], thanks to the compressive representation. As a result, parametric body models are used to tackle the same problem as ours from a different perspective. For this reason, we give in this section a brief

review of parametric body models. Readers can refer to Tian et al. [2022] for a more detailed survey.

Parametric body model represents human data with shape and pose parameters based on PCA technique, principal component analysis. The use of PCA can compress the shape space for a variety of human identities. Such model is built on top of large datasets such as CAESAR Robinette et al. [2002] and Hasler et al. [2009]. The famous parametric body models include SCAPE and SMPL.

SCAPE Anguelov et al. [2005] is the first parametric model to represent individual shape and pose-dependent shape. Such model is trained with the combination of pose data, a variety of pose performed by a particular person, and body shape data, different body identities in the similar pose. Some variants, such as S-SCAPE Jain et al. [2010], BlendSCAPE Hirshberg et al. [2012], Dyna Pons-Moll et al. [2015], explore and develop SCAPE by considering identity-specific pose deformation, vertex displacement and soft-tissue dynamics.



(a) Scan **(b)** SMPL **(c)** SMPL-D

Figure 2.2: Illustration of parametric model fitting. (a) Raw scan, fitting with (b) SMPL Loper et al. [2015] and (c) SMPL-D Pons-Moll et al. [2017]. The 3D model is from CAPE Ma et al. [2020].

SMPL Loper et al. [2015] is another representative parametric human model which is the most widely used in research community accompanying its variants. The shape parameters $\beta \in \mathbb{R}^m$ represent coefficients of the first m principal components, and the pose parameters, θ contain the relative rotation of child joints with respect to the parent node in kinematic tree. SMPL-X Pavlakos et al. [2019] jointly learns the body shape, hand pose and facial expression on

leveraging the hand model of MANO Romero et al. [2017] and the face model of FLAME Li et al. [2017]. SMPL-D Pons-Moll et al. [2017] models clothed person by considering cloth as displacement of each vertex. STAR Osman et al. [2020] ameliorates local blend issues from SMPL by constraining local joint impact on nearby vertices.

Other than shape modeling with statistical principal component and pose transformation with kinematic tree node, deep learning methods Jiang et al. [2020], Palafox et al. [2021] tackle the problem of decoupling the shape and pose with large-scale data. Furthermore, Lombardi et al. [2021] leverages neural implicit representation and physical kinematic model to learn a controllable body model.

Parametric or neural body model is obtained from pre-registered data. As chicken-egg problem, parametric model can also be used for the purpose of registration, *e.g.* Figure 2.2. As a non-rigid deformation task, optima of shape and pose parameters are not trivial to achieve without any warmup setup. For example, the optimization process is aided with the guide of initial markers Allen et al. [2003], Anguelov et al. [2005]. MoSh Loper et al. [2014] achieves fitting the sparse markers with BlendSCAPE Hirshberg et al. [2012]. Furthermore, MoSh++ Mahmood et al. [2019] considers the soft-tissue dynamics and improves the capture of hand by using SMPL-H model Romero et al. [2017]. Bogo et al. [2014, 2017] register scan data in static frame and in sequence by painting the texture on body. ClothCap Pons-Moll et al. [2017] handles the marker-less registration for clothed person relying on the segment of cloth with Markov random field and modeling cloth layer as the offset from naked surface of SMPL.

For registration, neural networks can be trained as a regressor for body shape and pose parameters to fit raw input. LBS-AE Li et al. [2019] fits the 3D model to the point cloud in the self-supervised manner. Moreover, IP-Net Bhatnagar et al. [2020a] leverages the neural implicit representation to predict 2-layer-surface from partial point cloud, and registers the inner part and outer part with SMPL Loper et al. [2015] and SMPL-D Alldieck et al. [2019b], Lazova et al. [2019], Pons-Moll et al. [2017]. LoopReg Bhatnagar et al. [2020b] learns the backward mapping from input scan data to the canonical model, and projects the point on canonical model back to the input scan with a diffused SMPL model in order to make the learning process differentiable and self-supervised.

Apart from registration, parametric body models contribute also to recover 3D human information, especially combining with neural network, *e.g.* the single-person Bogo et al. [2016], Kanazawa et al. [2018], Pavlakos et al. [2018], Omran et al. [2018] and the multiple-person Jiang et al. [2020a], Ugrinovic et al. [2021], Sun et al. [2021a] recovery from single color image and the video-based temporal 3D recovery Kocabas et al. [2020], Luo et al. [2020], Rajasegaran et al. [2021]. Although parametric models represent the body following the physical plausibility with the compact factors, the topology of human models has to be pre-defined which limits parametric models to go further in the reconstruction task, for example high-frequency detail recovery and cloth modeling. In Chapter 4 and 5, we examine how neural implicit representation handle shape completion and demonstrate comparison to parametric model fitting.

2.4 Shape Completion & Reconstruction from Partial Observation

3D human reconstruction from partial observation is fundamental and challenging to study for two reasons. First, we should consider all aspects to describe a human, such as shape, pose, cloth style and facial expression. Second, due to partial observation, ambiguity arise with occlusion issue. Moreover, custom-level depth camera such as Microsoft Kinect was introduced into public, 3D scanning technique has gained attention in the contexts Newcombe et al. [2011, 2015], Dai et al. [2017a,b], Litany et al. [2018], Chibane et al. [2020a]. However, classic methods fill a limited request in the setup that only small holes exist, with Laplace hole filling Nealen et al. [2006] or Poisson surface reconstruction Kazhdan et al. [2006], Kazhdan and Hoppe [2013]. Classic methods Newcombe et al. [2015], Dai et al. [2017a] can only update the canonical model from live visible domain but is still not able to provide a complete shape. Networks Chibane et al. [2020a] solve this ill-posed problem with statistical prior. Our goal in this thesis is to build dynamic human models from monocular camera with neural networks in order to simplify classic multi-view stereo reconstruction system. We thus review in this section classic methods 2.4.1, neural explicit methods 2.4.2 and neural implicit methods 2.4.3.

2.4.1 Classic Fusion

Fusion based methods date back to Izadi et al. [2011], Newcombe et al. [2011] for single-view depth reconstruction with one monocular depth sensor. The follow-up work, dynamic fusion Newcombe et al. [2015], develops the algorithm for a moving camera.

Given initial estimation of canonical-live correspondence from current canonical model to live frame, the warped surface $\hat{V}$ can be rendered into the live frame. This yields some new visible region in canonical model which updates the visible domain $u \in \Omega$. As a result, non-rigid warping and Truncated Signed Distance Field (TSDF) fusion are performed iteratively, such that the canonical model is more and more complete when the visible domain is updated with camera moving. The completion process relies on the minimization of the point-to-plane Chen and Medioni [1992], Low [2004] energy,

$$\sum_{u} \mathbb{L}((T(\pi(\hat{V}_u)) - \hat{V}_u)^\top \hat{N}_u) \tag{2.3}$$

where π is the projection operation, $\hat{N}$ is the estimation of the normal map and the transformation $T : \Omega \mapsto \mathbb{R}^3$. The Tukey loss is chosen as $\mathbb{L}$ in Newcombe et al. [2015].

DynamicFusion Newcombe et al. [2015] achieves integrating multi-frame depths into the canonical model to reconstruct surface and to track motion. In terms of human body surface reconstruction, BodyFusion Yu et al. [2017] integrates skeleton prior in order to achieve the consistency between non-rigid deformation and skeleton tracking. In the setting of fixed monocular depth camera, DoubleFusion Yu et al. [2018] leverages parametric model SMPL Loper et al. [2015] with the canonical model to improve outer surface fusion and tracking. HybridFusion Zheng et al. [2018] utilizes sparse inertial measurement units(IMU) in order to handle fast and occluded motion. RobustFusion Su et al. [2020] takes the advantage of data-driven prior, such as occupancy, pose&shape parameters and segmentation, to improve performance capture. Fusion4D Dou et al. [2016] and Motion2Fusion Dou et al. [2017] present high-speed performance capture systems leveraging learning-based 3D correspondence prediction to deal with fast motion reconstruction and topology change. KillingFusion Slavcheva et al. [2017] relies on regularization term which imposes the local rigidity by smoothening displacement vector field. SobolevFusion Slavcheva et al. [2018] defines the gradient flow in Sobolev space to alleviate over-smoothing

impact of regularization. Instead of fusing all previous frames, Function4D Yu et al. [2021] considers dynamic sliding window to fuse a initial surface and reconstructs completed surface with deep implicit network. In terms of large-scale scene reconstruction, BundleFusion Dai et al. [2017a] achieves on-the-fly surface reintegration with local-to-global hierarchical optimization. Fusion-based approaches rely on the multi-view information from sequential frames of moving camera or multiple fixed cameras. However, in the single monocular camera setup, fusion-based technique can not complete human shape due to limited visibility. Data-driven methods take the advantage of learned prior information to overcome this ill-posed problem. In Chapter 4 and 5, our proposed approaches tackle this problem with neural implicit representation.

2.4.2 Neural Explicit Modeling

In this section, we introduce neural network based human modeling. Firstly, we follow Yuan et al. [2018] to give the definition of shape completion. Let X be the input from partial observation of 3D sensor. Let Y be the dense representation of 3D shape, such as voxels, point clouds and meshes. Shape completion can be defined as learning the mapping $X \mapsto Y$, where the explicit correspondence between X and Y does not necessarily exist. Since Y is in the pre-defined topology, *i.e.* the number of point cloud, the resolution of voxel or the face connectivity of mesh is fixed before completion, this mapping is explicitly modeled by neural network.

Extending from 2D convolution to 3D convolution, the early works Han et al. [2017], Dai et al. [2017b], Yang et al. [2017], Stutz and Geiger [2018] focus on predicting occupancy or distance field in explicit voxel-grid with encoder-decoder network. Due to limited voxel resolution, Octree is adopted by Wang et al. [2017] to gradually voxelize object and increase final resolution.

Volumetric representation has huge memory cost and computational latency. Other than voxel, point cloud plays an important role in 3D shape completion because it is flexible to represent any topology with low cost. The pioneer work PCN Yuan et al. [2018] outputs dense and completed point clouds with encoder-decoder architecture and folding-based mechanism Yang et al. [2018] given partial input. TopNet Tchapmi et al. [2019] proposes the rooted tree structure to decode dense point clouds by using nodes of arbitrary structure hierarchically. Coarse-to-fine strategy is adopted in Wang et al. [2020], Liu et al. [2020b] in

order to preserve more details. Zhang et al. [2020] designs the separated feature aggregation and refinement component to uniformly assign completed points and reduce outlier.

Finally, the explicit approaches cost heavy latency to represent details in high resolution, otherwise cloth wrinkle or facial expression would be missing in the reconstruction. Such trade-off between memory cost and representation capability is not trivial to balance, especially for long-time network training phase. Due to the quantization effect of voxel and point cloud, mesh combining vertices and faces is more popular as a visualization format in computer vision and graphics. Thereby, the neural implicit representation gains more and more attention.

2.4.3 Neural Implicit Modeling

Instead of marching in explicit volumetric space, level set approaches Zhao et al. [2001], Osher and Sapiro [2002], Fedkiw et al. [2003] implicitly encode surface with the advantages of automatic resolution of surfaces and of handling various topologies. In addition, such method can be easily adapted with neural network architecture to handle complex surfaces. In general, given the query point $\mathbf{x} \in \mathbb{R}^3$ and a learnable feature z extracted from images Mescheder et al. [2019], Saito et al. [2019], Niemeyer et al. [2020], from point clouds Mescheder et al. [2019], Chibane et al. [2020a], Peng et al. [2020], Boulch et al. [2021], Atzmon and Lipman [2020] and from low-resolution voxels Mescheder et al. [2019], Chibane et al. [2020a], Peng et al. [2020], the occupancy Mescheder et al. [2019], Niemeyer et al. [2019], Genova et al. [2020], Deng et al. [2020a], level set Park et al. [2019], Chibane et al. [2020b], Gropp et al. [2020], Atzmon and Lipman [2020, 2021], Ma et al. [2022] and texture color Niemeyer et al. [2020], Yariv et al. [2020] can be modeled by neural networks in the form of $\mathcal{M} = f(\mathbf{x}, z)$. The surface $\mathcal{S}$ is represented as zero level sets predicted by neural network,

$$\mathcal{S} := \{\mathbf{x} \in \mathbb{R}^3 | f(\mathbf{x}, z) = 0\} \tag{2.4}$$

During train phase, query points $\mathbf{x}$ are sampled in 3D bounded space associating occupancy, signed distance or unsigned distance computed on the fly in the setup of supervised learning. The sampling strategy is of great importance to the reconstruction quality. Since implicit representation $\mathcal{M}$ varies fast

in the area around the iso-surface, it is more efficient to train points near the iso-surface with supervision rather than far away from the iso-surface. On the other hand, uniform sampling is beneficial to avoid ghost artifact when focusing only on the nearby surface sampling. Thereby, the combination of adaptive sampling around the surface and uniform sampling is widely used in the literatures Saito et al. [2019], Genova et al. [2019], Saito et al. [2020], Palafox et al. [2021].

The latent code z plays also an important role for the surface reconstruction. At early stage, the latent code is encoded globally with neural network Park et al. [2019], Mescheder et al. [2019], Niemeyer et al. [2019] where the concatenation of such global feature and position of $\mathbf{x}$ is fed into decoder part to predict implicit representation $\mathcal{M}$. So only the position of $\mathbf{x}$ can differ inputs to decoder in the same scene for different points. Due to the lack of local feature, the reconstruction surface is always missing high frequency details. Later, the latent code is extracted locally by using bi-linear or tri-linear interpolation on grid-like feature maps encoded by network Saito et al. [2019], Chibane et al. [2020a], Peng et al. [2020]. Genova et al. [2019, 2020] learn the local latent code automatically from the structured template. Ummenhofer and Koltun [2021] takes the advantage of multi-scale adaptive grid and Octree structure to efficiently learn the local feature with adaptive convolution kernel. Furthermore, the latent code, no matter global one Atzmon and Lipman [2020], Gropp et al. [2020] or local one Jiang et al. [2020b], can be optimized with some unsupervised point to surface loss as refinement process after long-time network training.

(a) Occupancy **(b)** Unsigned distance **(c)** Signed distance

Figure 2.3: Illustration of different supervision level. (a) Occupancy, (b) Unsigned distance field and (c) Signed distance field. In (b) and (c), the distance of the inner part is scaled with factor 5 to best illustrate it. The 3D model is from NPMs Palafox et al. [2021].

In terms of supervision level, we can provide the sign(occupancy) Figure 2.3(a) of query point, inside part or outside part, the unsigned distance Figure 2.3(b) from query point to surface, and signed distance Figure 2.3(c) during

training phase. For the reconstruction task Saito et al. [2019], Mescheder et al. [2019], Saito et al. [2020] from RGB input, occupancy is typically used as the implicit representation, while the signed distance is useful to train networks Atzmon and Lipman [2020], Gropp et al. [2020], Palafox et al. [2021] when partial 3D information is available, *e.g.* point cloud or depth. Unsigned distance Cai et al. [2020], Chibane et al. [2020b] has the benefit to model open surface.

The pioneer work of neural implicit network Mescheder et al. [2019] opens the possibility of reconstruction from image, from low-resolution voxel and from point cloud. Furthermore, Saito et al. [2019] recovers physically plausible human body from image and Saito et al. [2020] adopts coarse-to-fine strategy to extract high frequency details from patch of image. For the task of shape completion, partial data is processed into occupancy Chibane et al. [2020a] in voxel-grid and fed into 3D convolutional network hierarchically. Similar to voxel-based modeling, 3D convolutions lead a huge computational cost. NPMs Palafox et al. [2021] leverages implicit representation to learn a neural parametric model of human as well as to do joint optimization process to complete partial information from a single view, however the completion relies on the deformation from canonical model which limits prediction into the predefined topology. In Chapter 4, we propose to encode the feature map directly from 2D depth image which provides the intensive information such as shape, pose and garment detail. Moreover, we extend the spatial completion model into the temporal context which could aggregate more temporal information from the sequence to preserve details.

2.5 Temporal Modeling

The human performs motion with the temporal coherence, *e.g.* kinematic inertia, temporally consistent shape and cloth wrinkle folding. Frame by frame static modeling of motion would loose the kinematic properties. In addition, the information between the previous and next frames could contribute to the current frame. Therefore, 4D human modeling is typically of great importance in the community.

Pioneer works Akhter et al. [2012], Simon et al. [2014] consider the body or facial motion as a large matrix in which the column represents vertices in each frame. So motion completion or reconstruction is based on the matrix factorization with spatial and temporal basis. The human shape and motion variety are

limited with the dimensionality of matrix, so such modeling is only suitable to simple skeleton motions due to computational cost. Apart from rigged skeleton motion, the parametric human model, *e.g.* SMPL Loper et al. [2015], is widely used as a compressive representation of 3D shape for 4D modeling. Alldieck et al. [2018a] leverages parametric model to reconstruct the motion sequence from monocular video, Alldieck et al. [2018b, 2017] takes the advantage of silhouette or optical flow to explore the temporal coherence. Tung et al. [2017] propose to regress the shape and pose parameters under the guidance of key points and silhouette in the self supervised fashion. Kanazawa et al. [2019] learns a representation of 4D dynamics from video with the supervision of 2D pose detector Cao et al. [2019] in the semi-supervised manner, and outputs 3D poses of the current frame and nearby frames which are validated with an adversarial prior Goodfellow et al. [2014]. Parametric model-based methods discretize 4D space at fixed frame frequency and is not flexible to time-varying topologies. Template-based models Bermano et al. [2015], Zheng et al. [2017] require sometimes the semantic segmentation information in order to fit the model, and arise the problem for large deformation and temporal coherence, even in the setup of multi-view stereo Mustafa et al. [2015], Leroy et al. [2017], so such methods are restricted to the small movement or require the careful engineered work. With wide development of neural radiance field (NeRF) Mildenhall et al. [2020], Xian et al. [2021] includes the temporal term to model spatio-temporal video radiance field and D-NeRF Pumarola et al. [2021] learns the displacement from canonical space in the setup of radiance field. However, NeRF-like methods lack long-term constraint for 4D modeling in the radiance field and rely on a long-term optimization on the normalized space. OFlow Niemeyer et al. [2019] directly models the velocity field with neural network and learns the consistency between forward and backward flow under Ordinary Differentiable Equation Chen et al. [2018b] formulation. Flow-like work focuses on modeling the correspondence between canonical model and queried frame, but not on shape recovery and lacks fine details. Li et al. [2021] completes the shape of current and next frame together with the motion field based on explicit voxel representation, which is also restricted to fine details.

The 4D human modeling, especially in the continuous representation, is always missing due to the high dimensionality. In Chapter 4 and 5, we try to fill this gap by considering the input sequence all of at once and provide the representation of temporal dynamics, which is different from frame by frame

template modeling. And our methods allow us to make the movement of human physically plausible and to avoid the artefact due to the input noise of certain frame.

2.6 Coarse-to-Fine Detail Recovery

Coarse-to-fine is a common strategy in computer vision for 2D object recognition Lin et al. [2017], 3D scene reconstruction Sun et al. [2021b] and point cloud completion Wang et al. [2020]. This strategy can be naturally extended to recover fine human details, such as garment wrinkle, hair and facial expression. A low-level parametrization is built as a coarse result, on which the high frequency details are refined.

The coarse level result can be sometimes obtained with the parametric model such as SMPL Loper et al. [2015]. The fine details can be modeled as vertex displacement Pons-Moll et al. [2017], Alldieck et al. [2019a] from model based topology, some dense deformation Zhu et al. [2021] retained with subdivision of template mesh, or with implicit representation Zheng et al. [2021a] beyond 3D coarse geometry. However the pre-defined model topology restricts fine level result and sometimes additional information such as UV map is required as in Alldieck et al. [2019a] which declines flexibility to recover details. Surface normal plays also an important role to improve reconstruction quality, *i.e.* Zheng et al. [2019], Saito et al. [2020] include the normal map into the network encoding in order to extract high-frequency details or the normals estimated by Abrevaya et al. [2020] are added to the coarse result to enhance the geometry with normal mapping Cohen et al. [1998]. Similar to normal map, depth image Zeng et al. [2019], Newcombe et al. [2011] contains the fine level detail of 3D geometry and coarse-to-fine strategy has the benefit to extract such details.

As common strategy, coarse-to-fine pipeline covers almost all areas of computer vision, *e.g.* object detection Lin et al. [2017], optical flow Hu et al. [2016], Sun et al. [2018], depth estimation Yang et al. [2020], point cloud completion Wang et al. [2020], Liu et al. [2020b], 3D reconstruction Sun et al. [2021b] and so on. A pyramid of feature maps are extracted by neural network and the supervision is added in each level of pyramid. In most cases, such strategy works in the explicit manner, *i.e.* the resolution is multiplied by 2 for level up and the prediction of the previous level can play a role of initialization of the next level as in Hu et al. [2016], Lin et al. [2017], Sun et al. [2018]. For implicit neural

modeling, IF-Net Chibane et al. [2020a] collects the feature of query point in all levels of 3D feature voxel and feeds them all together to the decoder without the intermediate supervision. Saito et al. [2020] learns the global geometry from the coarse image and handle local details with high resolution image separately. So the coarse level result is not used as the initialization for the fine level, and the fine level supervision does not really refine or correct the coarse level result in the implicit context.

When extending to temporal domain, other than static resolution coarse-to-fine, temporal information can also be aggregated from different levels. In 2D context, video is represented in the different level with dilated convolution layer Farha and Gall [2019], Li et al. [2020] or grid-pooling technique Kahatapitiya and Ryoo [2021] to refine previous result. Sener et al. [2020] aggregates multi granularity temporal information with max pooling and attention as flexible representation for action prediction. In 3D context, existing works still prepare frame-wise representation and feed them to temporal model to achieve 3D reconstruction tasks. NeuralRecon Sun et al. [2021b] extracts the stack of feature maps and uses RNN to fuse the feature from different view coarse-to-fine. Transformer Vaswani et al. [2017] is also used to fuse the averaging feature from coarse level to fine result as in Bozic et al. [2021b], Stier et al. [2021], and opens the possibility to aggregate images from all available views. However, voxel representation for fusion operation suffers the trade-off between resolution and memory cost and the lack of intermediate supervision could not explore the benefit of coarse-to-fine.

In Chapter 5, we present a novel approach taking coarse-to-fine strategy and residual learning for detail recovery. We believe that coarse-to-fine strategy can balance the global information from coarse level and local cues from fine level to better represent 3D shape. Furthermore, we extend this strategy to temporal direction and examine how temporal continuity could contribute to shape completion along with this strategy.

Chapter 3

Reconstructing Human Body Mesh from Point Clouds with Adversarial GP Network

3.1 Introduction

Template-based human shape matching is a problem of broad interest in computer vision, for a variety of applications relevant to Augmented and Virtual Reality, surveillance and 3D media content production. It is relevant to various tasks such as dense shape alignment or tracking, shape estimation and completion from sparse or corrupt shape data. In this chapter, we present our adversarial GP network which takes unordered point cloud as input and outputs template-aligned mesh. During the training, we provide the registered data as supervision. The correspondence between two sets of point clouds can be established by finding the nearest neighbour between the query point from point cloud and output reconstruction, as shown in Figure 3.1.

This problem has been addressed with several classic approaches that either directly find dense correspondence using intrinsic surface embeddings Bronstein et al. [2006], Ovsjanikov et al. [2010], Kim et al. [2011] or use human body templates as geometric proxy to guide the matching Anguelov et al. [2005], Loper et al. [2015], Bogo et al. [2014, 2017], Zuffi and Black [2015]. Both approaches usually involve some form of non-convex optimization that is susceptible to ambiguities and local minima, and hand-crafted features to estimate the correspondence.

