3D Textured Shape Recovery with Learned Geometric Priors

Lei Li*
ETH Zurich
leilil@ethz.ch

Zhizheng Liu*
ETH Zurich
liuzhi@ethz.ch

Weining Ren*
ETH Zurich
weiren@ethz.ch

Liudi Yang*
ETH Zurich
liudyang@ethz.ch

Fangjinhua Wang
ETH Zurich
fangjinhua.wang@inf.ethz.ch

Marc Pollefeys
ETH Zurich
marc.pollefeys@inf.ethz.ch

Songyou Peng
ETH Zurich
songyou.peng@inf.ethz.ch

Abstract

3D textured shape recovery from partial scans is crucial for many real-world applications. Existing approaches have demonstrated the efficacy of implicit function representation, but they suffer from partial inputs with severe occlusions and varying object types, which greatly hinders their application value in the real world. This technical report presents our approach to address these limitations by incorporating learned geometric priors. To this end, we generate a SMPL model from learned pose prediction and fuse it into the partial input to add prior knowledge of human bodies. We also propose a novel completeness-aware bounding box adaptation for handling different levels of scales and partialness of partial scans.

1. Introduction

In this technique report, we demonstrate our solution to the 3rd SHApe Recovery from Partial textured 3D scans (SHARP) challenge, which wins the track 2 of the challenge and ranks 2nd overall. Recovering 3D textured shapes from partial scans is useful and practical in many applications such as augmented reality, but it is a challenging task especially with partial scans obtained from real world. The partial scans can have various sizes and appearances and may suffer from severe occlusion and motion blur.

Existing approaches on 3D shape recovery have shown promising results with an implicit function representation. For example, IF-Net [1] and ConvOcc-Net [6] first process the partial scan into spatial features and then representing the completed shape as an implicit occupancy map and the texture as an implicit texture field. While these approaches have great performance in simple recovery tasks, we observe two main failure cases in this year's challenge . 1) They often fail when the partial scan has too many missing parts. When a whole arm of the human is missing, they tend to predict an incomplete arm with many artifacts or nothing at all. 2) They can only handle objects with known sizes and similar shapes. However, in track 2 of the challenge, the objects have varying appearances and sizes. We hypothesize these happen as 1) Given limited training data, it is hard to infer the