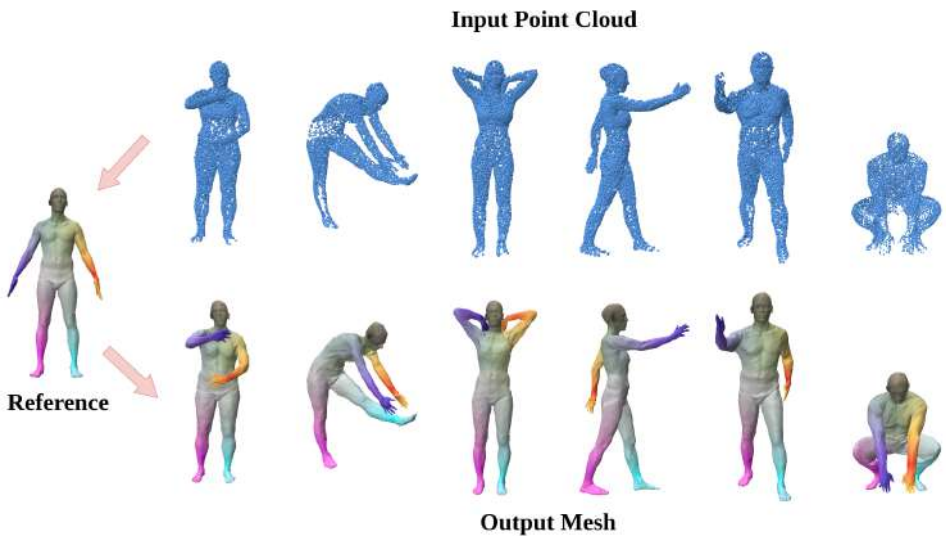


Figure 3.1: Our proposed approach takes point cloud as input and outputs the reference-aligned mesh. The correspondence can be built via the reference topology and shown in color.

Aiming for noise and initialization resilience and improvement in feature description has motivated an avenue of research in learning-based correspondence approaches to the human shape matching problem. These methods have the property to automate feature extraction and matching by mining large datasets, and can estimate correspondences by building automatic feature classifiers with e.g. random forests Rodolà et al. [2014], or simultaneously learn feature extraction and correspondence using DNNs Monti et al. [2017], Wei et al. [2016], Halimi et al. [2019], Litany et al. [2017], Tombari et al. [2010].

Many of these learning approaches rely on some form of human a priori knowledge. Most methods propose matching to an explicit shape deformation model, for which a reduced parameterization is predicted Tan et al. [2018], Li et al. [2019], Jiang et al. [2020] or whose mapping to the data is learned Marin et al. [2020], Prokudin et al. [2019] from observations.

Among the most successful approaches of inspiration to this work are those matching human shapes using an implicit deformation model which is entirely learned with no manually set components, as applied to humans Groueix et al. [2018a] or generic objects Groueix et al. [2018b]. By encoding matching to an underlying template as the expression of a learned global feature in a latent space automatically discovered by an auto-encoder, the model can be entirely automated and trained end-to-end for generic matching of two shapes, as opposed

to the previously described methods. As they use point-based DNN architectures Qi et al. [2017], these approaches can be applied to point cloud inputs without any surface consistency. All these properties afford greater robustness and generalization abilities, and allow this family of methods to outperform the latter on standard benchmarks. However in this process a weaker human shape and consistent surface prior is encoded than previous approaches, which leads to noisy, and sometimes non-realistic predicted human shapes, as confirmed by an analysis of the failure cases of these approaches. Some of the failures are mitigated using a post-processing step which consists in optimizing the shape matching features inferred by the DNN in the latent space, which improves the final result.

In this chapter, we explore within this family of approaches how local and global shape priors, commonly not encoded with point-wise architecture, can be reintroduced while maintaining the benefits of such an architecture (e.g. PointNet Qi et al. [2017]). We base our approach on a point-based auto-encoder similar to Groueix et al. [2018a], but with several key differences. To alleviate inference noise, we introduce a Gaussian Process decoder layer which inherently encodes surface smoothness and surface point coherence on the shape with lower point dimensionality on the surface, only to the price of a small pretraining phase. Second, more global consistency is built in the model by adding fully connected layers at the end of the decoder, which is made possible by the surface dimensionality reduction previously discussed. Third, to avoid inferring drastically non human shapes, we introduce an adversarial training phase inspired by Hu et al. [2018] which enforces consistency of human shape encodings in our latent space and helps to avoid overfitting. With these improved network characteristics and training procedures, we show that our approach provides results that are on par or better than state-of-the-art on the FAUST intra and inter challenges and illustrate the quality gain of our approach through an exhaustive ablation study illustrating the benefits of these three contributions.

3.2 Related Work

There exists a rich literature on registration and reconstruction of 3D data, many of which were reviewed in Section 2.2. Here, we focus our analysis on methods for registering human body shapes, following the classic distinction between

template-free and template-based methods. We then briefly discuss how Gaussian Processes have been combined with DNNs in previous work, and the use of adversarial training in the context of 3D vision.

Template-free methods Correspondences between non-rigid objects can be established by defining an intrinsic surface representation, which is invariant to bending. In the embedding space defined by this representation, the registration problem boils down to a non-convex optimization one. Examples of intrinsic representations are Generalized Multi-Dimensional Scaling (GMDS) Bronstein et al. [2006], heat kernel maps Ovsjanikov et al. [2010], Möbius transformations Kim et al. [2011]. Recent work tries to learn such representations, and therefore object-to-object correspondences, from data. While early approaches rely on random forests Rodolà et al. [2014], subsequent ones employ DNNs Monti et al. [2017], Wei et al. [2016]. For example, Deep Functional Maps Halimi et al. [2019], Litany et al. [2017] combine a deep architecture with point-wise descriptors Tombari et al. [2010] to obtain dense correspondences between pairs of shapes. These methods aim at matching arbitrary shapes. However, when focusing on particular instances like the human body, one can introduce more powerful class-specific shape priors.

Template-based methods When registering noisy and incomplete 3D human body data, one commonly relies on a predefined 3D body template acting as a strong shape prior. At registration time, the template surface is deformed in order to match the data. Many approaches rely on a statistical body model Anguelov et al. [2005], Loper et al. [2015] and define an objective function which is minimized via non-linear least squares Bogo et al. [2014, 2017], Zuffi and Black [2015]. However, these objective functions use hand-crafted error terms and are not as powerful as data-driven approaches. Recently, the wider availability of huge datasets of 3D body shapes Bogo et al. [2017], Varol et al. [2017] fostered the development of DNN-based methods. Mesh Variational Autoencoders Tan et al. [2018] learn a latent space for 3D human body representation, but their input is limited to fixed-topology shapes. LBS-AE Li et al. [2019] proposes a self-supervised approach for fitting 3D models to point cloud. The method relies on DNNs to learn a set of Linear Blending Skinning Magnenat-Thalmann et al. [1988] parameters. FARM Marin et al. [2020] establishes correspondences between shapes by automatically extracting a set of landmarks and then using

functional maps. Deep Hierarchical Networks Jiang et al. [2020] learn a 3D human body embedding which can then be fitted to data, leveraging a set of manually selected landmarks. Basis Point Sets Prokudin et al. [2019] propose an efficient point cloud encoding, which can then be combined with DNNs Huang et al. [2017] for shape registration and completion tasks.

GP and DNNs Gaussian Processes (GP) are popular in statistical learning for their generalization capabilities. In 3D vision, Lüthi et al. [2017] propose GPMMs, a morphable model based on GP, with applications to face modeling and medical image analysis. Recently, some studies Wilson et al. [2011],Wilson et al. [2016] try to interpret how DNNs can simulate the learning process of GP. For example, Deep GP Damianou and Lawrence [2013] focuses on probabilistic modeling of GP with DNNs, training the network via marginal likelihood. In this work, we leverage the interpolation and smoothness capabilities of GP in the context of 3D surface reconstruction.

Adversarial training After the introduction of Generative Adversarial Networks (GANs) Goodfellow et al. [2014], adversarial training has been widely used in computer vision. In 3D vision, HMR Kanazawa et al. [2018] applies adversarial learning to estimate 3D human body shape and pose from 2D images. CAPE Ma et al. [2020] uses it to learn a model of people in clothing. Fernández Abrevaya et al. [2019] and Shamai et al. [2019] use adversarial training to model faces in 3D. Hu et al. [2018] compare adversarial and L2-norm regularization for the task of image registration. To the best of our knowledge, our work is the first to propose adversarial training as a regularization term in the context of 3D registration.

In general, our work builds on 3D-CODED Groueix et al. [2018a], which uses a PointNet-like Qi et al. [2017] architecture to extract permutation-invariant point features. However it applies the point-wise decoders which are independent of each other. Thus we propose to strengthen the relationship of nearby points by using our GP layer and MLP layers. AtlasNet2 Deprelle et al. [2019] aims at improving upon 3D-CODED reconstructions by using a learnable template. However AtlasNet2 results exhibit artifacts similar to the ones of 3D-CODED in some challenging cases. In order to make the network predictions more robust, we propose to use adversarial training.

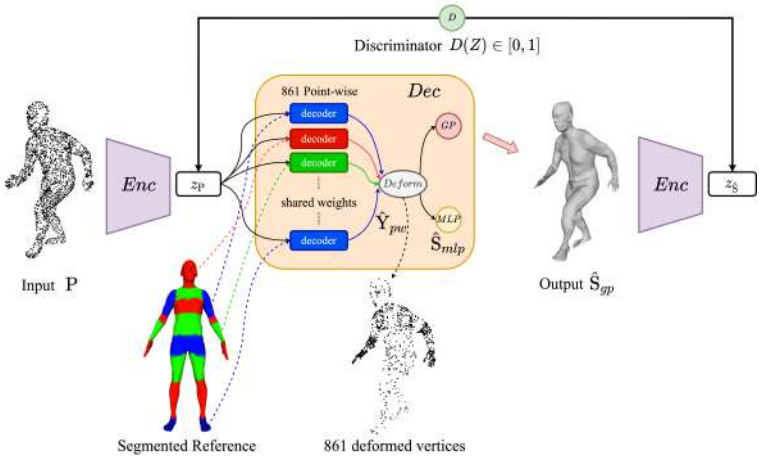


Figure 3.2: Overview of our Adversarial GP Network.

3.3 Method

Our approach takes as input an unordered set of n 3D points $\mathrm{P} \in \mathbb{R}^{n\times 3}$ and maps this set into a deformed instance $\mathrm{S} \in \mathbb{R}^{res\times 3}$ of a reference mesh with a fixed resolution res. The number n of input points can vary. This to allow for partial or incomplete shape description as typical with laser scan or depth data. In order to learn such a mapping we use a point-wise encoder-decoder architecture trained on standard human body datasets. This architecture presents two innovations to better enforce shape consistency: first a regularization layer that builds on Gaussian Process (GP) and second a global adversarial loss. The sections below detail the different components of our framework.

3.3.1 Network Architecture

As shown in Figure 3.2, our architecture encodes points P into a latent shape representation $\mathcal{Z}_p \in \mathbb{R}^{1024}$ which is then decoded into a deformation vector field Y defined over a mesh template to produce the shape S. Our objective is to balance global and local information with shape-wise and point-wise considerations. To this end, a PointNet Qi et al. [2017] like encoder is used as a backbone to extract the 1024-dimension latent shape feature $\mathcal{Z}_p$. On the decoder side, we first expect this global shape feature $\mathcal{Z}_p$ to predict the deformation of a subset of representative points on the reference mesh in a point-wise manner. More global considerations are then applied on this subset of points with both Gaussian Process interpolation and fully connected layers. Furthermore, to better constrain

the latent representation during training, the output vertices of the predicted deformed reference mesh S are fed into the encoder to verify whether they yield a latent feature $\mathcal{Z}_s$ close to the latent feature $\mathcal{Z}_p$ of the ground truth shape vertices.

Encoder

We extract the global feature $\mathcal{Z}_p$ with a simplified version of PointNet Qi et al. [2017]. The input points P are first processed by 3 hidden layers of size 64, 128 and 1024, respectively, followed by a max-pooling operator applied to the resulting point-wise features. Then, two linear layers of size 1024 lead to the latent space $\mathcal{Z}_\mathcal{P}$. All layers use batch normalization and *ReLU* (rectified linear unit) activation.

Decoder

The decoder takes as input the shape feature $\mathcal{Z}_p$ extracted by the encoder together with l 3D locations x_i of vertices distributed on the reference mesh. Point-wise decoders with shared weights are first used on the combinations $(\mathrm{x}_i, \mathcal{Z}_p)$. These decoders are composed of 3 hidden layers going from size 1027 to 513 and 256. The resulting features are projected into l individual vertex deformations y_i using 2 times hyperbolic tangent activation functions. Following point-wise decoders, two computation flows are applied in parallel on the resulting predicted vertex deformations y_i. One goes to GP layers that enforce local spatial consistency between vertices and the other goes to a fully-connected MLP layer that enforces a global constraint over vertices. We take the output of the GP flow as the final deformed instance.

3.3.2 Local and Global Spatial Consistency

Gaussian Process Interpolation

As mentioned before the decoder part includes a vertex interpolation technique based on Gaussian Process Williams and Rasmussen [2006]. To this aim, we assume here that deformations y_i of the reference mesh at vertex locations x_i are, up to a bias $\varepsilon \sim \mathcal{N}(0, \sigma)$, non linear functions $\mathrm{y}_i = f(\mathrm{x}_i) + \varepsilon$, which distributions are jointly Gaussian, with mean and covariance defined by the kernel

k:

$$k(\mathrm{x}_i, \mathrm{x}_j) = \gamma \exp(-\frac{\|\mathrm{x}_i - \mathrm{x}_j\|^2}{r}). \tag{3.1}$$

Under these assumptions, the joint distribution of l partial vertex observations Y and an unobserved vertex y_* over the deformed reference mesh can be expressed as:

$$\begin{bmatrix} \mathrm{Y} \\ \mathrm{y}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathrm{K}(\mathrm{X},\mathrm{X}) + \sigma^2 I, & \mathrm{K}_*^T(\mathrm{x}_*,\mathrm{X}) \\ \mathrm{K}_*(\mathrm{x}_*,\mathrm{X}), & k(\mathrm{x}_*,\mathrm{x}_*), \end{bmatrix}\right) \tag{3.2}$$

where $\mathrm{K}(,)$ denotes the covariances over the associated vertices x_i on the reference mesh:

$$\mathrm{K}(\mathrm{X},\mathrm{X}) = \begin{bmatrix} k(\mathrm{x}_1,\mathrm{x}_1) & \dots & k(\mathrm{x}_1,\mathrm{x}_l) \\ \vdots & \ddots & \vdots \\ k(\mathrm{x}_l,\mathrm{x}_1) & \dots & k(\mathrm{x}_l,\mathrm{x}_l) \end{bmatrix}, \; \mathrm{K}_*(\mathrm{x}_*,\mathrm{X}) = \begin{bmatrix} k(\mathrm{x}_*,\mathrm{x}_1) & \dots & k(\mathrm{x}_*,\mathrm{x}_l) \end{bmatrix}. \tag{3.3}$$

The posterior probability $P(\mathrm{y}_*|\mathrm{Y})$ can be inferred as a Gaussian distribution $\mathcal{N}(m(\mathrm{y}_*), var(\mathrm{y}_*))$ with:

$$m(\mathrm{y}_*) = K_*(K + \sigma^2 I)^{-1}\mathrm{Y} \tag{3.4}$$

$$var(\mathrm{y}_*) = k_* - K_*(K + \sigma^2 I)^{-1}K_*^T \tag{3.5}$$

where, to simplify our notation, $K_* = \mathrm{K}(\mathrm{x}_*,\mathrm{X})$, $K = \mathrm{K}(\mathrm{X},\mathrm{X})$ and $k_* = k(\mathrm{x}_*,\mathrm{x}_*)$. Taking the mean of this distribution as the predicted value we finally get:

$$\mathrm{y}_* = K_*(K + \sigma^2 I)^{-1}\mathrm{Y}. \tag{3.6}$$

In practice, to accelerate the GP computation and improve the reconstruction precision, we apply the above statistical reasoning individually over body parts instead of the full body. We follow for that purpose Basset et al. [2019] and segment the body topology into 19 small patches, including two patches on the elbow (see Figure 3.2). In addition, we do not consider absolute vertex locations as y_i but relative displacements with respect to the reference mesh instead. Note that we have finally 3 parameters: γ, r and σ for each body part. Thus we can use cross validation, more particularly in our case, *kernel selection*, to tune the GP parameters before the time-consuming gradient descent optimization during the neural network training.

In addition, the selected subset of l observation vertices impacts the final reconstruction of the full mesh. In order to select the most informative vertices

for that purpose, we pre-tune 19 kernels and select the observation vertices using 10 random meshes from the FAUST training dataset. We start at 10% resolution of the template, i.e. 689 vertices, and progressively add vertices to minimize the reconstruction error, finding an optimal value of 861 vertices.

Our network predicts therefore the deformations of this subset of $l = 861$ vertices, which are then completed by our GP layer. The GP layer consists of 19 body part components that exploit Equation (3.6) with pre-computed kernel matrices, see tuned Gaussian kernel parameter in Tab. 3.1. As explained before, the vertex of template is deformed by a point-wise decoder. While this is similar in spirit to Groueix et al. [2018a] and Deprelle et al. [2019], our approach differs in 2 aspects: (i) Instead of considering random points over the mesh surface during training, our approach focuses on a fixed subset of points – this allows us to better exploit the local spatial consistency of the reference mesh deformations; (ii) Instead of directly predicting the deformed template vertices, our point-wise decoder predicts the deformations (residuals) with respect to the template. The rationale here is that the residual space is generally easier to learn than the original coordinate space. In Figure 1, we show the segmented reference mesh and the 861 selected vertices deformed by the prediction of the point-wise decoder. Since the prediction is in the same order as the reference mesh, we can directly map the body part segmentation on the prediction of point-wise decoder.

Fully Connected Layer

The previous GP layer enforces local spatial constraints between mesh vertices by assuming joint Gaussian distributions that can be pre-learned from a few meshes. In order to complete this with more global considerations over the vertices of a shape, we also employ a fully-connected multi-layer perceptron as another interpolation flow. This MLP takes as input the $l = 861$ deformed vertices as predicted by the point-wise decoder. It is composed of a hidden layer of dimension 2048, followed by 2 times hyperbolic tangent activation functions, and one linear layer to interpolate to the resolution of the reference mesh, in practice 6890 vertices with the SURREAL synthetic data.

Patch	γ	r ($\times 10^{-2}$)	σ ($\times 10^{-1}$)
left forearm	7.81	1.47	1.99
left hand	4.31	1.65	2.20
left arm	0.20	4.00	2.15
left elbow	6.31	1.39	1.99
left foot	1.98	3.88	1.99
left leg	3.15	2.64	1.99
left thigh	1.98	2.20	2.20
right forearm	5.04	1.89	1.99
right hand	7.81	1.47	1.99
right arm	0.54	3.04	2.20
right elbow	6.98	1.39	1.74
right foot	5.04	1.14	1.95
right leg	1.98	6.48	1.99
right thigh	2.15	3.86	1.99
belly	12.81	1.33	1.74
crotch	1.15	2.44	2.20
head	3.04	1.68	2.20
torso	5.31	1.80	1.95
upper torso	23.15	1.35	5.49

Table 3.1: Tuned Gaussian Kernels.

3.3.3 Training Loss

In order to train our network we define a loss function $\mathcal{L}_r$ that accounts for the 3 outputs yielded by the decoder. The point-wise decoder computes the deformation field y_i over the subset of l mesh vertices on the reference mesh, while the GP and MLP layers output the deformed instances in the same resolution as the reference mesh. Hence:

$$\mathcal{L}_r(\hat{\mathrm{Y}}_{pw}, \mathrm{Y}_l, \hat{\mathrm{S}}_{gp}, \hat{\mathrm{S}}_{mlp}, \mathrm{S}) = L(\hat{\mathrm{Y}}_{pw}, \mathrm{Y}_l) + L(\hat{\mathrm{S}}_{gp}, \mathrm{S}) + L(\hat{\mathrm{S}}_{mlp}, \mathrm{S}) \quad (3.7)$$

where $L(\cdot,\cdot)$ denotes the standard mean-square error, $\hat{\mathrm{Y}}_{pw}$, $\hat{\mathrm{S}}_{gp}$, $\hat{\mathrm{S}}_{mlp}$ are the point-wise decoder, GP and MLP layer predictions respectively, Y_l is the ground truth deformation field over the reference mesh reduced to the l vertices predicted by the point-wise decoder and S is the ground truth deformed instance. In practice, we remark that the mesh obtained with the MLP layer is often blurry. However, the associated global constraint in the reconstruction loss appears to be beneficial in our experiments.

Adversarial Loss

In addition to the loss presented in the previous section, we investigate in this work the contribution of introducing an adversarial strategy Goodfellow et al. [2014] in the proposed framework. While the previous loss function enforces local and more global spatial consistency, it does not encode knowledge on what a regular shape should be. Hence artifacts can occur when considering data outside the training set, as in Figure 3.5 with test data. In order to better detect abnormal outputs, we therefore propose an additional adversarial loss.

Recall that, given an arbitrary input point cloud P, the encoder generates a latent feature $\mathcal{Z}_p$. From this latent feature, the decoder generates a deformed version, $\hat{\mathrm{S}}$, of the reference mesh. In principle, feeding the encoder with this set $\hat{\mathrm{S}}$ should yield a latent feature $\mathcal{Z}_{\hat{\mathrm{s}}(p)}$ statistically similar to $\mathcal{Z}_p$. We therefore express the adversarial loss as:

$$\mathcal{L}_a(\mathrm{P}, \hat{\mathrm{S}}) = \mathbb{E}_p[log(D(\mathcal{Z}_p)] + \mathbb{E}_{\hat{\mathrm{s}}(p)}[log(1 - D(\mathcal{Z}_{\hat{\mathrm{s}}})] \tag{3.8}$$

where $D(\cdot)$ is the discriminator trained to detect abnormal latent features. It projects the 1024-dimension point feature into 512 and then 256 dimensions with two hidden layers, and outputs a probability. The two hidden layers are activated by an *ELU* (Exponential Linear Unit) function followed by batch normalization; the output is activated by a *sigmoid* non linearity. The final loss for our network training is a combination of $\mathcal{L}_r$ and $\mathcal{L}_a$:

$$\mathcal{L}_t = \lambda_1 \mathcal{L}_r + \lambda_2 \mathcal{L}_a. \tag{3.9}$$

The training algorithm proceeds by iteratively updating the encoder-decoder and the discriminator as depicted below. The protocol followed in practice is detailed

in Section 3.4.

Algorithm 1: Training Algorithm

Input: Ground truth deformed instances S of the reference mesh

Initialization;

for *Training iterations* **do**

1. Sample a mini-batch of point cloud $\mathrm{P} \in \mathrm{S}$;

2. Compute the reconstruction $\hat{\mathrm{S}}(\mathrm{P})$;

3. Update $D(\cdot)$ by taking a learning step on loss $\mathcal{L}_a(\mathrm{P} \sim real, \hat{\mathrm{S}} \sim fake)$ (3.8);

4. Update then encoder and decoder by taking a learning step on loss $\mathcal{L}_t(\hat{\mathrm{S}} \sim real)$ (3.9);

end

3.4 Experimental Results

In this section, we first describe the datasets and the corresponding evaluation protocols. We then compare our approach against the state-of-the-art methods and provide a detailed analysis of our framework.

3.4.1 Datasets

We evaluate our framework for reconstructing human body meshes from point cloud data on the standard SURREAL Varol et al. [2017] and FAUST Bogo et al. [2014] datasets.

The SURREAL dataset is a large-scale synthetic dataset that consists of textured human body shapes in different 3D poses. We follow the protocol introduced in Groueix et al. [2018a] to generate our training data that consists of $230,000$ meshes.

The FAUST dataset provides 100 training and 200 testing human body scans of approximately 170000 vertices. They may include noise and holes, typically missing parts on the feet. The FAUST benchmark defines two challenges: the one on intra-, the other on inter-subject correspondences. We use the FAUST dataset only for testing purposes and do not use the provided scans for training.

Method	SURREAL-Chamfer	FAUST-inter
3D-CODED Groueix et al. [2018a]	1.33	2.88
AtlasNet2-Deformation Deprelle et al. [2019] 3D	1.17	2.76
AtlasNet2-Points 3D	1.11	3.05
AtlasNet2-Deformation 10D	1.01	2.77
AtlasNet2-Points 10D	1.01	2.85
Ours(MLP)	0.54	2.94
Ours(GP)	**0.35**	**2.73**
Ours(Adversarial GP)	0.50	2.76

Table 3.2: Results on the SURREAL validation set for human body reconstruction and on the FAUST-inter correspondence challenge. As in Deprelle et al. [2019], Groueix et al. [2018a], we report the symmetric Chamfer distance ($\times 10^{-3}$) for SURREAL validation. For FAUST, we report the Euclidean correspondence error in (cm). In FAUST, we apply the same refinement technique as in 3D-CODED to our MLP, GP and Adversarial GP.

3.4.2 Evaluation Protocol

We use the symmetric Chamfer distance between the predicted and ground-truth human shape to evaluate our framework on the SURREAL validation dataset. For our experiments on the FAUST dataset, we use the official test server to measure our accuracy. Throughout our experiments, we use the same training/test splits as 3D-CODED Groueix et al. [2018a]. We perform a line-search to find the initial orientation and the initial translation that gives the smallest Chamfer distance during testing FAUST.

3.4.3 Implementation Details

We implement our Adversarial GP network in PyTorch and train for 25 epochs from scratch. In practice, we set $\lambda_1 = 10$ and $\lambda_2 = 0.05$. We use the Adam optimizer with a learning rate of 0.001 for the Discriminator and 0.0005 for Encoder and Decoder. We set the batch size to 32. We follow 3D-CODED Groueix et al. [2018a] to add random translation between -3 cm and 3 cm to increase the robustness during training.

3.4.4 Comparison with Baselines

We report reconstruction and registration accuracy on the SURREAL Varol et al. [2017] and FAUST Bogo et al. [2014] datasets and compare our results to the state-of-the-art results of Groueix et al. [2018a] and Deprelle et al. [2019] in Table 3.2.

We further use the following baselines and versions of our approach in the evaluation:

- *MLP*: A multi-layer perceptron with 2 layers as described in Section 3.3.2 operating on the output deformations of the point-wise decoder.
- *GP*: Gaussian Process layer as described in Section 3.3.2 operating on the output deformations of the point-wise decoder.
- *Adversarial GP*: Adversarial network coupled with the Gaussian process and MLP layers that operates on the output deformations of the point-wise decoder (see Section 3.3.3).

We further compare our results to Deprelle et al. [2019], Groueix et al. [2018a] qualitatively to demonstrate the effectiveness of our method in Figure 3.4 and Figure 3.5.

Reconstruction.

We report our surface reconstruction results in comparison to Groueix et al. [2018a], Deprelle et al. [2019] on the SURREAL and FAUST datasets in Table 3.2. While providing accurate reconstructions, Groueix et al. [2018a] relies on point-to-point distance minimization, therefore lacking global context. To remedy this and encode global context, we apply an MLP on point-wise predictions. This would help encode global context, but in return, would ignore local dependencies. Our GP layer, on the other hand, aims at finding a balance between local and global context. As can be seen in Table 3.2, the GP layer yields the most accurate reconstruction results on the SURREAL validation set and in the FAUST Inter-Subject challenge.

Registration.

Our output mesh is reconstructed from an input point cloud and is aligned with a template shape. Therefore, our method could further compute registration to the human body by finding the closest point on the reconstruction. We evaluate our method on the FAUST Bogo et al. [2014] challenge, that includes 100-pairs of shapes to be matched. In FAUST, the input is real scan data in different orientations and translations and scans typically include noise and holes. In Table 3.3, we report the results of all published studies to date on the FAUST challenge.

Method	Intra (cm)	Inter (cm)
3D-CODED Groueix et al. [2018a]	1.985	2.878
Stitched puppets Zuffi and Black [2015]	1.568	3.126
LBS-AE Li et al. [2019]	2.161	4.08
FARM Marin et al. [2020]	2.810	4.123
BPS Prokudin et al. [2019]	2.327	4.529
FMNet Litany et al. [2017]	2.436	4.826
Convex-Opt Chen and Koltun [2015]	4.860	8.304
Our GP	2.349	2.734
Our Adversarial GP	1.904	2.759

Table 3.3: Results for the FAUST intra- and inter-subject challenges for human body registration.

Method	SURREAL-Chamfer	FAUST-intra	FAUST-inter
Adv+GP (w.o. MLP)	0.52	2.585	2.913
MLP+GP (w.o. Adv)	**0.37**	2.042	2.858
MLP+GP+L2 weight decay	5.40	6.068	7.58
MLP+GP+Dropout	0.38	2.236	2.984
Adv+MLP+GP (Adv GP)	0.50	**1.904**	**2.759**

Table 3.4: Numeric comparisons. We report the symmetric Chamfer distance ($\times 10^{-3}$) on the SURREAL validation dataset and Euclidean correspondence error (cm) in FAUST -intra/-inter challenges for the variants of our model. We further compare adversarial training to L2 weight decay (regularization term $\lambda = 5 \times 10^{-4}$) and dropout.

We do not include the results of DHNN as it requires manual selection of additional landmark points which is used to guide the optimization.

Importance of Adversarial Training.

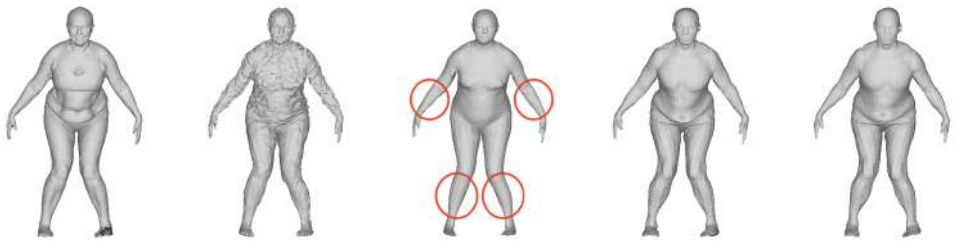
Although our GP network provides accurate reconstruction and registration results, we have observed in practice that it sometimes results in artifacts, as can be seen in a few cases in Figure 3.5. Our adversarial GP, on the other hand, is able to correct these artifacts and results in physically plausible human shape reconstructions, as demonstrated in Figure 3.5. This is in part due to the fact that adversarial training prevents overfitting to the SURREAL training data and achieves good generalization across datasets. In Figure 3.6, we show that using the MLP network along with the GP layer further regularizes the training of our Adversarial GP framework. Therefore, in practice, we also employ an MLP during training of our Adversarial GP.

In Table 3.4, we further analyze the influence of adversarial loss on the reconstruction accuracy. Using an adversarial loss yields more accurate results

Method	3D-CODED	AtlasNet2	GP	Adv GP
without refinement	6.29	4.72	**4.71**	4.964
with refinement	3.048	-	**2.734**	2.873
with refinement+ high-res template	2.878	2.76	2.815	**2.759**

Table 3.5: Comparison to 3D-CODED Groueix et al. [2018a] and AtlasNet2 Deprelle et al. [2019] with and without refinement. We report Euclidean correspondence errors on the FAUST-inter challenge in (*cm*). The refinement is based on optimizing the global feature to minimize the symmetric Chamfer distance. We follow Groueix et al. [2018a] to register the scan to a high-resolution template.

on the FAUST dataset. While resulting in lower accuracy on the SURREAL dataset, adversarial training helps to prevent overfitting by ensuring that the distributions of the input data and reconstruction are similar. In Figure 3.5, we demonstrate that adversarial training in practice results in physically more plausible and realistic shapes. To demonstrate the effectiveness of adversarial training as a regularization mechanism, we further compare it to standard regularization techniques of L2-weight decay and dropout in Table 3.4. To confirm the choice of our Adversarial training strategy, we compare to the qualitative results of different training strategies in Figure 3.6.



(a) input scan **(b)** MLP **(c)** MLP+Laplace **(d)** GP **(e)** Adv GP

Figure 3.3: Smoothness of GP. From left to right, (a) input scan, reconstruction in standard resolution of (b) MLP, (c) MLP smoothened by the Laplacian operator, (d) GP, and (e) Adversarial GP.

Influence of Gaussian Kernel Regularization.

In Figure 3.3, we present qualitative reconstruction results obtained with different decoders to further support our quantitative analysis in Table 3.2. While the MLP decoder results in a blurry shape, Laplacian denoising results in a shrinkage in the volume, especially in the limbs. GP and Adversarial GP, on the other hand, provide high-fidelity reconstructions.

Refinement.

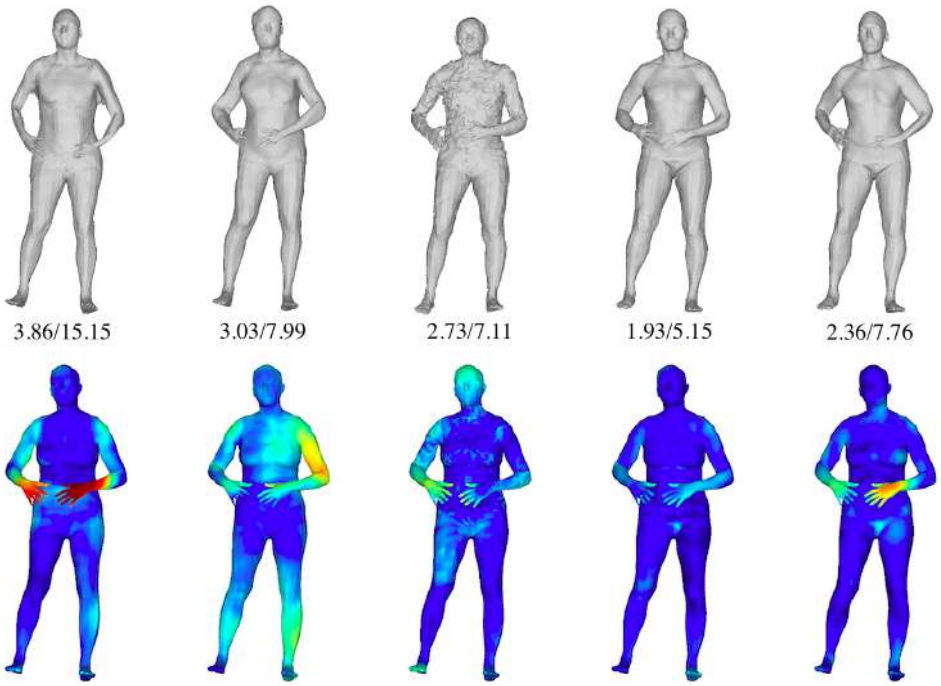
During evaluation, we follow the same refinement strategy of 3D-CODED Groueix et al. [2018a], that minimizes the Chamfer distance between reconstructions and inputs. Consequently, a nearest neighbor search is performed to find correspondences and match shapes. To highlight the benefit of refinement, we show in Table 3.5 our results in comparison to Groueix et al. [2018a] and Deprelle et al. [2019] with and without refinement. Refinement results in better accuracy for our method, as expected, and our approach provides better results in comparison to Groueix et al. [2018a] and Deprelle et al. [2019] in all cases. When we use a high resolution template for the nearest neighbor step, we gain an additional accuracy improvement for Adversarial GP, but not for GP. The result could not be always improved by using a high resolution template due to the fact that the FAUST-inter challenge computes the Euclidean distance between the prediction and sparse landmarks. Since the Euclidean distance is more tolerant of the artifacts in Figure 3.5 than geodesic distance, Adversarial GP can not make great improvement in FAUST challenge.

Qualitative Results.

In Figure 3.4, we show qualitative results on SURREAL validation of the variants of our approach, such as MLP, GP and Adversarial GP, in comparison to 3D-CODED Groueix et al. [2018a] and AtlasNet2 Deprelle et al. [2019]. Our method yields better reconstruction accuracy than Groueix et al. [2018a] and Deprelle et al. [2019] and provides realistic surface reconstructions by jointly accounting for local and global context. In Figure 3.5 and 3.6, we show the qualitative comparisons on FAUST test data. Since the output meshes share the same topology as the reference, the correspondence can be built via the reference topology as shown in Figure 3.1.

3.5 Conclusion

In this chapter, we have presented an encoder-decoder neural network architecture to reconstruct human body meshes from point cloud data, by learning dense human body correspondences. Deep Neural Networks (DNNs) achieve state-of-the-art results with generic point-wise architectures; but in doing so,



(a) 3D-CODED **(b)** AtlasNet2 **(c)** MLP **(d)** GP **(e)** Adv GP

Figure 3.4: Qualitative evaluation for human shape reconstruction. From left to right, reconstruction in standard resolution of (a) 3D-CODED Groueix et al. [2018a], (b) AtlasNet2 Deprelle et al. [2019], (c) Our MLP, (d) our GP and (e) our Adversarial GP. And we report the heatmaps and mean/max Euclidean reconstruction error in (*cm*) with respect to the ground truth.

they exploit much weaker human body shape and surface priors with respect to methods that explicitly model the body surface with 3D templates. We investigated the impact of adding back such stronger shape priors by proposing a novel dedicated human template matching process. Our architecture enforces shape coherence and surface smoothness with a specialized Gaussian process layer. Moreover our adversarial training framework allows for generalization across datasets and reconstructs high-fidelity human meshes. The choice of these elements was grounded on an extensive analysis of DNNs failure modes in widely used datasets like SURREAL and FAUST. We validated and evaluated the impact of our novel components on these datasets, showing a quantitative improvement over state-of-the-art DNN-based methods, and qualitatively better results. This work opens the possibility to add the spatial constraint during training a neural network and densify the points from a sparse representation simultaneously.

However, we observe that the proposed method relies heavily on the global

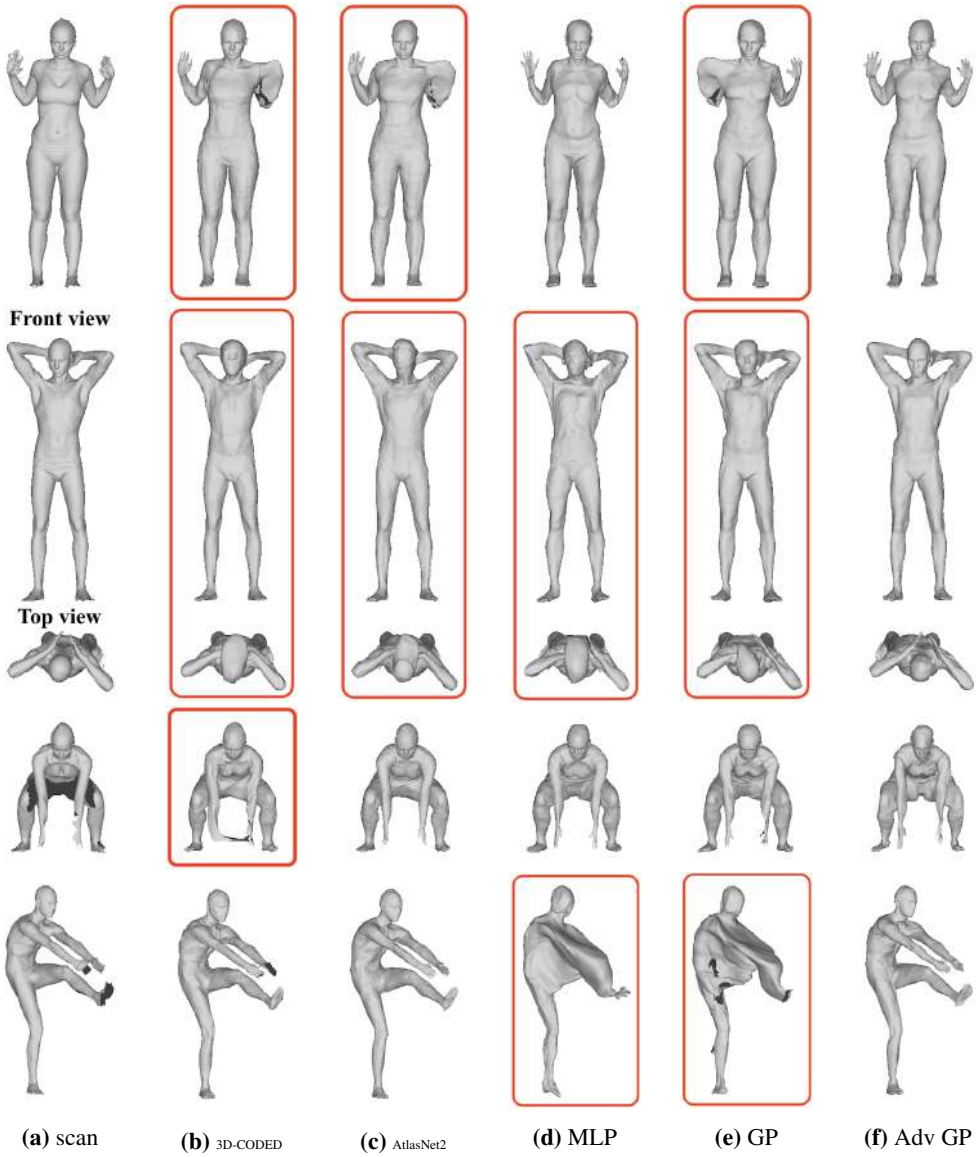
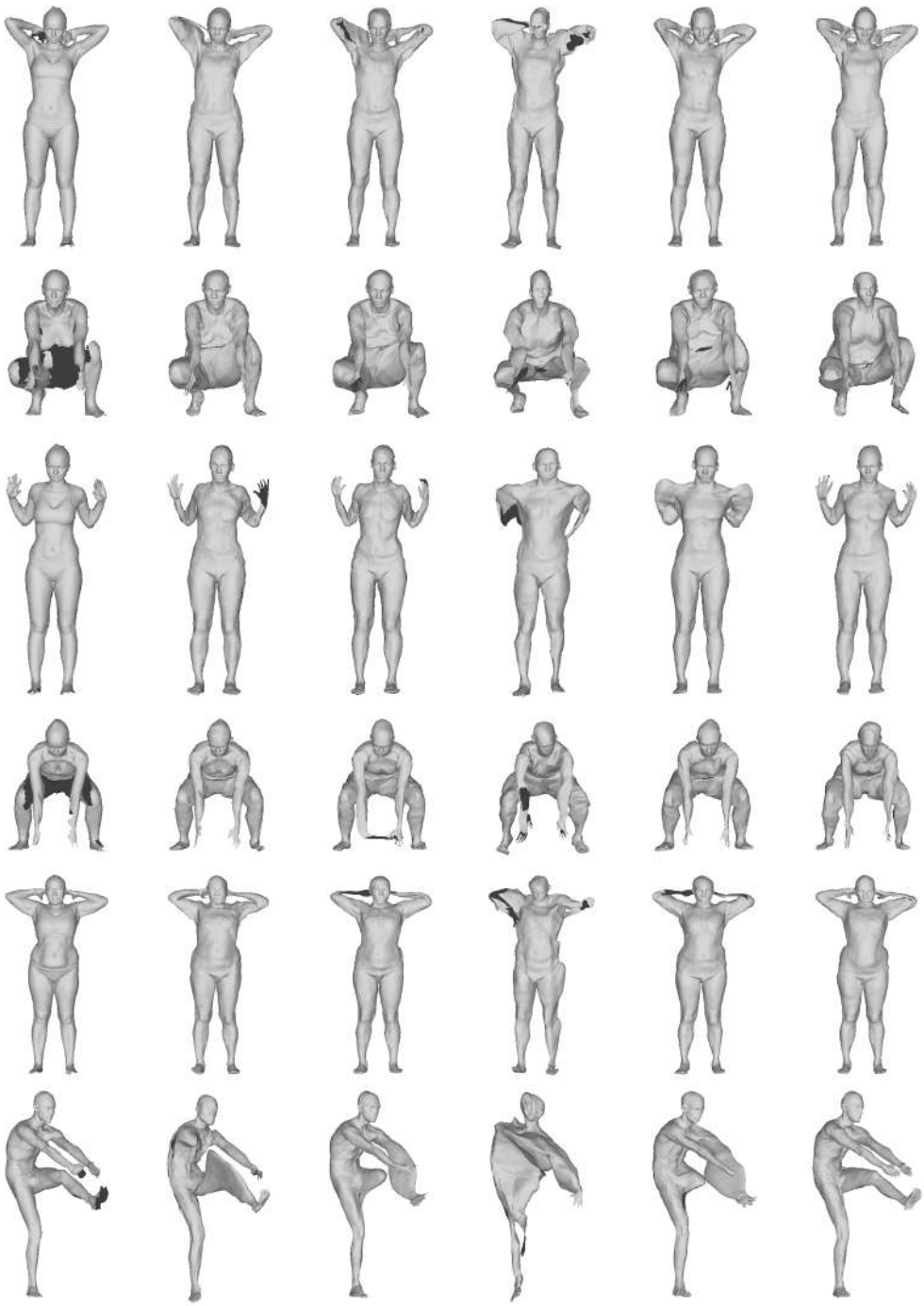


Figure 3.5: Challenging cases in FAUST. From left to right, (a) input scan, reconstruction in high resolution of (b) 3D-CODED Groueix et al. [2018a], (c) AtlasNet2 Deprelle et al. [2019], (d) Our MLP, (e) our GP and (f) our Adversarial GP. We highlight the failure cases with red box.



(a) scan **(b)** w/o MLP **(c)** w/o Adv **(d)** L2 **(e)** Dropout **(f)** Adv GP

Figure 3.6: Quantitative results in FAUST challenge. From left to right, (a) input scan, reconstruction in high resolution of (b) Adv+GP (without MLP), (c) MLP+GP (without Adv), (d) MLP+GP+L2, (e) MLP+GP+Dropout and (f) Adversarial GP.

feature learned by the encoder which lacks the representation capability of unseen pose. Failure reconstruction arises when the pose of human is never seen during training. Another limitation lies on the pre-defined topology which is not suitable for various shapes, especially dressed person. Therefore, we move to neural implicit modeling in Chapter 4 and 5 to flexibly represent different human shapes and poses.

Future work will apply the proposed framework to problems like motion sequence alignment and tracking as natural extension. The Gaussian Process can not only encode spatial consistency but also explore temporal coherence for dynamic model. We believe that considering the spatial and temporal priors together would have further benefit for motion sequence.

Another interesting direction with GP layer would be style transfer or identity transfer. In the transfer problem, it is difficult to handle a very dense representation of human shape. Here GP layer can serve as a compression and densification technique to reduce the dimensionality of human vertices. Unlike PCA-like method compresses human shape with parameters based on statistical variation, our GP layer gives the mapping between the sparse and dense representation with the function of geometry localization.

Chapter 4

Spatio-Temporal Human Shape Completion With Implicit Function Networks

4.1 Introduction

In this chapter, we examine the problem of 3D human shape estimation from incomplete 3D observations, *e.g.*depth images, under motion obtained from a single camera. This under-constrained problem requires additional information which can be provided by a learned model but also by leveraging observations over time when considering temporal sequences. We investigate how to benefit from both through data driven spatio-temporal modeling. The input is a sequence of depth-like partial observations, our network will predict the occupancy of any query point in space at any time stamp, which allows us to complete the representation in both space and time domains, *e.g.* Figure 4.1.

Building human shape models from incomplete 3D observations over time is a challenging task with many applications in augmented and virtual reality or telepresence applications, among others. Part of the difficulty lies in the choice of the shape representation, which can be global and encode high level features of human shape such as pose, or more local features to identify details on the surface of observed subjects. In fact many shape models used to complete partial shape data combine both aspects to leverage both of their advantages, *e.g.*by constraining local shape refinement with a global pose and body model of humans underlying to the shape, *e.g.* Loper et al. [2015], Pons-Moll et al. [2017], Zhang et al. [2017], Alldieck et al. [2018b]. But how to balance those aspects is

often manually hardwired in the existing methods, especially with classic pre-learning inference and reconstruction models.

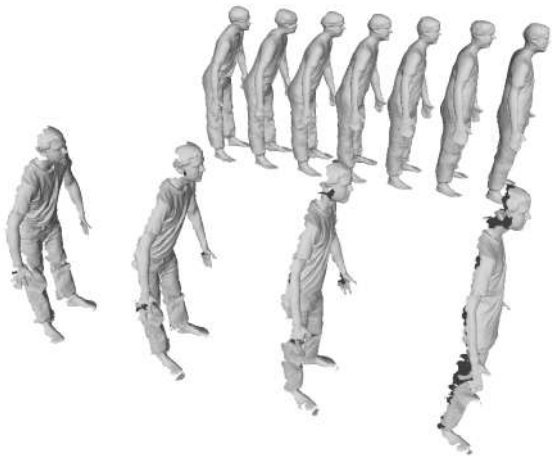


Figure 4.1: Given incomplete temporal 3D observations as input, here 4 samples, our STIF-Nets reconstruct their complete 3D shape and provide unseen interpolated frames.

The advent of Deep neural Networks (DNN) has brought a whole new set of possibilities to enhance inference and tackle single-view 3D reconstruction problems with data-driven priors, but has simultaneously made the representation problem even more open, because allowing DNNs to account and optimally operate on 3D information has proven to be non-trivial. This is even more prominent when one tackles 4D space-time applications due to the added dimensionality, and the even more massive amount of training data needed to train models in this context. In fact this has been such a barrier that the literature in data driven 4D shape modeling is quite scarce at the time of this writing.

With this in mind, a particular category of implicit methods for 3D shape representation is rapidly gaining attention Mescheder et al. [2019], Chibane et al. [2020a]. By encoding the shape implicitly using an indicator function parameterized by MLPs to express the occupation at a given point in space, these methods have succeeded in reducing the dimensionality of the network needed for 3D shape inference problems and allowed a continuous representation of 3D shapes to be embedded in the inference problem. Expectedly, these desirable characteristics have translated to 3D shape inference results of very promising quality. To our knowledge, they have yet to be extended to spatio-temporal evolution of shapes, for which the reduction of dimension could provide a key benefit.

Our intent is to bridge this gap, by providing a new implicit spatio-temporal model by which better shape inference and completion can be achieved, given temporal depth sequences. In so doing, we target improved shape and motion quality by going beyond static shape priors to learn spatio-temporal shape-motion priors. To this goal, our model uses a U-Net encoder to produce an image-dimensioned feature map, similarly to Mescheder et al. [2019], Niemeyer et al. [2019], Chibane et al. [2020a], Saito et al. [2019], but instead of only encoding a per-pixel implicit depth indicator function, our features parameterize a per-pixel implicit space-*time* depth evolution indicator function. To balance global and local temporal aspects in our model, we use the bidirectional GRU Chung et al. [2014] to connect latent global features encoded by the U-Net from previous and subsequent frames, an architecture we coin U-GRU Encoder.

Using two databases of human motion in and without clothing, we assess the qualitative gain of this architecture on analysing monocular video sequences for 3D dynamic human shape estimation, comparing with per-frame estimation methods over the set of input frames, and also examining temporal densification, which can be performed by sampling shape estimates at intermediate time stamps of the implicit spatio-temporal function. Our experiments not only show high quality results for the interpolation task, comparable to results obtained from per-frame methods had they been provided with intermediate inputs; but also show improvement of quality of the 3D models retrieved with actually observed frames, with respect to per-frame methods.

4.2 Related Work

A dense analysis of shape completion can be found in Section 2.4. Here we focus on previous works that are most related to ours in the spatial and temporal completions.

4.2.1 Spatial Shape Completion

In order to complete a shape given a partial observation at a given time, methods have been proposed that differ with respect to the shape representations they consider. These representations can be either discrete with *e.g.*point based representations or continuous as with neural implicit functions that encode occupancy.

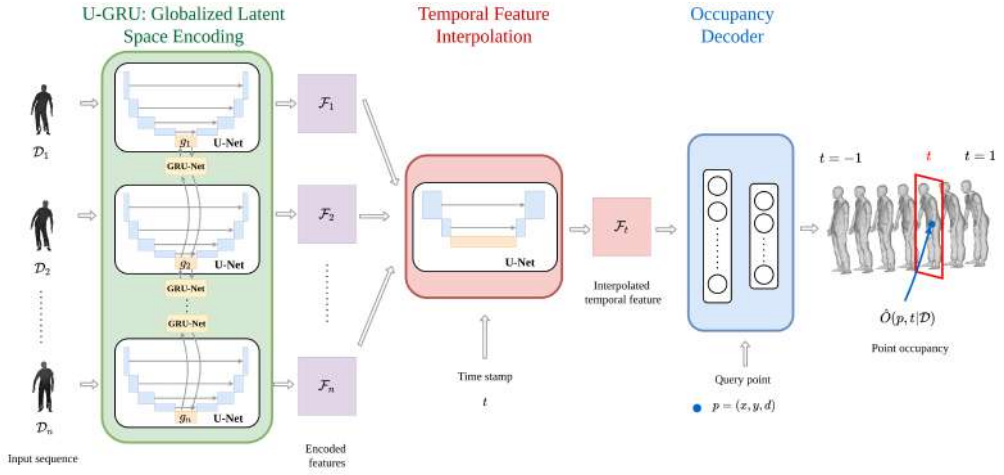


Figure 4.2: STIF-Nets Overview. Our architecture is articulated among three phases: (1) simultaneous encoding-decoding of input sequence frames to a set of features using U-Nets whose latent spaces are interconnected thanks to GRU networks (§4.3.1); (2) Temporal interpolation of the feature space with U-Net encoder (§4.3.2); (3) Occupancy decoding along the viewing line (§4.3.3).

With explicit point based representations, strategies were explored that use prior assumptions, such as parametric or template shape models, to guide shape completion. For instance, Prokudin et al. [2019], Bogo et al. [2016], Kanazawa et al. [2018] reconstruct the human body by inferring the parameters of the SMPL model Loper et al. [2015], a popular parametric model for undressed humans. In another work, LBS-AE Li et al. [2019], Linear Blending Skinning parameters are learned from point cloud in a self-supervised way. Relaxing somewhat the constraints on the shape model, other approaches use a template, for example a mesh, to model human shapes. In this line of work approaches such as 3D-CODED Groueix et al. [2018a], Deprelle et al. [2019], Halimi et al. [2020], Zhou et al. [2020] deform a template using global features extracted from partial observations with PointNet Qi et al. [2017]. BPS Prokudin et al. [2019] learns the 3D point cloud descriptor and is able to reconstruct the SMPL-topology mesh from it. While strong prior models clearly help the completion task they also limit its applicability to reduced shape spaces, *e.g.*undressed human bodies.

Implicit representations have also been largely exploited to model occupancy in 3D. In the discrete case of voxels, the grid regularity allows the extension of CNN-based methods to 3D and the ability to infer human shapes with data-driven strategies as in Varol et al. [2018], Zheng et al. [2019]. Voxel based representations suffer anyway from complexity issues and recent strategies have

taken a continuous approach with implicit representations, *e.g.* Mescheder et al. [2019], Xu et al. [2019], Saito et al. [2019], Chibane et al. [2020a], Saito et al. [2020], Deng et al. [2020a]. In this case, occupancy is modeled at any 3D location and learned with ground-truth examples. OccNet Mescheder et al. [2019], a seminal work in this category, can be used to complete shapes but is missing local input features, which are crucial to preserve human shape details. SAL Atzmon and Lipman [2020], IGR Gropp et al. [2020] and SALD Atzmon and Lipman [2021] learn the implicit representation of human, but rely on optimizing the latent code to fit the uniformly sampled dense point cloud. NASA Deng et al. [2020b] encodes the articulated human conditioning only on pose with implicit function, which is identity-dependent. LEAP Mihajlovic et al. [2021] and SCANimate Saito et al. [2021] learn the human implicit representation by using the Linear Blending Skinning. However, LEAP Mihajlovic et al. [2021] requires key joint information which is identity- and pose-dependent. In general, optimization-based post-processing stage in Atzmon and Lipman [2020], Gropp et al. [2020], Atzmon and Lipman [2021], Deng et al. [2020b], Saito et al. [2021] could not be easily applied with partial view input. Recently, IF-Net Chibane et al. [2020a], which stacks 3D convolutions to extract features at different scales for query points, can preserve human body details. It is in practice compute-hungry to train, as a consequence of the voxel representation required to perform 3D convolutions. While we use a similar strategy, we lift the problem to the spatio-temporal domain and reduce the complexity by basing our network in the 2D pixel domain as for Saito et al. [2019] but with depth-time implicit queries. We also note in this category of work Bhatnagar et al. [2020a] which combines IF-Net and SMPL to register shapes, with still the aforementioned limitations.

4.2.2 Spatio-Temporal Shape Completion

Besides model based approaches that perform temporal shape predictions using parametric representations based on rigged skeletons *e.g.*Martinez et al. [2017], Zhou et al. [2018], Chiu et al. [2019], Cao et al. [2020] or more elaborate models such as SMPL *e.g.*Tung et al. [2017], Alldieck et al. [2018a,b], few works consider spatio-temporal completion per-se. Akhter et al. [2012] can complete shape sequences both spatially and temporally by using decompositions over spatio-temporal basis and by assuming temporally consistent shape models, *i.e.*with fixed topology, for that purpose. The approach targets more

sparse missing data in temporal trajectories than large completions of shapes with inconsistent topologies as we do. Considering point trajectories, flow based methods have been proposed that model spatio-temporal shape evolution and can therefore perform predictions, *e.g.*Niemeyer et al. [2019], Yuan et al. [2020], Jiang et al. [2020c], Yang et al. [2019]. DynamicFusion Newcombe et al. [2015] reconstructs 3D information by using the fusion of multiview depth image, which is not in our case of single view completion. In this category, OccFlow Niemeyer et al. [2019] is close to our objective with an approach that predicts spatial-temporal occupancies. However, considering point trajectories implies temporal correspondences which are often difficult to obtain and also sparse missing data instead of significant completions.

4.3 Method

Our goal in this chapter, given a monocular sequence of input depth frames representing incomplete shapes, is to infer a set of complete and temporally densified or predicted shapes. By incomplete inputs, we mean that frames are typically presumed to be obtained from time of flight cameras or front scans with depth sensing technologies, with back and occluded data missing.

Let $\mathcal{D} = \{\mathcal{D}_i\}_{i\in\{1,\cdots,\mathrm{n}\}}$ be the discrete time sequence of input depth frames of resolution res $\times$ res. As we seek benefit from the lean and continuous parameterization an implicit representation offers, following several similar papers Mescheder et al. [2019], Chibane et al. [2020a] we model the problem with a DNN representing the occupancy regression function of a query point $p = (x, y, d)$ with continuous pixel coordinates $(x, y) \in [1, \mathrm{res}]^2$, and continuous depth $d \in \mathbb{R}$, and given a time stamp in the continuous interval $t \in [-1, 1]$ representing the initial frame interval $[1, \mathrm{n}]$. This function produces predictions of the occupancy $\hat{\mathcal{O}}(p, t|\mathcal{D}) \in [-1, 1]$ of the query point given the input depth images $\mathcal{D}$.

Note that, contrary to other similar inspirational works which focus on 3D volumetric inference Chibane et al. [2020a], we choose a similar 2D discrete support grid similar to Saito et al. [2019] instead, for targeted memory and computational efficiency improvements necessary to deal with the additional temporal dimension. We also explicitly focus our method on the output surface at the zero crossing of the occupancy function $\hat{\mathcal{O}}(p, t|\mathcal{D})$, similar to TSDF-based methods Innmann et al. [2016], which can be efficiently extracted using

a marching cubes Lorensen and Cline [1987], Lewiner et al. [2003] algorithm, and evaluate points in the vicinity of the surface, as opposed to volume-centric approaches Chibane et al. [2020a] which tend to infer volumetric occupancy functions in $[0,1]$ for regular volumetric grids.

In order to make the problem tractable and decompose the function according to its main factors, we build our network architecture along three phases, illustrated in Figure 4.2. In the first (§4.3.1), we decode a set of 2D feature maps $\mathcal{F} = \{\mathcal{F}_i\}_{i\in\{1,\cdots,\mathrm{n}\}}$ from each 2D input depth image $\mathcal{D}_i$ of identical resolution res $\times$ res but introduce a global correlation to allow the network to learn global motion features linking them. We input the queried continuous time variable t in the second phase (§4.3.2), and jointly use it with the full set of feature maps to decode a t-specific interpolated 2D feature map $\mathcal{F}_t$ also matching the input resolution res $\times$ res. This feature map is then used jointly with the query point p to decode the final occupancy result $\hat{\mathcal{O}}(p,t|\mathcal{D}) \in [-1,1]$, as described in §4.3.3.

4.3.1 Globalized Latent Space Encoding

In this phase, we want to allow the network to extract relevant feature maps $\mathcal{F}_i$ for each input $\mathcal{D}_i$ that preserve some detail, while simultaneously allowing the method to be aware of global aspects such as the underlying subject motion. To this goal we opt for a 2D U-Net encoder-decoder structure Ronneberger et al. [2015] per-input frame, which projects its inputs on a low dimensional latent space and lifts it back to an output matching the input size, using four symmetric downsampling convolution and upsampling deconvolution layers. U-Net also has the property to balance global aspects of the frame with local ones, using skip connections between matched convolution and deconvolution layers, that allow to preserve local and high frequency details for creation of the feature map, while still allowing for efficient training. Accounting for the expected symmetry between the features extracted for the various input frames, we propose to train the n U-Net instances with shared weights.

We however still need to account for shared temporal aspects and interframe motion. To this end, we link each U-Net's latent space vector with those of both temporally adjoining frames using a bidirectional Gated Recurrent Unit, or GRU, to learn interframe residuals of the latent space. The intent is to force interframe phenomena to be treated as a global, transpixel phenomena. The

choice of GRU is motivated by its use in Natural Language Processing, where it was shown to perform similarly to LSTMs with fewer parameters and easier training Chung et al. [2014], Cho et al. [2014]. We show the combination of U-Net and GRU, which we coin U-GRU, to significantly improve training results (Tab. 4.3). Thus phase 1 of our network can be seen as a global feature decoder solution from the input frame set to the set of intermediate feature maps, which are all individually made aware of interframe cues:

$$\mathcal{F} = \text{U-GRU}(\mathcal{D}). \tag{4.1}$$

4.3.2 Temporal Feature Interpolation

With the feature map set still global to the entire input sequence, we propose in a second phase to extract an interpolated feature map $\mathcal{F}_t$ which is specialized for the queried time t. We concatenate all feature maps together and add t weighed by a constant normalization factor $c_t \times t$ as an additional constant input channel $\mathcal{T}$ to every pixel of the map, and feed this aggregate to a simpler U-Net Ronneberger et al. [2015] with a pixel-wise 1×1 convolution operator, two levels of downsampling convolutions and upsampling operations, to decode $\mathcal{F}_t$ from $\mathcal{F}$:

$$\mathcal{F}_t = \text{U-NET}(\mathcal{F}, \mathcal{T}). \tag{4.2}$$

With this architecture choice, the network can learn its temporal interpolation function while automatically adjusting between both global and local per-pixel components of the interpolation. So the U-Net here is able to reduce the aliasing effect of the interpolated feature. We believe that the Temporal Feature Interpolation could remedy the missing information, *e.g.*hole, noise or occlusion, from one single frame by considering temporal information from previous and next frames(Figure 4.8(e)(g)). Note again that the network can be trained with any continuous $t \in [-1, 1]$ where -1 stands for the first given frame and 1 represents the last frame.

4.3.3 Occupancy Decoder

This third and last phase focuses on spatial decoding of the occupancy $\hat{\mathcal{O}}(p, t|\mathcal{D})$ of a given query point $p = (x, y, d)$. We bilinearly interpolate a feature $\mathcal{F}_t(x, y)$

specific to the real-valued (x, y) from feature map $\mathcal{F}_t$ for query point p. Then we associate the depth query value d weighed by normalizing constant $c_d \times d$ with $\mathcal{F}_t(x, y)$ as the input for an MLP regressor with the following characteristics. The aggregate feature $\{\mathcal{F}_t(x, y), d\}$ of query point p at time t is sent to two linear layers. The first one is activated using the widely used RELU and second one using TANH which conveniently produces occupancy values $\hat{\mathcal{O}}(p, t|\mathcal{D})$ in the target interval $[-1, 1]$:

$$\hat{\mathcal{O}}_t(p|\mathcal{D}) \triangleq \text{MLP}(\text{U-NET}(\text{U-GRU}(\mathcal{D}), \mathcal{T}), p). \tag{4.3}$$

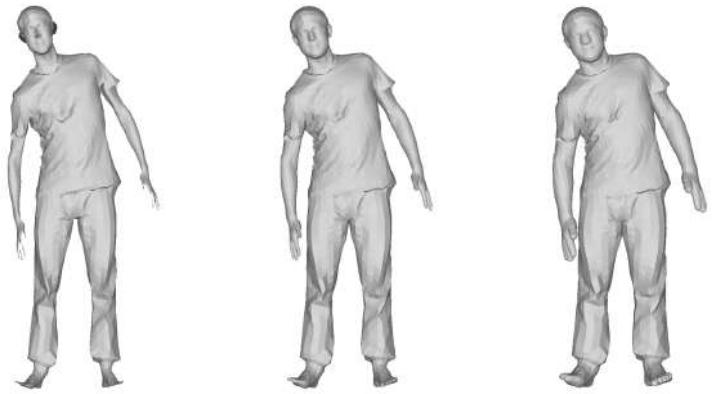
4.3.4 Training

The proposed network can be trained for various tasks, *i.e.*shape completion of input frames, temporal interpolation or densification of frames. We propose a uniform supervised training procedure for all of these cases. For this we consider that, for a given batch of ground truth training sequences B, we are given occupancy samples $\mathcal{O}_{p,j}$ with a randomized point set $p \in P_j$ and their matching inputs $\mathcal{D} = \{\mathcal{D}_j\}_{j \in 1,\cdots,\mathrm{m}}$, from a set of ground truth frames with time stamps $\{t_j\}_{j \in 1,\cdots,\mathrm{m}}$. Typically this set will include time stamps that match the input frames and some additional training examples regularly interspaced between input frames. The training can then be realized by minimizing a mean square loss over the set of network parameters θ:

$$\theta^* = \arg\min_\theta \sum_B \sum_{t \in T} \sum_{p \in P_j} \|\hat{\mathcal{O}}(p, t|\mathcal{D}) - \mathcal{O}_{p,j}\|^2. \tag{4.4}$$

To account for the continuous nature of the occupancy function $\hat{\mathcal{O}}(p, t|\mathcal{D})$, we create more temporal samples than given in training, by drawing t from a randomized, denser set T within the training interval, and use j of the time stamp closest to t in the above training procedure.

Point sampling strategy. The choice of the training point set P_j is an important one. Naive strategies would be to use uniformly randomized or regularly spaced samples over the whole sequence's bounding box to present the training with positive and negative examples. This is however quite inefficient as it wastes most of the advantage of modeling the occupancy as an implicit function, the main point being to decorrelate the training complexity from dense 3D space sampling that would occur with regular grid CNNs. Mescheder et al.



(a) o = +0.5 **(b)** o = 0 **(c)** o = -0.5

Figure 4.3: Point sampling strategy. From left to right, (a) shrunk surface, (b) origin surface, and (c) expanded surface. o stands for occupancy value.

[2019], Chibane et al. [2020a] use a Gaussian sampling strategy at the vicinity of the surface and train with a classification loss. We propose a simpler yet experimentally efficient sampling which leverages our surface level set parameterization, by providing samples from 3 distinct surfaces, see Figure 4.3, in the vicinity of the true surface: one corresponding to the true surface location with training label 0, the expanded and the shrunk surfaces with positive and negative displacement along the surface normal, with respective label sets in $o \in \{-0.5, +0.5\}$. The shrinking and expansion factor along the normal we choose is the label o multiplied by a constant scale factor l. n_s samples are used for every surfaces, and we also sample n_s points from inner part of shrunk surface with $o = +1$ and $3 \times n_s$ points from outer part of expanded surface with $o = -1$ as we empirically observe the need for more negative samples with the expansion.

4.3.5 Implementation Details and Inference

We implement our STIF-Nets in PyTorch and train it from scratch. The global network architecture is shown in Figure 4.2. Here, the details of the three blocks are shown in Figure 4.4. In practice, $\mathrm{n} = 4$ depth images with the dimension of $(1, reso, reso)$ are processed with U-GRU Encoder. Then, 4 feature maps and 1 temporal channel are concatenated along the channel dimension. The constant temporal channel is expanded with $c_t \times t$. So the input to the Feature Interpolation phase is with the dimension of $(4 \times 128 + 1, reso, reso)$. Once the feature map is interpolated at time stamp t, the feature of point p can be queried on

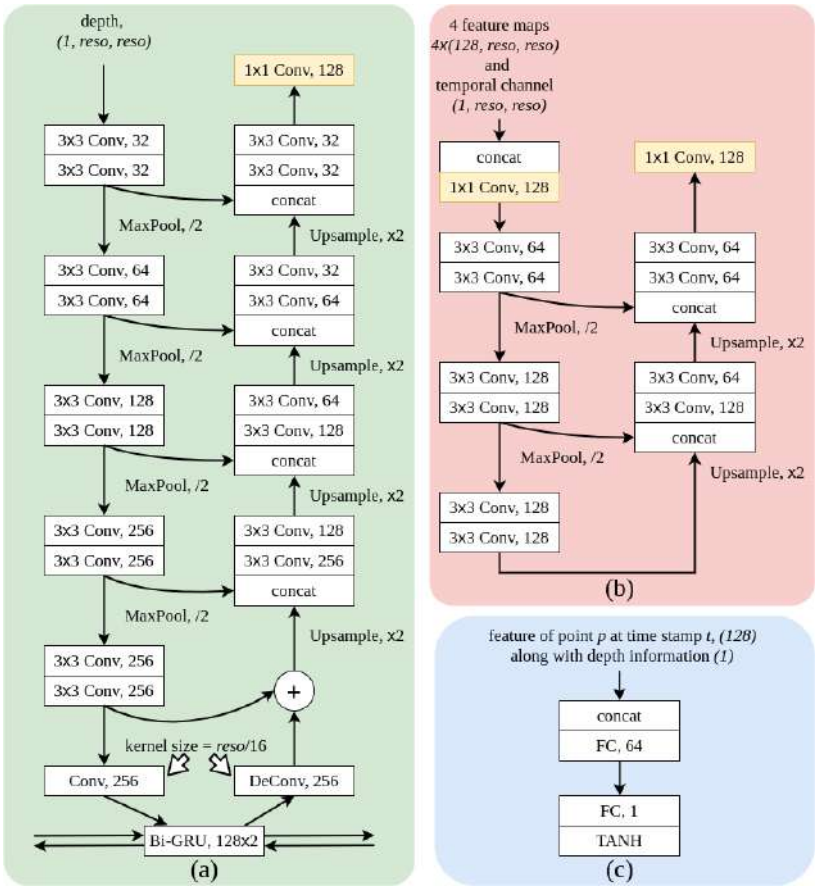


Figure 4.4: STIF-Nets architecture: (a) U-GRU Encoder, (b) Temporal Feature Interpolation U-Net and (c) Occupancy Decoder. Conv, FC, reso and concat stand for 2D convolution, fully-connected layer, resolution and concatenation operation. The three yellow blocks are not activated with any activation function while the other Conv layers in (a) and (b) and the the first FC in (c) are activated with RELU. The output is activate with TANH to bound the occupancy prediction in the interval [-1, 1]. Here, we ignore the dimension of batch size.

the feature map with the horizontal and vertical coordinates. This feature and the depth information, $c_d \times d$, are concatenated, which are sent to the Occupancy Decoder. During the training, we set the sampling number $\mathrm{n}_s = 300$, the expanded/shrunken length $l = 0.02$ in the normalized space, two coefficients $c_t = \mathrm{res}, c_d = \max(\mathrm{res}, 256)$. Due to the limitation of GPU memory for sequential data, we set the batch size to 1 and we drop the Batch Normalization in the original U-Net implementation. We use Adam optimizer with the learning rate of 0.0001. During inference, we set the resolution of 3D occupancy grid to 256^3 for all occupancy-based methods, except the one reported as depth/grid resolution $512/512$ in Tab. 4.4, which is different from res for the depth image. Marching cubes Lorensen and Cline [1987], Lewiner et al. [2003] is applied to extract the zero-level set of the computed occupancy grid as a surface mesh.

4.4 Experimental Evaluation

In order to evaluate STIF-Nets we conducted quantitative and qualitative comparisons on the shape completion task given depth image sequences, this for both input frames and new interpolated, focusing on body shapes in motion. In the following we provide numerical results as well as ablation studies that shed light on how the main components of STIF-Nets impact the performance. We first detail the data and metrics used in our evaluation.

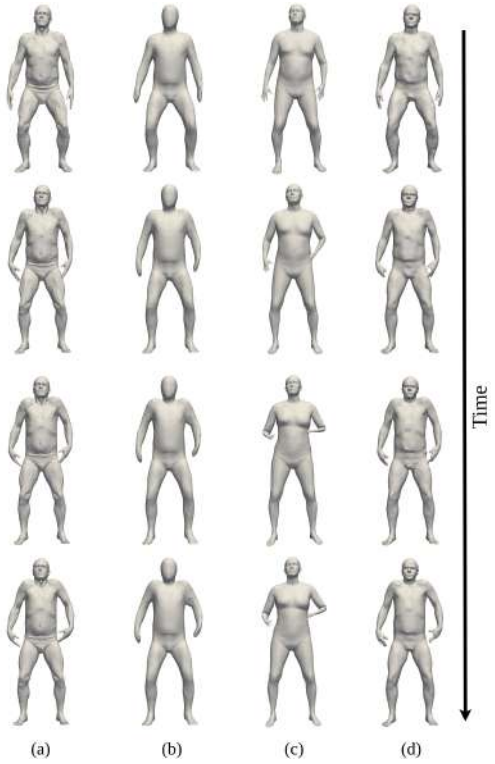


Figure 4.5: Sequence reconstruction. From left to right, (a) ground truth, reconstruction of (b) OccFlow Niemeyer et al. [2019], (c) BPS Prokudin et al. [2019] and (d) our STIF-Nets.

4.4.1 Data and metric

We collected human motion data from a clothed human dataset CAPE Ma et al. [2020] which is based on 4D capture ClothCap Pons-Moll et al. [2017],with two clothing styles dressed on each character, and an undressed human dataset DFAUST Bogo et al. [2017]. Both datasets contain real scans that were captured at 60fps and fit with the SMPL model Loper et al. [2015], which provides our ground truth surface models. From the scans we created front view depth images

Data	CAPE		DFAUST	
Method	IoU ↑	Chamfer ↓	IoU ↑	Chamfer ↓
3D-CODED Groueix et al. [2018a]	0.455	0.591	0.578	0.347
OccNet Mescheder et al. [2019]	0.488	0.476	0.604	0.340
IF-Net Chibane et al. [2020a]	0.787	0.155	0.822	0.134
Chibane et al. [2020a](×2)	0.804	0.143	0.840	0.127
OccFlow Niemeyer et al. [2019]	-	-	0.740	0.231
BPS Prokudin et al. [2019]	-	-	0.761	0.197
Our STIF	**0.822**	**0.123**	**0.858**	**0.111**

Table 4.1: Spacial completion with IoU and Chamfer-L1 distances ($\times 10^{-1}$) for 4 interframe intervals.

where the depth of a pixel is determined by the front-facing scanned 3D point that is closest to the pixel viewing-line. Note that we preserve the hole and noise of real scans in our processed depth image in order to test the robustness of our method, see Figure 4.1, 4.2 and 4.7.

Training and Test Sets: The training set includes 8 characters, 2 male and 2 female from each dataset. For each character, we selected 3 or 4 motion sequences for a total of 28 sequences. Within each motion sequence, we extracted 6 sub-sequences composed of 4 frames. These sub-sequences are of 2 types: 4-interval sequences with interframe intervals of 4 and approximately 200ms durations; and 10-interval sequences with interframe intervals of 10 and approximately 500ms durations. In addition, 3 frames, taken randomly within each sequence were added to the 4 frames with the objective to more robustly train interpolation. Both 4 and 10-interval sequences were used in the training for a total of 336 input sequences. Note that we train our STIF-Nets only once across dressed and undressed characters and short-term and long-term sequences. The test set includes 9 characters, 4 from CAPE and 5 from DFAUST, who perform 54 input sequences. 2 characters in the test set were completely unseen during training. The seen shape characters perform the different motion styles from the training set.

Metric: We evaluated the completions using 2 metrics: The volumetric intersection over union (IoU) and a surface based Chamfer-L1 distance. Note that numerical values were computed with the meshes obtained by the Marching cubes algorithm applied on the occupancies predicted by STIF-Nets. In practice, we noticed that the IoU metric, which is volumetric, hardly differentiates the approaches whereas the Chamfer distance, a surface metric, provides more insights though being more sensitive to noise.

4.4.2 Frame Completion

Using the data and the metrics mentioned in the previous section we conducted comparisons of STIF-Nets with representative state of the art methods: one point based method, 3D-CODED Groueix et al. [2018a] and two recent implicit function based methods, OccNet Mescheder et al. [2019] and IF-Net Chibane et al. [2020a]. We retrain the 3D-CODED, OccNet and IF-Net on our dataset and the numeric results with 4 frames intervals are shown in Table 4.1. Note that our models provide clearly better results with both IoU and Chamfer distance. IF-Net processes the partial scan data into voxel occupancy. To fairly compare with IF-Net, we set the same total resolution of voxels as our processed depth image, and also compare with IF-net inputting the voxels of 2 times resolution. Note that our method outperforms both. In addition, we evaluate OccFlow Niemeyer et al. [2019] and BPS Prokudin et al. [2019] on DFAUST. OccFlow is pretrained on DFAUST dataset and BPS is pretrained on CAESAR Robinette et al. [2002] which is a very large undressed human dataset, as supplied by authors.

Static completion comparisons are presented in Figure 4.8. They show that both 3D-CODED Groueix et al. [2018a] and OccNet Mescheder et al. [2019] have difficulties preserving shape details, as a result of the missing local features in these methods. We also prepare our static model which replaces the U-GRU encoder by the U-Net and remove the second Feature Interpolation phase. The qualitative results in Figure 4.8 show that all static approaches including ours and IF-Net Chibane et al. [2020a] present artefacts, often resulting from holes and noise in the raw input scans. On the other hand, dynamic approaches appear more robust, with our STIF-Nets outperforming the naive temporal baseline. The Feature Interpolation phase is able to reduce the aliasing effect of bilinear sampling of feature map in the Occupancy Decoder, which preserves the high frequency feature on the face, belly and even cloth. We also prepare a naive dynamic baseline which drops the GRU and the dimensionality in the simpler Temporal Feature Interpolation is (32, 32, 32, 64, 64, 64, 32, 32, 32, 128). The naive dynamic baseline is not able to handle the temporal information without U-GRU Encoder. Some additional qualitative comparisons are shown in Figure 4.9, which include three new identities in the datasets. Remark that the missing part of head in the last example of Figure 4.9 degrades the reconstruction of the two static baselines, but the STIF-Nets could fix it by considering temporal information.

Data	CAPE		DFAUST	
Method	IoU ↑	Chamfer ↓	IoU ↑	Chamfer ↓
Our Static(depth)	0.651	0.512	0.712	0.380
(neighbour)	0.753	0.158	0.777	0.154
(latent)	0.788	0.157	0.812	0.154
3D-CODED Groueix et al. [2018a]	0.456	0.592	0.578	0.349
OccNet Mescheder et al. [2019]	0.488	0.475	0.604	0.337
IF-Net Chibane et al. [2020a]	0.791	0.158	0.826	0.143
Chibane et al. [2020a](×2)	**0.806**	0.150	0.841	0.143
Our STIF	**0.806**	**0.139**	**0.842**	**0.133**

Table 4.2: Temporal interpolation with IoU and Chamfer-L1 distances ($\times 10^{-1}$) for 4 interframe intervals.

The best comparison with OccFlow and BPS would require full scan data instead partial scans, which is significantly less challenging than our scenario, where the performance of these two methods degrades. In Figure 4.5, OccFlow is not able to preserve the high frequency feature on the human body and the BPS descriptor is sensitive to the noise on the real scan data.

We also experiment the influence of interval frames on the completion result. Note that we do not train again our STIF-Nets for 2, 6 or 8 interframe intervals. In Figure 4.6, as can be expected the dynamic model efficiency reduces with increasing inter frame intervals, but this gap is not large between short-term sequence and long-term sequence.

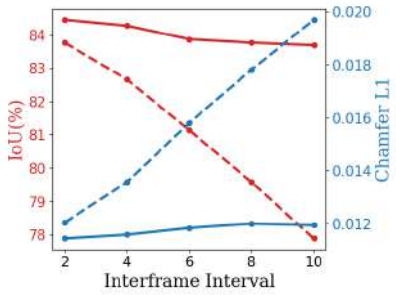


Figure 4.6: Evaluation of our STIF-Nets with different interframe intervals. The solid line —— stands for completion task and the dashed one - - - stands for interpolation task.

Our method outperforms the best method in the compared baseline, IF-Net, both in speed and performance. In our tests, the IF-Net average per-frame computation time is 4.791 s/frame, whereas our static baseline and STIF-Nets run in 0.343 and 0.349 s/frame on a GeForce RTX 2080Ti, using the same total input and occupancy grid resolution. Our STIF-Nets pays only 6ms per frame penalty for using the temporal information.

	Input Frame Completion				Interpolation			
Data	CAPE(dressed)		DFAUST(undressed)		CAPE		DFAUST	
Method	IoU ↑	Chamfer ↓	IoU ↑	Chamfer ↓	IoU ↑	Chamfer ↓	IoU ↑	Chamfer ↓
Our Static + basic sampling	0.788	0.143	0.829	0.129	0.784	0.178	0.825	0.176
Our Static	0.804	0.128	0.845	0.113	0.788	0.157	0.812	0.154
Our Native Dynamic	0.713	0.183	0.737	0.180	0.733	0.176	0.760	0.175
Our STIF	**0.822**	**0.123**	**0.858**	**0.111**	**0.806**	**0.139**	**0.842**	**0.133**

Table 4.3: Quantitative comparisons with IoU and Chamfer-L1 distances ($\times 10^{-1}$) for 4 inter-frame intervals on both completion and interpolation tasks.

depth(^{2})/grid(^{3}) resolution	IoU ↑	Chamfer ↓
128/256	0.810	0.143
256/256	0.843	0.116
512/256	0.846	0.113
512/512	0.858	0.104

Table 4.4: Impact of the depth image and occupancy grid resolution for shape completion with IoU and Chamfer-L1 distances ($\times 10^{-1}$).

4.4.3 Frame Interpolation

We also experimented the ability to interpolate between the input frames from partial data. For the interpolation task, the evaluation is performed at the 3 middle frames within the 4 intervals of each sequence. From the static case, a naive interpolation baseline can be achieved using different strategies: nearest neighbor frame, depth image interpolation or latent representation interpolation. Numerical comparisons in Table 4.2 show that the latter performs the best and we therefore only report results with latent space interpolation in Table 4.2 and 4.3 for other static methods. Table 4.2 also demonstrates that STIF-Nets outperforms static interpolation with both metrics. Figure 4.10 illustrates frame interpolation. The spatio-temporal modeling is able to preserve the volume and the high frequency features even on the Interpolation task which is difficult for static models. In Figure 4.6, the interframe interval notably influences the interpolation performance, which degrades gracefully given that we did not increase the supervision of interpolation task during the training of long interval.

4.4.4 Ablation Studies

Table 4.3 reports on two crucial elements in our method: sampling and dynamic modeling. It shows that our proposed sampling strategy benefits with respect to the Gaussian sampling from surface used in the IF-Net Chibane et al. [2020a].

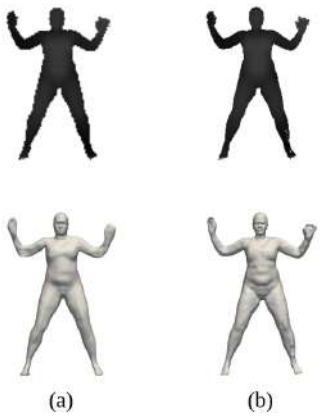


Figure 4.7: Qualitative results with (a) 128^2-resolution depth image and (b) 256^2-resolution depth image for STIF-Nets.

Again STIF-Nets quite significantly outperforms the static approach quantitatively. To illustrate the efficiency of our dynamic modeling, we prepare a naive dynamic baseline which drops the GRU in Encoder, considers the 4 frames as 4 channels to extract the feature map and uses a Temporal Feature Interpolation with lower dimensionality. For both spatial completion and temporal interpolation tasks, our STIF-Nets better preserve the reconstruction volume, higher frequency facial & surface details and input similarity(Figure 4.8 and 4.9). The Feature Interpolation phase is able to predict the feature map at any queried time stamp in order to fill the gap between spatial completion and temporal interpolation.

We also experiment the STIF-Nets with different resolution of input depth image. Table 4.4 reports that increasing the input depth image resolution could benefit the reconstruction accuracy and the 3D occupancy grid resolution, used for the Marching cubes Lorensen and Cline [1987], Lewiner et al. [2003] during inference, plays as well an important role. Figure 4.7 shows that more details on the face and belly are extracted by STIF-Nets with high resolution depth image than the low resolution one.

4.5 Conclusion

In this chapter, we have presented STIF-Nets, a deep network architecture to model shapes from incomplete observations. STIF-Nets builds on neural implicit function representations, which has proved efficient for shape modeling.

The key contribution with respect to existing works is to lift these representations to the spatio-temporal domain, hence leveraging information over time and enabling shape completions over both the spatial and temporal domains. In this chapter we propose to generalize implicit encoding to spatio-temporal shape inference with STIF-Nets, where temporal redundancy and continuity is expected to improve the shape and motion quality. Experiments demonstrate that STIF-Nets contributes with improved robustness, shape quality and generalization abilities with respect to purely spatial strategies. There still exist some directions to investigate as the future work.

How could we exploit more from temporal modeling? We have examined the performance of STIF-Nets with different interframe intervals in Section 4.4, which reflects that the completion is degraded for long-term interval. Due to the limitation of GPU memory, we can not currently extend STIF-Nets to a large number of frames. So we believe that long distance frame could not necessarily contribute to current frame. Instead of exploring long-term sequence, we will discuss the possibility of aggregating the corresponding local feature and extracting the high-frequency details in 4-frame sequence in Chapter 5.

Another interesting direction would be motion prediction with neural implicit modeling. We have already investigated one potential application of our STIF-Nets for motion temporal interpolation in Section 4.4. The enhancement result is physically plausible. Other than that, the temporal information can be used to predict the following frames of 3D model by querying the time stamp larger than 1. To our knowledge, existing works of motion prediction Butepage et al. [2017], Ghosh et al. [2017], Mao et al. [2019] rely only on the skeleton without considering dense representation of human. So we believe that STIF-Nets can trigger new research direction in 4D shape modeling.

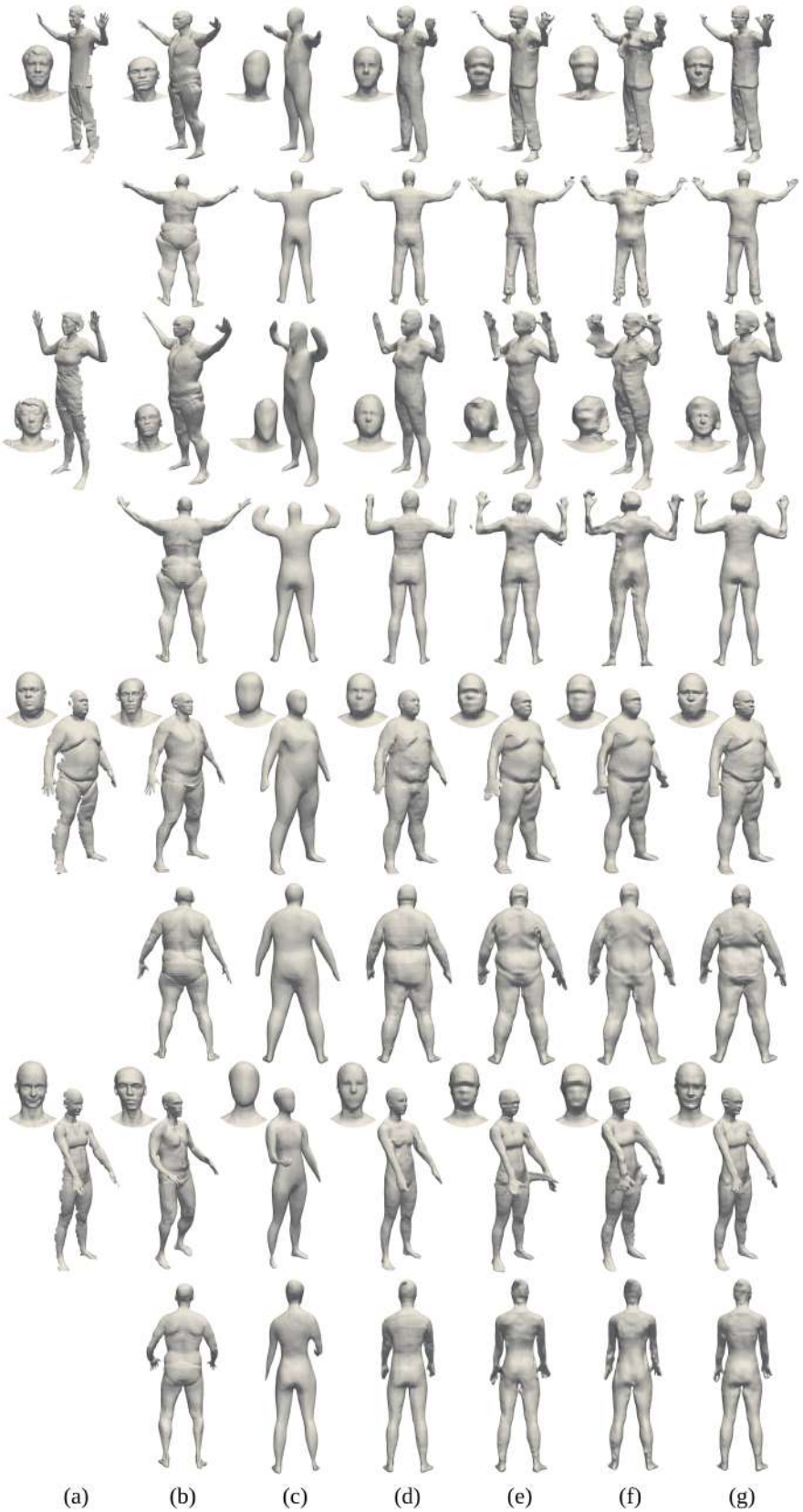


Figure 4.8: Qualitative results with front-view completions. From left to right, (a) partial scan, reconstruction of (b) 3D-CODED Groueix et al. [2018a], (c) OccNet Mescheder et al. [2019], (d) IF-Net Chibane et al. [2020a], (e) our static, (f) our naive dynamic and (g) our STIF-Nets.

Figure 4.9: Qualitative result for completion task. From left to right, (a) front-view partial scan, the reconstruction (b) our static method with Gaussian sampling, (c) our static method with our sampling, (d) the naive dynamic baseline and (e) our STIF-Nets.

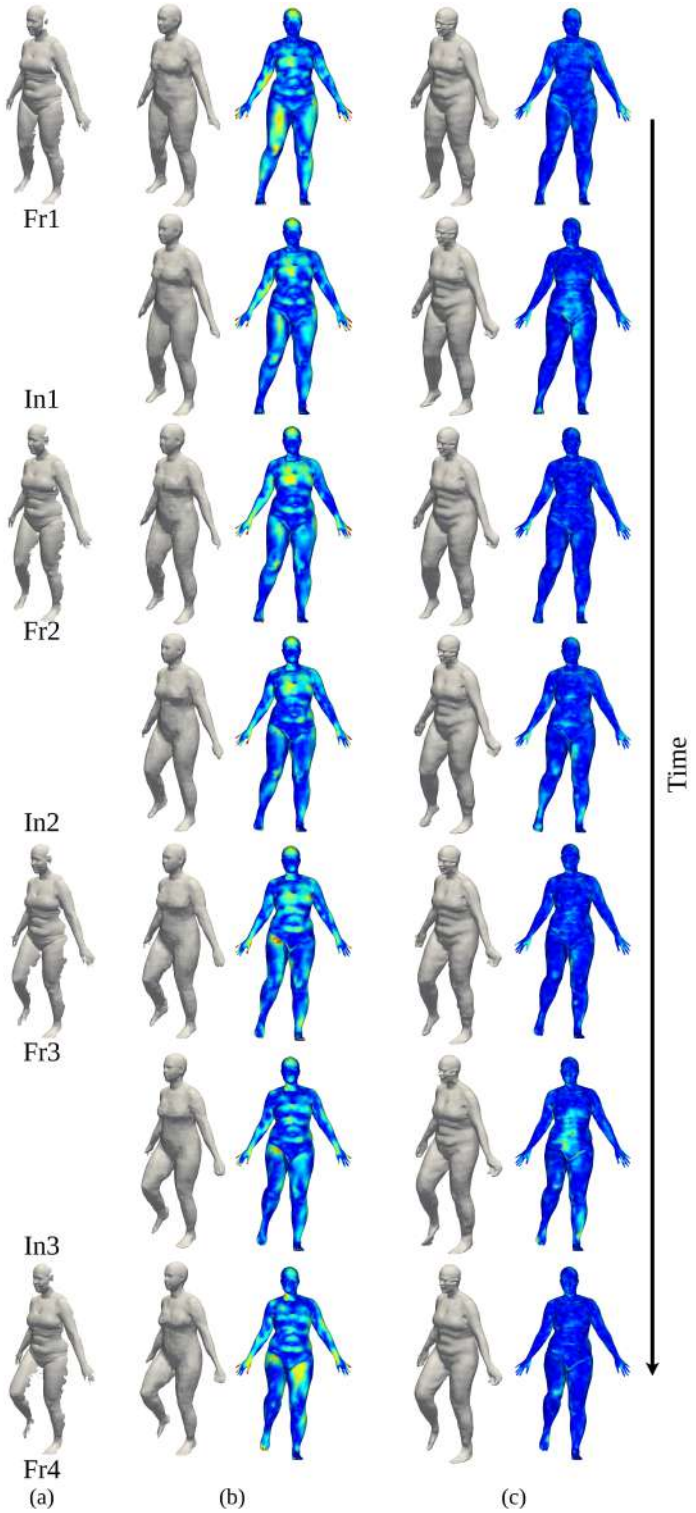


Figure 4.10: Qualitative results for a 4 frame input sequence with frame completion (Fr1, Fr2, Fr3, Fr4) and interpolation (In1, In2, In3). From left to right, (a) Partial scan, reconstruction/heatmap of (b) IF-Net Chibane et al. [2020a], and of (c) our STIF-Nets. For the heatmap, we compute the Chamfer-L1 distance from reconstruction to the ground truth and we set 0.03 as the maximum error.

Chapter 5

Pyramidal Signed Distance Learning for Spatio-Temporal Human Shape Completion

5.1 Introduction

In the previous chapter, we have investigated how does the spatial-temporal implicit modeling perform for the task of spatial shape completion and temporal frame enhancement. Completing shape models is a problem that arises in many applications where perception devices provide only partial shape observations. This is the case with single depth cameras which only perceive the front facing geometric information. Completing the 3D geometry in such a situation while retaining the observed geometric details is the problem we consider in this chapter. We particularly focus on human shapes and take benefit of the camera ability to provide series of depth images over time. Here we take the same input as in the Chapter 4, the network coarse-to-finely outputs the signed distance fields over time.

The task is challenging since different and potentially conflicting issues must be addressed. Strong priors on human shapes and their clothing are required as large shape parts are usually occluded in depth images. However, such priors can in practice conflict with the capacity to preserve geometric details present in the observations. Another issue lies in the ability to leverage information over time with local geometric patterns that can be either temporally consistent or time varying, as with folds on human clothing.

Existing methods fall into two main categories with respect to their global or local approach to the human shape completion problem. On the one hand, model-based methods build on parametric human models which can be fitted to incomplete observations, for instance SMPL Loper et al. [2015] Dyna Pons-Moll et al. [2015] or NPMs Palafox et al. [2021]. Approaches in this category can leverage strong priors on human shapes and hence often yield robust solutions to the shape completion problem. However, parametric models usually impose that the estimated shapes lie in a low dimensional space, which limits the ability to generalize over human shapes, in particular when considering clothing. Another consequence of the low dimensional shape space is that local geometric details tend to be filtered out when encoding shapes into that space.

On the other hand, neural network based approaches, *e.g.*NDF Chibane et al. [2020b], IF-Net Chibane et al. [2020a] or STIF Zhou et al. [2021], use implicit representations to model the observed surfaces. Such representations give access to a larger set of shapes, although this set is still bounded by the coverage in the training data. They also present better abilities to preserve geometric details as they are, by construction in these approaches, local representations. Besides, in contrast to model based methods, the local aspect of the representation impacts the robustness with estimated shapes that can often be spatially inconsistent.

In order to retain the benefit of local representations while providing better robustness and spatial coherence we propose a novel hierarchical coarse-to-fine strategy that builds on implicit neural representations Park et al. [2019], Mescheder et al. [2019]. This strategy applies to the spatial domain as well as to the temporal domain. We investigate the temporal dimension since redundancy over time can contribute to shape completion, as demonstrated in Zhou et al. [2021]. This dimension naturally integrates into the pyramidal spatio-temporal model we present. Our approach considers as input consecutive depth images of humans in a temporal sequence and estimates in turn a sequence of 3D distance functions that maps any point in the space time cube covered by the input frames to its distance to the observed human shape. This distance mapping function is learned over temporally coherent 3D mesh sequences and through a MLP decoder that is fed with multi-scale spatio-temporal features, as encoded from the input depth images. Our experiments demonstrate the spatial and temporal benefit of the hierarchical strategy we introduce with both local and global shape properties that are substantially better preserved with respect to the state of the art.

The main contributions are to show the benefit of a pyramidal architecture that uses a residual pyramid of features in the context of implicit surface regression and to propose a tailored training strategy that exploits this residual architecture by using losses that are distributed at each scale of the feature map both spatially and temporally.

5.2 Related Work

The set of methods proposed in the literature to estimate full human body surfaces from a sequence of partial observations has been reviewed in Section 2.4. Other than dividing the related work in spatial and spatial-temporal approaches as in Section 4.2, here we focus on the most related methods in four categories: model-based method, regression-based methods, hybrid methods and temporal modeling. We detail these different approaches and discuss their strength and limitations.

Model-based methods build on learned parametric human models that are fitted to incomplete observations. The parametric model generally represents a 3D undressed human with global shape and pose parameters. The surface can be represented explicitly using a triangulated surface similarly to SMPL Anguelov et al. [2005], Loper et al. [2015], Pavlakos et al. [2019], Xu et al. [2020] or Dyna Pons-Moll et al. [2015], or implicitly using either a parameterized occupancy or a signed distance function Palafox et al. [2021], Atzmon and Lipman [2020], Gropp et al. [2020], Atzmon and Lipman [2021]. Model based methods that decouple the body shape parameters from the pose like Loper et al. [2015], Palafox et al. [2021] can be used to efficiently exploit temporal coherence in a sequence by estimating a single shape parameter vector for the whole sequence and different pose for each frame. The low dimensionality of the parameterized space inherently limits the ability of these models to represent details such as clothing or hair present in the data. To model the details beyond the parameterized space Pons-Moll et al. [2017], von Marcard et al. [2018], Alldieck et al. [2018a,b], Bhatnagar et al. [2020b] add an estimate a cloth displacement map to represent 3D dressed humans, which significantly improves the accuracy of the reconstructed surface but often still impose some constraint on the topology of the surface and thus do not allow to fully explain complex observed data.

Regression-based methods directly predict the shapes from the incomplete data. This allows to predict less constrained surfaces and keep more of

the details from the input data than model-based methods. The representation of the predicted surface can be explicit Prokudin et al. [2019] or implicit which allows for changes in the topology Chen and Zhang [2019], Park et al. [2019]. This approach has been used for non-articulated objects Liu et al. [2019], Mescheder et al. [2019], Xu et al. [2019], Niemeyer et al. [2020], Peng et al. [2020], Chibane et al. [2020b] and articulated human body Saito et al. [2019, 2020], Chibane et al. [2020a], Huang et al. [2020], Deng et al. [2020b], Zhou et al. [2021], Prokudin et al. [2019]. The implicit surface can be represented using a (truncated) signed distance function, unsigned distance functions Chibane et al. [2020b] or an occupancy function as in IF-Net Chibane et al. [2020a]. Learning truncated signed distance functions tends to lead to more precise surface than learning occupancy as the signed distance provides a richer supervision signal for points that are sampled near but not on the surface. One main challenge with these approaches is to efficiently exploit the temporal coherence to extract surface detail from previous frames.

Hybrid methods that mix various of these approaches have been proposed to combine their advantages. IP-Net Bhatnagar et al. [2020a] generates the implicit surface with Chibane et al. [2020a] and registers the completed surface with SMPL+D Alldieck et al. [2019b], Lazova et al. [2019] and thus inherits the limited ability of model based method to represent geometric details. Function4D Yu et al. [2021] combines a classical dynamic fusion method with a post processing step to repair holes in the surface using a learned implicit surface regression. While these methods improve results w.r.t each of the combined methods, they anyway tend to retain part of their drawbacks.

Exploiting temporal coherence is a main challenge for regression based methods, as mentioned above. Using a pure feed-forward approach where a set of frames is fed into the regressor becomes quickly unmanageable as the size of the temporal window increases. Oflow Niemeyer et al. [2019] exploits the temporal coherence by estimating a single surface for the sequence initial frame which is deformed over all frames using a dense correspondence field conditioned on the whole sequence inputs. This strategy enforces by construction temporally consistency, but it also prevents temporal information to benefit to the shape model. STIF Zhou et al. [2021] address the problem of temporal integration using a recurrent GRU layer to aggregate information. Such layer can however only be used at the low-resolution features and hence surface details do not propagate from one frame to the next.

In this chapter, we suggest a pyramidal approach that can exploit both spatial and temporal dimensions. Taking inspiration from pyramidal strategies applied successfully to other prominent vision problems such as multi-view stereo Yang et al. [2020] and optical flow Sun et al. [2018] we devise a method that aggregates spatio-temporal information in a coarse-to-fine manner, propagating features from low to high resolution through up-sampling, concatenation with higher resolution features and the addition of residuals.

5.3 Network Architecture

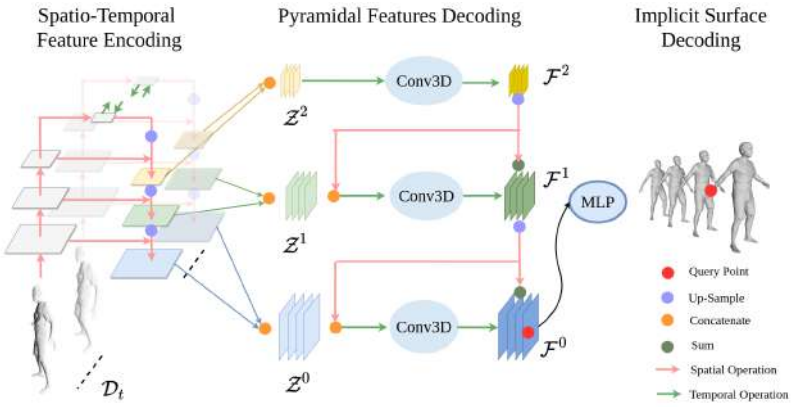


Figure 5.1: Architecture. Our network consists of 3 phases: the depth images are processed with Spatio-Temporal Feature encoding backbone in §5.3.1 into three-level-feature maps. These feature maps are latter aggregated with 3D convolution considering the temporal dimension in the coarse-to-fine manner in §5.3.2. The aggregated feature map is used to predict the signed distance field with MLP in §5.3.3.

We wish to compute the completion of a sequence of shapes observed from noisy depth maps. These maps, noted $\mathcal{D} = \{\mathcal{D}_t\}_{t\in\{1,\dots,n\}}$, provide a truncated front point cloud of the human shape in motion. Our output, by contrast, is a corresponding sequence of complete shapes encoded as a set of per-frame Truncated Signed Distance functions, using neural implicit functions $\text{SDF}_t(p)$ with $t \in \{1,\dots,n\}$, which can be each queried for arbitrary 3D points $p = (x,y,d)$. The surface can then be extracted using standard algorithms Lorensen and Cline [1987], Lewiner et al. [2003].

In the following, we explain how the $\text{SDF}_t(p)$ functions can be coded with a network inspired from point cloud SDF networks Chibane et al. [2020b], Park et al. [2019], but which also jointly leverages the analysis of small temporal

frame subgroups of size n. We also propose a more general pyramidal coarse-to-fine architecture to balance global information and local detail in the inference, which was previously only demonstrated with explicit TSDF encodings for a different incremental reconstruction problem Sun et al. [2021b]. Our network is articulated around three main phases represented in Fig. 5.1, a feature pyramid extraction phase (§5.3.1), a pyramidal feature decoding phase (§5.3.2), and an implicit surface decoding phase (§5.3.3), detailed as follows.

5.3.1 Spatio-Temporal Feature Encoding

To build a hierarchical inference structure, a time-tested pattern (*e.g.* Yang et al. [2020], Sun et al. [2018]) is to build a feature hierarchy or feature pyramid Lin et al. [2017] to extract various feature scale levels. The previous chapter shows that a measurable improvement to this strategy, in the context of temporal input images sequences, is to link the coarsest layer in each input image's pyramid with a bidirectional recurrent GRU component Cho et al. [2014], Chung et al. [2014], instead of building independent per-frame pyramids. This allows the model to learn common global characteristics and their mutual updates at a modest computational cost. We modify the backbone encoder U-GRU Zhou et al. [2021], which for each input depth map $\mathcal{D}_t$ with $t \in \{1,\dots,\mathrm{n}\}$, yields a set of per-frame feature maps $\mathcal{Z}_t^0$, $\mathcal{Z}_t^1$, $\mathcal{Z}_t^2$ at three levels of respectively high, mid and low resolutions. In the rest of the discussion we will only use temporal aggregates of each feature level $\mathcal{Z}^0 = \{\mathcal{Z}_t^0\}, \mathcal{Z}^1 = \{\mathcal{Z}_t^1\}$, and $\mathcal{Z}^2 = \{\mathcal{Z}_t^2\}$ with $t \in \{1,\dots,\mathrm{n}\}$, such that we can denote their extraction:

$$\{\mathcal{Z}^0, \mathcal{Z}^1, \mathcal{Z}^2\} = \text{U-GRU}(\mathcal{D}_1, \dots, \mathcal{D}_\mathrm{n}), \tag{5.1}$$

as illustrated in Fig. 5.1. The feature map here contains 3 dimensions: x, y and t in the three different levels of the pyramid. However, the coarsening is only done in the spatial domain at this stage so that all scale feature maps contain the same number of frames n. More details on these feature map's dimensions are provided in section 5.5.

5.3.2 Pyramidal Feature Decoding

In the next stage, our goal is to allow the network to progressively decode and refine spatio-temporal features of the sequence that will be used for TSDF decoding. To this goal, we first process the coarsest feature aggregate $\mathcal{Z}^2$ with a full 3D (2D + t) convolution to extract a sequence feature $\mathcal{F}^2$:

$$\mathcal{F}^2 = \text{Conv3D}(\mathcal{Z}^2). \tag{5.2}$$

We then subsequently process the finer feature levels $l \in \{1,0\}$, first by upsampling the previous-level sequence feature map in the spatial domain using bilinear interpolation, noted $\mathcal{U}^l$, then concatenating it with the aggregate feature $\mathcal{Z}_l$ as input to 3D convolutions, as shown in Fig. 5.1. The output of the 3D convolutions is the residual correction of the previous-level aggregated feature, echoing successful coarse-to-fine architectures Yang et al. [2020], Sun et al. [2018]:

$$\mathcal{U}^l = \text{Up-Sample}(\mathcal{F}^{l+1}), \tag{5.3}$$

$$\mathcal{F}^l = \mathcal{U}^l + \text{Conv3D}(\{\mathcal{U}^l, \mathcal{Z}^l\}). \tag{5.4}$$

5.3.3 Implicit Surface Decoding

The final high-resolution feature $\mathcal{F}^0$ obtained as output of the previous process serves as a latent vector map, from which we sample a latent vector at a given query point's image coordinates. Similar to §4.3.3, the selected latent vector is then decoded using an MLP, and provides the TSDF value for that point. More formally, given a query point (x,y,d) and a time frame $t \in \{1,\ldots,\text{n}\}$ this signed distance function is computed by first sampling bi-linearly at the corresponding 2D location (x,y) the slice $\mathcal{F}_t^0$ of the high resolution 3D feature map that corresponds to frame t of the temporal window. We can note this sampler f_t^0:

$$f_t^0(x,y) = B(\mathcal{F}_t^0; x, y), \tag{5.5}$$

After obtaining the queried feature vector in the high resolution scale, we concatenate it with depth feature z. The signed distance function can then be computed by

$$\text{SDF}_t^0(x,y,d) = \text{MLP}(f_t^0(x,y), d), \tag{5.6}$$

where an MLP is used to decode the TSDF value for a given pixel x, y, and frame t of the high-resolution feature map, and an implicitly accounted d. Operational details and constants are discussed in the next section.

5.4 Training Strategies

In the previous section, we discussed the network used and inference path for a given set of n input frames. The same path can be used for supervised training, but pyramidal architectures are usually trained with intermediate loss objectives that guide the coarser levels. This is easy to do with a supervised loss when the coarser objectives are just a down-sampled version of the expected result, *e.g.*for object detection, depth map or optical flow inference Yang et al. [2020], Sun et al. [2018]. A main contribution here is to show the benefit of such schemes with an implicit representation of the surface reconstructed from a residual pyramid of features. One of the key difficulties lies in characterizing the intermediate, lower resolution contribution to the final result of coarser layers of the implicit decoder. To this goal we present four training strategies, with several temporal and pyramidal combinations, and discuss their ablation in the experiments. A visual summary of these strategies is provided in Fig. 5.2.

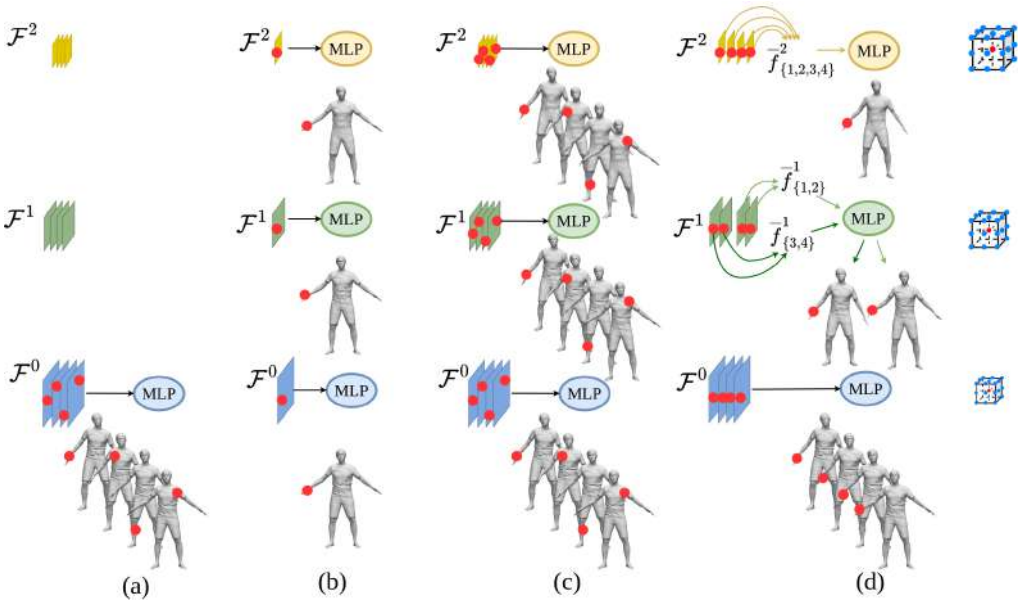


Figure 5.2: Training strategy variants, from left to right: (a) temporal naive, (b) static pyramidal, (c) temporal pyramidal and (d) spatio-temporal pyramidal. More detailed explanations can be found in §5.4.3.

5.4.1 Common Training Principles

We render depth maps of 3D raw scans from a training set to associate depth maps observations to ground truth computed SDF values for a given set of query points in the 3D observation space. Following general supervision principles of implicit networks Saito et al. [2019], Zhou et al. [2021], we provide training examples for three groups: on-surface points, points in the vicinity of the surface, and other examples uniformly sampled over the whole observed 3D volume. On-surface training examples are obtained simply by sampling n_s vertices from ground truth 3D model and associating them with SDF value 0. For surface vicinity points at any given time t, as shown in Fig. 5.2, for each of the n_s vertices, we randomly draw 4 training points with their ground truth SDF values among points on a 26-point 3D grid centered on a ground truth surface point, in which the closest projected grid edge length matches the observed feature map pixel size, as illustrated in the right most column in Figure5.2. For the third, uniform point group, we randomly draw n_u sample training points covering the whole acquisition space, for which we provide the SDF to the ground truth surface. In our experiments, we keep $n_s = 400$ and $n_u = 800$ constant regardless of training scenario.

5.4.2 Pyramidal Training Framework

For explicit pyramidal training, we need to supply ground truth SDFs for all levels of the pyramid, including the coarser ones. To this goal, for a set of sampled ground truth surface points at time t, we extend the set of surface-vicinity training samples by randomly drawing them from three different 26-point-grid cubes centered on the ground truth surface point, instead of only one. Each of these grid cubes corresponds to a network pyramid level feature $\mathcal{F}_t^l$ with $l \in \{0,1,2\}$ and its closest projected grid edge length matches the finest feature map pixel size $\mathcal{F}_t^0$. We still draw 4 samples at each feature level l, among the 26 possibilities.

For each of the m query points $p_i = (x_i, y_i, d_i)$ in the training set so constructed among the three training groups, noting its frame t_i and pyramid level l, we generalize equations 5.5 and 5.6 to provide an MLP decoding and training path of SDFs at each possible resolution level and frame in the temporal

window:

$$f^l_{t_i}(x_i, y_i) = B(\mathcal{F}^l_{t_i}; x_i, y_i) \tag{5.7}$$

$$\text{SDF}^l_{t_i}(x_i, y_i, d_i) = \text{MLP}(f^l_{t_i}(x_i, y_i), d_i) \tag{5.8}$$

Given the ground truth SDF value s^l_i computed from ground truth 3D models for every training point p_i for level l, the loss can then be defined as:

$$L = \sum_l \lambda^l \frac{1}{m^l} \sum_{i=1}^{m^l} \|\text{SDF}^l_{t_i}(x_i, y_i, d_i) - s^l_i\|^2. \tag{5.9}$$

where λ is the weight to balance the final loss, in practice we set $\lambda^0 = 4$ for the finest level, $\lambda^1 = 1$ and $\lambda^2 = 0.1$ for the coarsest level.

5.4.3 Simpler Variants of the Pyramidal Approach

Based on the previous framework, various training variants can then be devised by ablating the 2D+t convolutional pyramid training or the supervision of all pyramid levels versus only limiting the loss to the finer level, as shown in Fig. 5.2.

- **Static pyramidal variant.** We use a simple training variant derived from this loss, by considering single one input frame $K = 1$ and consequently replacing 3D (2D+t) convolutions by 2D convolutions, in effect keeping the aggregation from each single feature map $\mathcal{Z}^0$, $\mathcal{Z}^1$, and $\mathcal{Z}^2$ decoded from the input feature pyramid to the corresponding $\mathcal{F}^0$, $\mathcal{F}^1$, and $\mathcal{F}^2$. The pyramid is trained including training samples for all pyramid levels $\mathcal{F}^l$ in the loss L from (5.9). The static pyramidal baseline is illustrated in Fig. 5.2(b).
- The **temporal naive variant** keeps the (2D+t) convolutions for a temporal frame group $t_i \in \{1, \ldots, K\}$ but only supervises the finest pyramid level.
- The **temporal pyramidal variant** includes the loss terms for all 3 pyramid levels, and uses unmatched, untracked randomized sample training points in each frame of the processed temporal frame group. The temporal naive and temporal pyramidal variants are illustrated respectively in Fig. 5.2(a) and (c).

5.4.4 Full Spatio-Temporal Pyramidal Training

While the previous pyramidal variant account for spatial hierarchical aspects in training, no component in the previous training schemes accounts for hierarchical temporal aggregation or weak surface motion priors in the training losses. One would expect such terms to help the training balance global spatio-temporal aspects against local spatio-temporal details for the underlying surface in motion.

Here we propose a temporal coarse-to-fine spatio-temporal training supervision for $\mathrm{n} = 4$ time frames, extending the previous temporal pyramidal variant. In this training strategy we add gradual pooling of the queried features from feature maps $\mathcal{F}_t^l$ to provide 4 supervision signals at the finer level with map $\mathcal{F}^0$, 2 supervision signals at the mid-level $\mathcal{F}^1$ from frame groups 1, 2 and 3, 4, and one signal global to the sequence for the coarsest level $\mathcal{F}^2$, averaging the feature over all 4 frames. At training time, for a point $p = (x, y, d)$ among the uniform group of training points, this can be done simply by retrieving the four features in the four temporal maps corresponding to point p in the coarse map $\mathcal{F}^2$ and the two averages by retrieving the two features from temporal frames 1, 2 and 3, 4 respectively. To perform this in a meaningful way for points on and at the vicinity of the surface however, the features pooled temporally should be aggregated from a coherent point trajectory in the temporal sequence. For this we leverage training datasets for which a temporal template fitting is provided (*e.g.*with SMPL Loper et al. [2015]) to work with temporally registered vertices. Intuitively, this allows the proposed network to account for a weak prior on underlying point trajectory coherence when producing estimated SDF sequences at inference time.

To formalize this sampling strategy we can first denote v_i^t the 3D position of the i^{th} vertex in the temporal registered model in frame t. Then, similarly to what is done in section 5.4.2, each of the query points p_i^1 generated in the vicinity to the surface in the first frame is obtained by choosing a point on a 26-point 3D grid centered around a vertex of index v_i^1 on the fitted ground truth model in the first frame. However instead of sampling the query point independently for each of the 4 frames in the temporal window, we reuse the same vertex indices and the same offset w.r.t to that vertex throughout the 4 frames to obtain query points $(p_i^t)_{t=1}^4$ along a trajectory that follows the body motion. i.e. $p_i^t = v_i^t + p_i^1 - v_i^1$. We can then formalized the averaging of the features extracted across several

frames in a set S using the point trajectories $(p_i^t)_{t\in S}$ as:

$$\overline{f}_S^l = \frac{1}{|S|}\sum_{t\in S} f_t^l(x_i^t, y_i^t), \tag{5.10}$$

with $p_i^t = (x_i^t, y_i^t, d_i^t)$. As shown in Figure 5.2(d), for each point sample we use features $\overline{f}^2_{\{1,2,3,4\}}$ to compute a low-level loss and the features $\overline{f}^1_{\{1,2\}}$ and $\overline{f}^1_{\{3,4\}}$ to compute two mid-level losses. For each of these pooled features, we use the mean of the point's ground truth SDFs as supervision signal after MLP decoding.

5.5 Implementation Details

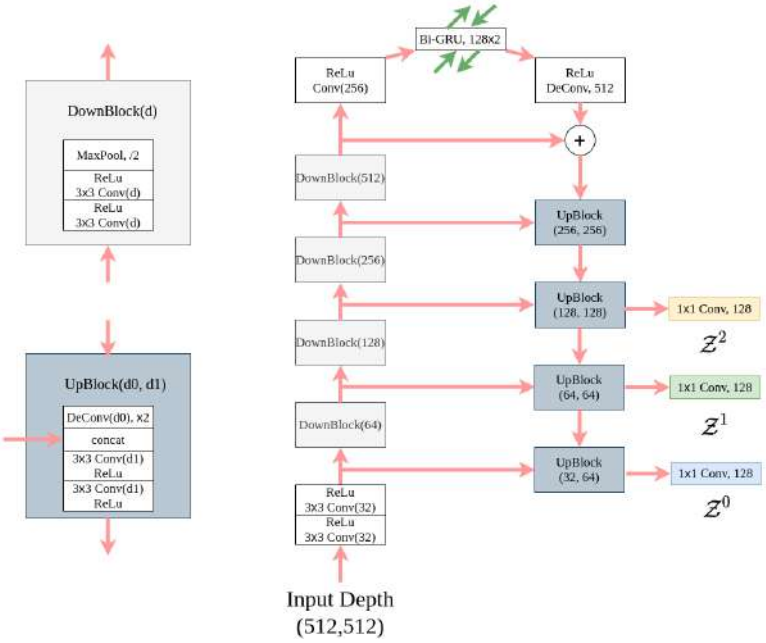


Figure 5.3: Spatio-Temporal Feature Encoding.

As illustrated in Figure 5.3, the Spatio-Temporal Feature Encoding consists of 4 down blocks to the global information, and 4 up blocks. We adapt U-Net Ronneberger et al. [2015]-like encoder Zhou et al. [2021] as backbone, for the purpose of hierarchical learning, combined with bidirectional GRU Cho et al. [2014], Chung et al. [2014] for pyramidal feature extraction in §5.3.1. In this architecture, the 3 last up blocks are associated with 1×1 convolution to output the features $\mathcal{Z}^l, l = \{0, 1, 2\}$ at different levels. Figure 5.4(a) shows how

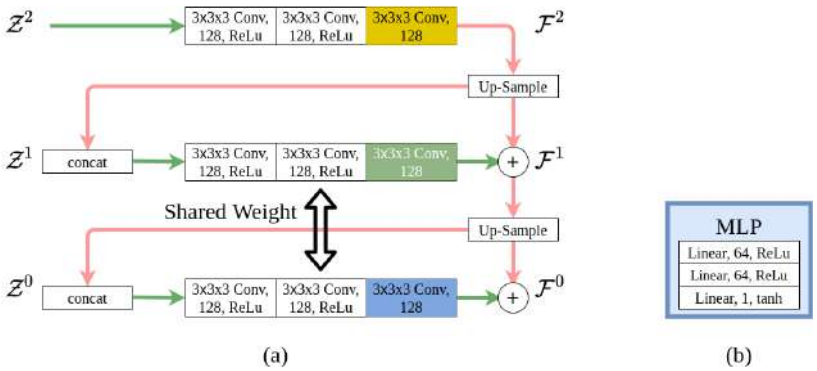


Figure 5.4: Detailed Architecture of (a)Pyramidal Feature Decoding and (b) Implicit Surface Decoding.

the features $\mathcal{Z}^l$ are fed into three 3D convolution blocks in §5.3.2. Each 3D convolution block contains 3 convolutions layers with $3 \times 3 \times 3$ kernels. We use zero padding spatially and temporally and the weight of these 3D convolutions are shared for $l = 0$ and 1. Figure 5.4(b) illustrates the Implicit Surface Decoding in §5.3.3. The layer dimension of the MLP is 64, 64 and 1 respectively. The depth feature z is multiplied by 128 in order to get a similar activation ranges as the values in the feature vector $f_t^0(x,y)$. The final output layer of the MLP is activated with *tanh* in order to bound the signed distance in the range $[-1,1]$. The pre-computed SDF is scaled with the factor of 75 and we set the truncate value as 1 in the normalized space. In practice, this multiplication improves also the numerical stability, otherwise the loss would be too small during the training. While it would be possible to compute the inference results over an input sequence in a sliding window fashion, for computational trade-off with quality we compute the results on consecutive groups of $\mathrm{n} = 4$ frames. A training batch is composed of 4×512^2 depth images and 4×2800 query points per scale for the SDF evaluation. It takes 2.86 seconds with GPU memory footprint about 7.86GB for a batch training.

5.6 Experiments

We provide quantitative and qualitative comparisons on real-world scan data. In particular, we discuss the results of the different training strategies introduced in §5.4. We also show numerical comparisons, on raw scan data, between our method and both learning-based method and model-based methods.

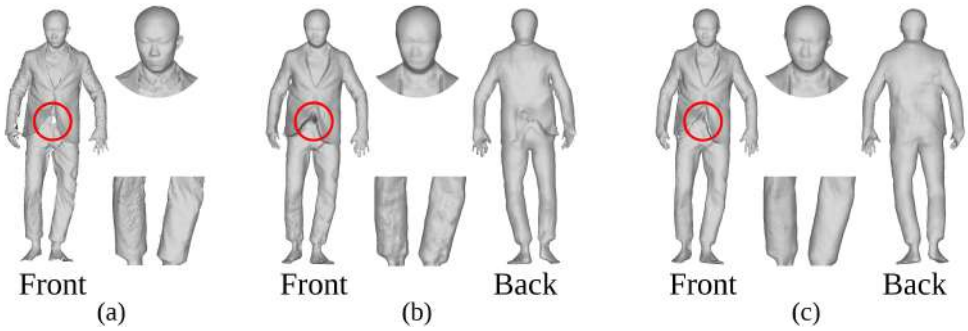


Figure 5.5: Shape completion on CAPE(left) and DFAUST(right). (a) front-view partial scan. Completion with: (b) our static pyramidal method, (c) our spatio-temporal pyramidal method.

5.6.1 Dataset

Focusing on dynamic human shape completion, we collect data from the CAPE Ma et al. [2020] dataset for dressed humans with different clothing styles for characters, and the DFAUST Bogo et al. [2017] dataset for undressed humans, as in competitive methods Palafox et al. [2021], Zhou et al. [2021]. Both are captured at 60fps and provide temporally coherent mesh sequences alongside the raw scan data. We render depth images in resolution 512^2 from the raw scans in order to preserve the measurement noise. We also pre-compute pseudo ground-truth signed distances using the mesh models to avoid topological artifacts present in the raw data. Meshes in these datasets are actually obtained by fitting the SMPL Loper et al. [2015] model to the raw scan data. However, thanks to the local reasoning, our network tries to capture the partwise space spanned by the local geometric patterns from the fitted models and is agnostic to the associated shapewise parametric space, here SMPL.

5.6.2 Training Protocol

We follow the training protocol of Zhou et al. [2021]. We take 2 male and 2 female characters from each dataset, in total 8 characters. Each character performs 3 or 4 motion styles, in total there are 28 motion sequences. 6 subsequences with 4 frames are extracted in each motion sequence. In addition, the 4-frame sub-sequence are in 2 styles: short-term and long-term. The short-term input sub-sequence covers about 200ms duration while the long-term one covers about 500ms duration. Note that we train only one model for the characters in the two datasets and short-term as well as long-term sub-sequences.

Data	CAPE		DFAUST	
Method	IoU ↑	Chamfer-L1 ↓	IoU ↑	Chamfer-L1 ↓
(i) naive temporal (a)	0.777	0.174	0.858	0.115
(ii) static pyramidal (b)	0.788	0.172	0.871	0.112
(iii) temporal pyramidal (c)	0.808	0.182	0.873	0.118
(iv) spatio-temporal pyramidal (d)	**0.839**	**0.161**	**0.898**	**0.103**
(v) spatio-temporal occupancy (e)	0.800	0.168	0.872	0.109

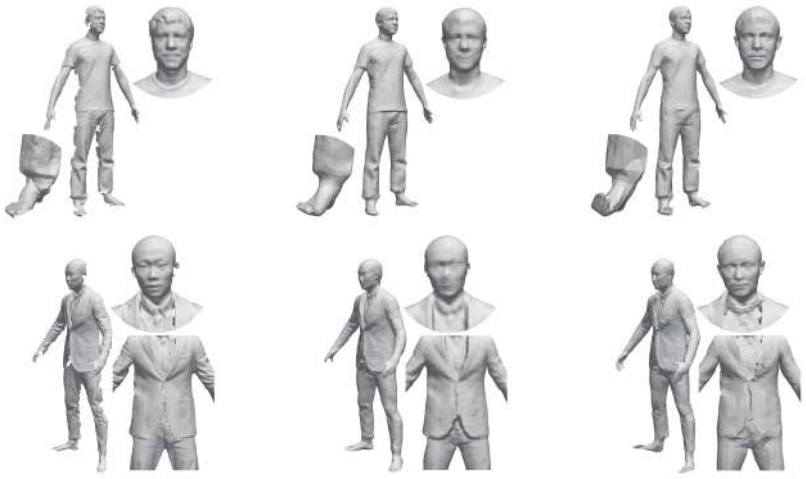
Table 5.1: Pyramidal Approach Variants. Spatial completion with IoU and Chamfer-L1 distances ($\times 10^{-1}$) in the real 3D space.

5.6.3 Ablation and Variants Comparison

We validate our method by testing the different variants of §5.4 on 152 frames from new released CAPE data and 100 frames of two unseen identities from DFAUST (see Table 5.1). These results show that the spatio-temporal pyramidal training strategy in §5.4.4 and illustrated in Figure 5.2(d) yields the best results. In Table 5.1, from row(i) to row(iii), the pyramidal training is more effective than the naive training. In row(iv), we leverage the coarse-to-fine strategy not only spatially but also temporally and obtain the improvement from row(ii) to row(iv), especially for the more difficult case of clothed humans with the CAPE dataset. By comparing row(iv) and (v), it shows the effectiveness of our neural signed distance representation with respect to occupancy. In order to better illustrate the contribution of the temporal information, Figure 5.5 show how the spatio-temporal pyramidal model can correct artefacts still present with the static pyramidal only.

5.6.4 Local Pattern Reasoning

As mentioned in Section 5.6.1, we render depth images from the raw scans, *e.g.* Figure 5.6(a), and preserve therefore measurement noise. The signed distances are precomputed from watertight meshes, as obtained from the full scans and SMPL Loper et al. [2015], see Figure 5.6(c). Although the SMPL fitting can be locally imperfect, as a result of the global fitting, our network learns locally and naturally tends to reproduce the input depth information as optimal predictions on average over all parts in all the training models. This can be observed in Figure 5.6 where the network better predicts the foot than the SMPL fitting on the full scan.



(a) Scan **(b)** Ours **(c)** SMPL

Figure 5.6: Local pattern reasoning. From left to right, (a) partial scan, (b) completion with our method and (c) watertight pseudo ground truth with SMPL Loper et al. [2015] on full scans.

5.6.5 Learning-based Method Comparisons

Baselines We first compare our method with learning-based baselines in the following categories: 1. *Static methods* as 3D-CODED Groueix et al. [2018a], ONet Mescheder et al. [2019], BPS Prokudin et al. [2019], IF-Net Chibane et al. [2020a] and 2. *Dynamic methods* as OFlow Niemeyer et al. [2019] and STIF Zhou et al. [2021]. 3D-CODED and BPS are point-based methods which take point cloud as input and output the template-aligned mesh. ONet, IF-Net, OFlow and STIF-Net are implicit function learning methods as ours. IF-Net relies on the 3D convolution so it pre-processes the partial observation input into voxel grid. And STIF-Net inputs the depth image as ours. To fairly compare with IF-Net and STIF-Net, we use the same resolution of input. The final mesh surface is extracted with marching cubes Lewiner et al. [2003], Lorensen and

Data	CAPE		DFAUST	
Method	IoU ↑	Chamfer-L1 ↓	IoU ↑	Chamfer-L1 ↓
3D-CODED Groueix et al. [2018a]	0.455	0.591	0.578	0.347
ONet Mescheder et al. [2019]	0.488	0.476	0.604	0.340
IF-Net Chibane et al. [2020a]	0.853	0.121	0.876	0.107
BPS Prokudin et al. [2019]	-	-	0.761	0.197
OFlow Niemeyer et al. [2019]	-	-	0.740	0.231
STIF Zhou et al. [2021]	0.834	0.113	0.865	0.104
ours	**0.880**	**0.100**	**0.914**	**0.092**

Table 5.2: Comparison with learning-based methods with IoU and Chamfer-L1 distances ($\times 10^{-1}$) in the real 3D space on Zhou et al. [2021] benchmarck.

Figure 5.7: Static Shape completion, top two identities are from CAPE and bottom two are from DFAUST. From left to right, (a) front-view partial scan, completion with (b) 3D-CODED Groueix et al. [2018a], (c) IF-Net Chibane et al. [2020a] and (d) our method. We set the maximum reconstruction to ground truth Chamfer-L1 distance as 2cm for heatmaps.

Cline [1987].

Results We follow the evaluation protocol of STIF-Net Zhou et al. [2021], evaluating on 9 characters, including 2 unseen ones w.r.t. training data, with Intersection over Union(IoU) and Chamfer-L1 distance. In Tab 5.2, we improve both IoU and Chamfer-L1 distance on both datasets. We would like to highlight the benefit from 2 aspects: the dynamic modeling and the pyramidal learning

Method	IoU ↑	Chamfer-L2 ↓
OpenPose+SMPL	0.68	0. 243
IP-Net Bhatnagar et al. [2020a]	0.82	0. 034
NPMs Palafox et al. [2021]	0.83	**0.022**
ours	**0.89**	0.029

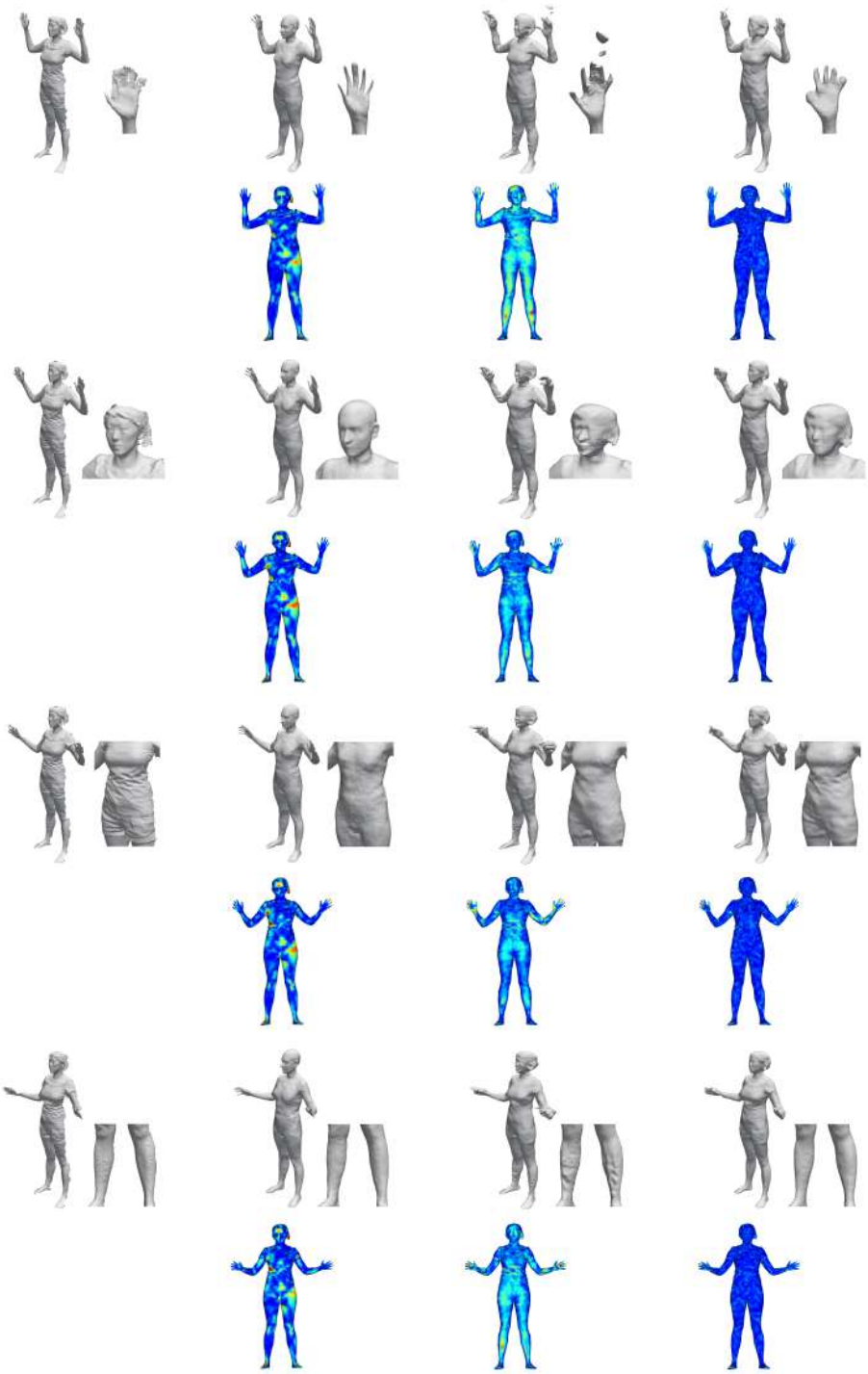
Table 5.3: Comparison with model-based methods with IoU and Chamfer-L2 distances ($\times 10^{-3}$) for CAPE data in the normalized space on Palafox et al. [2021] benchmark. Reuse the table of Palafox et al. [2021].

strategy. In Figure 5.7, our method largely improves the qualitative results from the static methods, especially for detail preservation. Our pyramidal learning strategy still retains more detail than dynamic state-of-the-art Zhou et al. [2021], the wrinkle, face details in Figure 5.8 and 5.9 and it could correct some artifacts, *e.g.*hand, leg in Figure 5.8 and 5.9.

5.6.6 Model-based Method Comparisons

Baselines We compare with model-based methods OpenPose Cao et al. [2019] + SMPL Loper et al. [2015], IP-Net Bhatnagar et al. [2020a] and NPMs Palafox et al. [2021]. Such methods require latent space optimization process. For more details, IP-Net extracts the surface from partial observation input with learning-based method Chibane et al. [2020a], then optimizes the SMPL+D Alldieck et al. [2019b], Lazova et al. [2019] model parameters to fit the surface. NPMs learns the neural parameter model from real-world data and synthetic data and optimize such parameters to fit the partial observation during inference. Note that IP-Net Bhatnagar et al. [2020a] and NPMs Palafox et al. [2021] train on 45000 frames from not only CAPE, but also from the synthetic data DeformingThings4D Li et al. [2021] and motion capture data AMASS Mahmood et al. [2019], while we train on 1344 frames from the real scan data of CAPE and DFAUST only.

Results We follow the evaluation protocol of NPMs Palafox et al. [2021] on 4 identities in 2 cloth styles with IoU and Chamfer-L2 distance. In Table 5.3, our method improves the numeric result on IoU but not on the Chamfer-L2 loss. One may note that since our method is devoid of a parametric or latent control space, it has more freedom to reconstruct details, see Figure 5.8. NPMs improves over a first inference phase by using a latent-space optimization which specifically minimizes for an L2 surface loss. Thus, NPMs is slightly better than ours on Chamfer-L2 metric but increases the inference time. While lacking such a stage, the results produced by our method are still observed to be generally competitive with the latter optimization methods.



(a) scan **(b)** NPMs **(c)** STIF **(d)** Ours

Figure 5.8: Motion completion on CAPE dataset. From left to right, (a) front-view partial scan, completion with (b) NPMs Palafox et al. [2021], (c) STIF Zhou et al. [2021] and (d) our method. We set the maximum reconstruction to ground truth Chamfer-L1 distance as 2cm for heatmaps.

(a) scan **(b)** OFlow **(c)** STIF **(d)** Ours

Figure 5.9: Motion completion on DFAUST dataset. From left to right, (a) front-view partial scan, completion with (b) OFlow Niemeyer et al. [2019], (c) STIF Zhou et al. [2021] and (d) our method. We set the maximum reconstruction to ground truth Chamfer-L1 distance as 2cm for heatmaps.

duration	100ms	200ms	300ms	400ms	500ms
STIF Zhou et al. [2021]	0.857/0.107	0.852/0.108	0.855/0.109	0.852/0.111	0.850/0.111
Ours	**0.900/0.095**	**0.900/0.095**	**0.900/0.096**	**0.899/0.097**	**0.898/0.097**

Table 5.4: Impact of sequence duration using the test data. Metrics are IoU/Chamfer-L1 ($\times 10^{-1}$)

5.7 Impact of Sequence Duration

During the training, we use both the short and long term intervals, resulting in sequences of 200ms and 500ms respectively. To give more insights on the impact of the sequence duration, we provide results with the same trained network but with different frame interval values for testing on the STIF Zhou et al. [2021] benchmark, in Table 5.4. The comparison with STIF shows that our architecture improves for all frame intervals and is more robust for long-term sequence completion.

5.8 Conclusion

In this chapter, we have proposed a complete framework for coarse-to-fine treatment and training of implicit depth completion from partial depth maps. Existing methods that solve this problem can provide robustness, with for instance model-based strategies that rely on parametric human models, or precision with learning approaches that can capture local geometric patterns using implicit neural representations. We investigated how to combine both properties with a novel pyramidal spatio-temporal learning model. This model exploits neural signed distance fields in a coarse-to-fine manner, this in order to benefit from the ability of implicit neural representations to preserve local geometry details while enforcing more global spatial consistency for the estimated shapes through features at coarser levels. In addition, our model also leverages temporal redundancy with spatio-temporal features that integrate information over neighboring frames. We provided a discussion and analysis of various training strategies. We demonstrated the substantial quality improvement that can be obtained by using an architecture that uses a residual pyramid of features in the context of implicit surface regression when combined with a tailored training strategy distributed losses each scale of the feature map both spatial and temporally.

Compared to the parametric model-based methods, the proposed method is flexible to recover the local details while the model-based approaches lacks the

representation capability due to the pre-defined model topology. Compared to the learning-based methods, our method outperform the state-of-the-art methods on human shape completion task thanks to our pyramidal neural signed distance learning strategy. We confirm the contribution of both the coarse-to-fine and temporal aggregation strategy through ablation studies.

Since the proposed method focuses on the local feature aggregation instead of global consistency, it is limited to handle the severe occlusion. Classic topology analysis techniques Chazal et al. [2016], Oudot [2017] are useful to handle the topology change, but the computation of persistence diagram for 3D object is expensive. We could add some topological loss as the regularization term Shit et al. [2021] to preserve the topological consistency in the future work. All in all, we believe that the proposed implicit pyramid schemes can be applied in a larger context for other problems in the realm of 3D vision and reconstruction.

Chapter 6

Conclusion

In this thesis, we have presented novel approaches profiting from large-scale human datasets to generate the human body geometry corresponding to the input signals, including point cloud and sequential depth images. We investigated the following aspects: (1) how to add the shape prior into neural network which could improve the performance and the robustness; (2) how to represent continuously the sequential input in both spatial and temporal directions; (3) how to design a pyramidal training strategy to recover fine details from different level representations. All proposed approaches achieve the state-of-the-art performance and trigger the new research direction in 3D human modeling.

6.1 Summary

We examine the contribution and drawback of this thesis as follows:

Chapter 3 presented a novel encoder-decoder network to tackle the problem of reconstructing human body mesh from sparse point clouds, for dense human correspondence. We carefully studied the failure case of the state-of-the-art methods Groueix et al. [2018a], Deprelle et al. [2019] in FAUST Bogo et al. [2014] challenge and came up with constraining neural network with predefined Gaussian Process layer as shape prior. Furthermore, the Gaussian Process layer enables to smoothen the reconstructed surface with respect to the fully connected layers. In addition, we proposed to train the network with adversarial loss which allowed to improve the generalization across datasets and the reconstruction fidelity.

Similar to 3D-CODED Groueix et al. [2018a], to handle the ambiguity of unstructured point cloud input, the pre-processing is applied for GP network, *e.g.* the randomly line-search the translation and rotation, which consumes the

time. The PointNet-like Qi et al. [2017] encoder only extracts the global feature from point cloud which is not able to preserve the local details on the human body and is limited to represent unseen pose. Once the point cloud input contains large missing part, the GP network fails to give the plausible result. Thus, we turn to the depth-like input in Chapter 4 and 5 for the completion task.

Chapter 4 contributed a novel spatio-temporal completion approach for human motion, tackling the problem of continuously reconstructing the sequential input from partial observation. We proposed a three-phase process to decompose the whole task according to the factors, (i) encoding a sequence of 2D depth image and linking the global feature through all frames; (ii) given the continuous query time stamp, interpolating the time-specific 2D feature map from the full set of encoded features; (iii) given the query point, decoding the occupancy value by using the local feature on the interpolated feature map along with the depth information. The proposed method outperforms the per-frame approaches, resulting the artefacts due to the noise or hole from raw input, for shape completion task and shows the advantage for enhancing the sequence densification by sampling the intermediate time stamp on spatio-temporal features.

The connection between global features from full set of frames was considered, but the local feature aggregation across frames is also worth studying. For specific body part, the corresponding part in other frames could typically contribute more to the detail recovery than the global information. The proposed method is also restricted to extract the high-frequency details due to the lack of representation ability of occupancy supervision.

Finally, Chapter 5 proposed the pyramidal strategy for learning signed distance in the coarse-to-fine manner. We started with the encoder architecture of Chapter 4 to prepare the pyramidal feature maps in three levels. Each level containing a set of frame features, would be refined with 3D convolution. The refined features of previous level was combined with the pre-encoded features as the initialization for the following refinement 3D convolution until the third level. In the setup of four-frame sequence, the supervision was added to the network progressively level by level, *i.e.* the coarse level supervision was averaging the ground truth of the corresponding points across four frames, the middle level was averaging the two first frames and two last frames and the full supervision was provided for the finest level. We demonstrated how the training strategy enables to learn the rich latent space from sequential input and recover more detailed results with respect to that in Chapter 4.

In this work, the correspondence across input frames is required during training, which is an additional information. Furthermore, due to memory cost, the proposed method can not be extended to long-time sequence, same case for the method of Chapter 4. The implicit neural modeling relies heavily on the local information sampled by the query point, which lacks the global constraint. In other words, the human topology can not be guaranteed when the extreme case occurs, *e.g.* the severe occlusion.

We will discuss the future work related to the solution to the limitations and some new directions in the next section.

6.2 Future Work

Temporal modeling the human shape is hard problem in terms of the long-term memory, pose/shape variation and so on. The proposed methods in Chapter 4 and 5 have only solved part of them. We will discuss the potential directions.

The aforementioned approaches were built on open source datasets captured with high quality scan system. The small scalability of the capture system limits the variety of pose and the range of motion. Moreover, the dataset here is very "clean", *i.e.* human is well semantically segmented without the occlusion from other objects in the capture space. In the case of flexible setup, some recovery system from RGB data is well studied in Saito et al. [2019, 2020], but the reconstruction from single view RGB signal can not guarantee the geometry consistency. Multi-stereo system Leroy et al. [2017] enables to capture the high fidelity reconstruction with the tolerance of the calibration and synchronization of multiple cameras. Currently, the fusion-like approaches together with neural network Yu et al. [2021], Zheng et al. [2021b] achieve the high quality result with very sparse RGB-D sensors. Could we achieve the coherent results with a less constrained setup, for example a moving mobile phone? More and more mobile phones can record depth information in the specific camera mode in order to balance the foreground and background effect on photo. It could be an interesting direction to explore the neural network for the low resolution depth image captured by mobile phone.

In Chapter 4 and 5, the proposed methods rely on the local point-specific feature reasoning which enables to recover details and to represent the unseen shape and unseen pose. However, the global constraint is missing to preserve

the topological consistency when occlusion occurs. How to consider both topological consistency and local details? The trivial solution is considering more 3D supervision such as the 3D convolution in Chibane et al. [2020a] but arises a huge memory cost. The topological analysis Chazal et al. [2016], Oudot [2017] is another tool to preserve topological consistency, but the corresponding persistence diagram is normally time consuming in 3D case. The topology-preserving loss Shit et al. [2021] can also be added as the regularization term during training. Therefore, it is a good direction to explore, combining implicit modeling and topological analysis method, in order to benefit from each other.

Some technical contribution can be extended to the larger context. The Gaussian Process layer in Chapter 3 is a good manner to densify the points following the radial basis function kernel to figure out the topology-aware shape. So the style transfer or identity transfer can apply this layer to alleviate the costly long-term training process. It would be interesting to test its capability in this direction.

Bibliography

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4620–4628, 2019.

Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 618–626, 2018a.

Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-sheng Hua. Boosting semantic human matting with coarse annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8572, 2020a.

Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Humangan: A generative model of human images. In *2021 International Conference on 3D Vision (3DV)*, pages 258–267. IEEE, 2021.

Kristijan Bartol, David Bojanić, Tomislav Petković, and Tomislav Pribanić. A review of body measurement using 3d scanning. *IEEE Access*, 2021.

Douglas Lanman and Gabriel Taubin. Build your own 3d scanner: 3d photography for beginners. In *ACM SIGGRAPH 2009 Courses*, pages 1–94. 2009.

Masahiro Yamaguchi, Koki Wakunami, and Mamoru Inaniwa. Computer generated hologram from full-parallax 3d image data captured by scanning vertical camera array. *Chinese Optics Letters*, 12(6):060018, 2014.

Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4dcomplete: Non-rigid motion estimation beyond the observable surface. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 12706–12716, 2021.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. *ACM Transactions on Graphics (TOG)*, 24(3):408–416, 2005.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transaction on Graphics (TOG)*, 34(6):248:1–248:16, October 2015.

Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Skinning with dual quaternions. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 39–46, 2007.

Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Geometric skinning with approximate dual quaternion blending. *ACM Transactions on Graphics (TOG)*, 27(4):1–23, 2008.

Chun-Liang Li, Tomas Simon, Jason Saragih, Barnabás Póczos, and Yaser Sheikh. Lbs autoencoder: Self-supervised fitting of articulated meshes to point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11967–11976, 2019.

Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. 3d-coded : 3d correspondences by deep deformation. In *Proceedings of the European Conference on Computer Vision*, pages 235–251, 2018a.

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):1–15, 2017.

Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, oct 2019a.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. IEEE, 2011.

Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015.

Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 910–919, 2017.

Angela Dai, Matthias Nießner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface re-integration. *ACM Transactions on Graphics 2017 (TOG)*, 2017a.

Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7287–7296, 2018.

Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. Faust: Dataset and evaluation for 3d mesh registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3794–3801, 2014.

Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic faust: Registering human bodies in motion. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 6233–6242, 2017.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. *arXiv preprint arXiv:1904.03278*, 2019.

Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision*, sep 2018.

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 of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

Theo Deprelle, Thibault Groueix, Matthew Fisher, Vladimir Kim, Bryan Russell, and Mathieu Aubry. Learning elementary structures for 3d shape generation and matching. In *Advances in Neural Information Processing Systems*, pages 7435–7445, 2019.

Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2020a.

Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4627–4635, 2017. doi: 10.1109/CVPR.2017.492.

https://renderpeople.com/3d-people. 2018.

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

Marc Pollefeys, Reinhard Koch, Maarten Vergauwen, and Luc Van Gool. Metric 3d surface reconstruction from uncalibrated image sequences. In *European Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, pages 139–154. Springer, 1998.

Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59(3):207–232, 2004.

Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 963–968. Ieee, 2011.

Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011.

Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Gipuma: Massively parallel multi-view stereo reconstruction. *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V*, 25(361-369):2, 2016.

Michael Goesele, Brian Curless, and Steven M Seitz. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2402–2409. IEEE, 2006.

Nader Salman and Mariette Yvinec. Surface reconstruction from multi-view stereo. *Lecture notes in computer science*, 2009.

Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996.

Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 781–796, 2018.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5525–5534, 2019.

Jiayu Yang, Wei Mao, Jose M. Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020.

Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020.

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34, 2021.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

Nicola D'Apuzzo. 3d body scanning technology for fashion and apparel industry. In *Videometrics IX*, volume 6491, pages 203–214. SPIE, 2007.

Mao-Jiun J Wang, Wen-Yen Wu, Kuo-Chao Lin, Shi-Nine Yang, and Jun-Ming Lu. Automated anthropometric data collection from three-dimensional digital human models. *The International Journal of Advanced Manufacturing Technology*, 32(1):109–115, 2007.

Jianfeng Wang, Cha Zhang, Wenwu Zhu, Zhengyou Zhang, Zixiang Xiong, and Philip A Chou. 3d scene reconstruction by multiple structured-light based commodity depth cameras. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5429–5432. IEEE, 2012.

Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992.

Brett Allen, Brian Curless, and Zoran Popović. Articulated body deformation from range scan data. *ACM Transactions on Graphics (TOG)*, 21(3):612–619, 2002.

Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. *ACM transactions on graphics (TOG)*, 22(3):587–594, 2003.

Dragomir Anguelov, Praveen Srinivasan, Hoi-Cheung Pang, Daphne Koller, Sebastian Thrun, and James Davis. The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces. *Advances in neural information processing systems*, 17, 2004.

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.

Xiaohan Shi, Kun Zhou, Yiying Tong, Mathieu Desbrun, Hujun Bao, and Baining Guo. Mesh puppetry: cascading optimization of mesh deformation with inverse kinematics. In *ACM SIGGRAPH 2007 papers*, pages 81–es. 2007.

Yuri Pekelny and Craig Gotsman. Articulated object reconstruction and markerless motion capture from depth video. In *Computer Graphics Forum*, volume 27, pages 399–408. Wiley Online Library, 2008.

Tsz-Ho Kwok, Kwok-Yun Yeung, and Charlie CL Wang. Volumetric template fitting for human body reconstruction from incomplete data. *Journal of Manufacturing Systems*, 33(4):678–689, 2014.

Will Chang and Matthias Zwicker. Automatic registration for articulated shapes. In *Computer Graphics Forum*, volume 27, pages 1459–1468. Wiley Online Library, 2008.

Silvia Zuffi and Michael J Black. The stitched puppet: A graphical model of 3d human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3537–3546, 2015.

Alexander M. Bronstein, Michael M. Bronstein, and Ron Kimmel. Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5):1168–1172, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0508601103. URL https://www.pnas.org/content/103/5/1168.

Khaled Khairy and Jonathon Howard. Spherical harmonics-based parametric deconvolution of 3d surface images using bending energy minimization. *Medical image analysis*, 12(2):217–227, 2008.

Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *Proceedings of the European Conference on Computer Vision*, pages 356–369. Springer, 2010.

Tamal K Dey, Kuiyu Li, Chuanjiang Luo, Pawas Ranjan, Issam Safa, and Yusu Wang. Persistent heat signature for pose-oblivious matching of incomplete models. In *Computer Graphics Forum*, volume 29, pages 1545–1554. Wiley Online Library, 2010.

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009.

Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009.

M. Aubry, U. Schlickewei, and D. Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2011.

Mathieu Andreux, Emanuele Rodola, Mathieu Aubry, and Daniel Cremers. Anisotropic laplace-beltrami operators for shape analysis. In *European Conference on Computer Vision*, pages 299–312. Springer, 2014.

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 (ToG)*, 31(4):1–11, 2012.

Or Litany, Tal Remez, Emanuele Rodola, Alex Bronstein, and Michael Bronstein. Deep functional maps: Structured prediction for dense shape correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5659–5667, 2017.

Emanuele Rodolà, Luca Cosmo, Michael M Bronstein, Andrea Torsello, and Daniel Cremers. Partial functional correspondence. In *Computer graphics forum*, volume 36, pages 222–236. Wiley Online Library, 2017.

Nicolas Donati, Abhishek Sharma, and Maks Ovsjanikov. Deep geometric functional maps: Robust feature learning for shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8592–8601, 2020.

Souhaib Attaiki, Gautam Pai, and Maks Ovsjanikov. Dpfm: Deep partial functional maps. In *2021 International Conference on 3D Vision (3DV)*, pages 175–185. IEEE, 2021.

Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1802–1811, 2017.

Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8958–8966, 2019.

Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 602–618, 2018.

Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7163–7172, 2019.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 4267–4276, 2021.

Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5554–5564, 2022.

Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1450–1459, 2021a.

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

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 561–578. Springer International Publishing, October 2016.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

Nicolas Ugrinovic, Adria Ruiz, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Body size and depth disambiguation in multi-person reconstruction from single images. In *2021 International Conference on 3D Vision (3DV)*, pages 53–63. IEEE, 2021.

Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11179–11188, 2021a.

Or Litany, Alex Bronstein, Michael Bronstein, and Ameesh Makadia. Deformable shape completion with graph convolutional autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1886–1895, 2018.

Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the International Conference on Computer Vision*, 2021.

Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923*, 2022.

Kathleen M Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, and Scott Fleming. Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary. Technical report, Sytronics Inc Dayton Oh, 2002.

Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. In *Computer graphics forum*, volume 28, pages 337–346. Wiley Online Library, 2009.

Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. Moviereshape: Tracking and reshaping of humans in videos. *ACM Transactions on Graphics (TOG)*, 29(6):1–10, 2010.

David A Hirshberg, Matthew Loper, Eric Rachlin, and Michael J Black. Coregistration: Simultaneous alignment and modeling of articulated 3d shape. In *Proceedings of the European Conference on Computer Vision*, pages 242–255. Springer, 2012.

Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. volume 34, pages 120:1–120:14, August 2015.

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)*, 2019.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):245, 2017.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. URL https://doi.org/10.1145/3130800.3130813.

Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *European Conference on Computer Vision*, pages 598–613. Springer, 2020.

B. Jiang, J. Zhang, J. Cai, and J. Zheng. Disentangled human body embedding based on deep hierarchical neural network. In *IEEE Transactions on Visualization and Computer Graphics*, volume 26, pages 2560–2575, 2020.

Sandro Lombardi, Bangbang Yang, Tianxing Fan, Hujun Bao, Guofeng Zhang, Marc Pollefeys, and Zhaopeng Cui. Latenthuman: Shape-and-pose disentangled latent representation for human bodies. In *2021 International Conference on 3D Vision (3DV)*, pages 278–288. IEEE, 2021.

Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):220, 2014.

Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Proceedings of the European Conference on Computer Vision*. Springer, August 2020a.

Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019b.

Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *International Conference on 3D Vision (3DV)*, sep 2019.

Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2020b.

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018.

Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018.

Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020a.

Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020.

Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3d representations. *arXiv preprint arXiv:2111.07868*, 2021.

Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5868–5877, 2017b.

Andrew Nealen, Takeo Igarashi, Olga Sorkine, and Marc Alexa. Laplacian mesh optimization. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 381–389, 2006.

Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006.

Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG)*, 32(3):29, 2013.

Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011.

Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992.

Kok-Lim Low. Linear least-squares optimization for point-to-plane icp surface registration. *Chapel Hill, University of North Carolina*, 4(10):1–3, 2004.

Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu. Hybridfusion: real-time performance capture using a single depth sensor and sparse imus. In *Proceedings of the European Conference on Computer Vision*, pages 384–400, 2018.

Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *European Conference on Computer Vision*, pages 246–264. Springer, 2020.

Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. Motion2fusion: Real-time volumetric performance capture. *ACM Transactions on Graphics (TOG)*, 36(6):1–16, 2017.

Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1395, 2017.

Miroslava Slavcheva, Maximilian Baust, and Slobodan Ilic. Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2646–2655, 2018.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2021.

Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pages 728–737. IEEE, 2018.

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 *Proceedings of the IEEE international conference on computer vision*, pages 85–93, 2017.

Bo Yang, Hongkai Wen, Sen Wang, Ronald Clark, Andrew Markham, and Niki Trigoni. 3d object reconstruction from a single depth view with adversarial

learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 679–688, 2017.

David Stutz and Andreas Geiger. Learning 3d shape completion from laser scan data with weak supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1955–1964, 2018.

Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions On Graphics (TOG)*, 36(4):1–11, 2017.

Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018.

Lyne P Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. Topnet: Structural point cloud decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 383–392, 2019.

Xiaogang Wang, Marcelo H Ang Jr, and Gim Hee Lee. Cascaded refinement network for point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 790–799, 2020.

Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11596–11603, 2020b.

Wenxiao Zhang, Qingan Yan, and Chunxia Xiao. Detail preserved point cloud completion via separated feature aggregation. In *European Conference on Computer Vision*, pages 512–528. Springer, 2020.

Hong-Kai Zhao, Stanley Osher, and Ronald Fedkiw. Fast surface reconstruction using the level set method. In *Proceedings IEEE Workshop on Variational and Level Set Methods in Computer Vision*, pages 194–201. IEEE, 2001.

Stanley Osher and Guillermo Sapiro. Dynamic visibility in an implicit framework. 2002.

Ronald P Fedkiw, Guillermo Sapiro, and Chi-Wang Shu. Shock capturing, level sets, and pde based methods in computer vision and image processing: a review of osher's contributions. *Journal of Computational Physics*, 185(2):309–341, 2003.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

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

Alexandre Boulch, Pierre-Alain Langlois, Gilles Puy, and Renaud Marlet. Needrop: Self-supervised shape representation from sparse point clouds using needle dropping. In *2021 International Conference on 3D Vision (3DV)*, pages 940–950. IEEE, 2021.

Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

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*, October 2019.

Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020.

Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *Proceedings of the European Conference on Computer Vision*. Springer, August 2020a.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2020b.

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.

Matan Atzmon and Yaron Lipman. Sald: Sign agnostic learning with derivatives. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=7EDgLu9reQD.

Baorui Ma, Yu-Shen Liu, and Zhizhong Han. Reconstructing surfaces for sparse point clouds with on-surface priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7154–7164, 2019.

Benjamin Ummenhofer and Vladlen Koltun. Adaptive surface reconstruction with multiscale convolutional kernels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5651–5660, 2021.

Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020b.

Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. Learning gradient fields for shape generation. In *European Conference on Computer Vision*, pages 364–381. Springer, 2020.

Ijaz Akhter, Tomas Simon, Sohaib Khan, Iain Matthews, and Yaser Sheikh. Bilinear spatiotemporal basis models. *ACM Transactions on Graphics (TOG)*, 31(2):17, 2012.

Tomas Simon, Jack Valmadre, Iain Matthews, and Yaser Sheikh. Separable spatiotemporal priors for convex reconstruction of time-varying 3d point clouds. In *Proceedings of the European Conference on Computer Vision*, pages 204–219. Springer, 2014.

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision*, pages 98–109, Sep 2018a. doi: 10.1109/3{DV}. 2018.00022.

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018b.

Thiemo Alldieck, Marc Kassubeck, Bastian Wandt, Bodo Rosenhahn, and Marcus Magnor. Optical flow-based 3d human motion estimation from monocular video. In *German Conference on Pattern Recognition*, pages 347–360. Springer, 2017.

Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5236–5246, 2017.

Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

Amit Bermano, Thabo Beeler, Yeara Kozlov, Derek Bradley, Bernd Bickel, and Markus Gross. Detailed spatio-temporal reconstruction of eyelids. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

Qian Zheng, Xiaochen Fan, Minglun Gong, Andrei Sharf, Oliver Deussen, and Hui Huang. 4d reconstruction of blooming flowers. In *Computer Graphics Forum*, volume 36, pages 405–417. Wiley Online Library, 2017.

Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. General dynamic scene reconstruction from multiple view video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 900–908, 2015.

Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Multi-view dynamic shape refinement using local temporal integration. In *Proceedings of the IEEE international conference on computer vision*, pages 3094–3103, 2017.

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 *European conference on computer vision*, pages 405–421. Springer, 2020.

Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL https://proceedings.neurips.cc/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 936–944, 2017.

Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021b.

Hao Zhu, Xinxin Zuo, Haotian Yang, Sen Wang, Xun Cao, and Ruigang Yang. Detailed avatar recovery from single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 2021a.

Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2019.

Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. Cross-modal deep face normals with deactivable skip connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4979–4989, 2020.

Jonathan Cohen, Marc Olano, and Dinesh Manocha. Appearance-preserving simplification. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 115–122, 1998.

Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2315–2324, 2019.

Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5704–5712, 2016.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. 2018.

Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019.

Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Kumara Kahatapitiya and Michael S Ryoo. Coarse-fine networks for temporal activity detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8385–8394, 2021.

Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *European Conference on Computer Vision*, pages 154–171. Springer, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Advances in Neural Information Processing Systems*, 34, 2021b.

Noah Stier, Alexander Rich, Pradeep Sen, and Tobias Höllerer. Vortx: Volumetric 3d reconstruction with transformers for voxelwise view selection and fusion. In *2021 International Conference on 3D Vision (3DV)*, pages 320–330. IEEE, 2021.

Maks Ovsjanikov, Quentin Mérigot, Facundo Mémoli, and Leonidas Guibas. One point isometric matching with the heat kernel. *Computer Graphics Forum*, 29(5):1555–1564, 2010. doi: 10.1111/j.1467-8659.2010.01764.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2010.01764.x.

Vladimir Kim, Yaron Lipman, and Thomas Funkhouser. Blended intrinsic maps. *ACM Transactions on Graphics (TOG)*, 30(4), jul 2011.

E. Rodolà, S. Bulò, T. Windheuser, M. Vestner, and D. Cremers. Dense non-rigid shape correspondence using random forests. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4184, 2014.

Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5115–5124, 2017.

Lingyu Wei, Qixing Huang, Duygu Ceylan, Etienne Vouga, and Hao Li. Dense human body correspondences using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1544–1553, 2016.

Oshri Halimi, Or Litany, Emanuele Rodola, Alex M Bronstein, and Ron Kimmel. Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4370–4379, 2019.

Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. Variational autoencoders for deforming 3d mesh models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5841–5850, June 2018. doi: 10.1109/CVPR.2018.00612.

Riccardo Marin, Simone Melzi, Emanuele Rodolà, and Umberto Castellani. FARM: Functional automatic registration method for 3d human bodies. In *Computer Graphics Forum*, volume 39, pages 160–173. Wiley Online Library, 2020.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4332–4341, October 2019.

Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 216–224, 2018b. doi: 10.1109/CVPR.2018.00030.

Yipeng Hu, Eli Gibson, Nooshin Ghavami, Ester Bonmati, Caroline M Moore, Mark Emberton, Tom Vercauteren, J Alison Noble, and Dean C Barratt. Adversarial deformation regularization for training image registration neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 774–782. Springer, 2018.

Nadia Magnenat-Thalmann, Richard Laperrire, and Daniel Thalmann. Joint-dependent local deformations for hand animation and object grasping. In *Proceedings of Graphics Interface*, volume '88. Citeseer, 1988.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

Marcel Lüthi, Thomas Gerig, Christoph Jud, and Thomas Vetter. Gaussian process morphable models. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1860–1873, 2017.

Andrew Gordon Wilson, David A Knowles, and Zoubin Ghahramani. Gaussian process regression networks. In *Proceedings of the International Conference on Machine Learning*, 10 2011.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378, 2016.

Andreas Damianou and Neil Lawrence. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

Victoria Fernández Abrevaya, Adnane Boukhayma, Stefanie Wuhrer, and Edmond Boyer. A decoupled 3d facial shape model by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–10, October 2019. URL https://hal.archives-ouvertes.fr/hal-02064711.

Gil Shamai, Ron Slossberg, and Ron Kimmel. Synthesizing facial photometries and corresponding geometries using generative adversarial networks. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15:1–24, 10 2019. doi: 10.1145/3337067.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

Jean Basset, Stefanie Wuhrer, Edmond Boyer, and Franck Multon. Contact Preserving Shape Transfer For Rigging-Free Motion Retargeting. In *MIG 2019 - ACM SIGGRAPH Conference Motion Interaction and Games*, pages 1–10, October 2019. doi: 10.1145/3359566.3360075. URL https://hal.archives-ouvertes.fr/hal-02293308.

Qifeng Chen and Vladlen Koltun. Robust nonrigid registration by convex optimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2039–2047, 2015.

Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Oshri Halimi, Ido Imanuel, Or Litany, Giovanni Trappolini, Emanuele Rodolà, Leonidas Guibas, and Ron Kimmel. The whole is greater than the sum of its nonrigid parts. In *Proceedings of the European Conference on Computer Vision*, 2020.

Boyao Zhou, Jean-Sebastien Franco, Federica Bogo, Bugra Tekin, and Edmond Boyer. Reconstructing human body mesh from point clouds by adversarial gp network. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision*, September 2018.

Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in Neural Information Processing Systems 32*, pages 492–502. 2019.

Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Proceedings of the European Conference on Computer Vision*, pages 612–628. Springer, 2020b.

Marko Mihajlovic, Yan Zhang, Michael J. Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2021.

Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2021.

Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r11Q2SlRW.

Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1423–1432. IEEE, 2019.

Zhe Cao, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Proceedings of the European Conference on Computer Vision*, 2020.

Shuaihang Yuan, Xiang Li, Anthony Tzes, and Yi Fang. 3dmotion-net: Learning continuous flow function for 3d motion prediction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020.

Chiyu Jiang, Jingwei Huang, Andrea Tagliasacchi, and Leonidas Guibas. Shapeflow: Learnable deformations among 3d shapes. In *Advances in Neural Information Processing Systems*, 2020c.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2019.

Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Proceedings of the European Conference on Computer Vision*, pages 362–379, 2016.

William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4): 163–169, 1987.

Thomas Lewiner, Hélio Lopes, Antônio Wilson Vieira, and Geovan Tavares. Efficient implementation of marching cubes' cases with topological guarantees. *Journal of graphics tools*, 8(2):1–15, 2003.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017.

Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017.

Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019.

Boyao Zhou, Jean-Sebastien Franco, Federica Bogo, and Edmond Boyer. Spatio-temporal human shape completion with implicit function networks. In *Proceedings of the International Conference on 3D Vision*, 2021.

Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020.

Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019.

Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. 2019.

Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020.

Frédéric Chazal, Vin De Silva, Marc Glisse, and Steve Oudot. *The structure and stability of persistence modules*, volume 10. Springer, 2016.

Steve Y Oudot. *Persistence theory: from quiver representations to data analysis*, volume 209. American Mathematical Soc., 2017.

Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. cldice - a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16560–16569, June 2021.

Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6239–6249, 2021b.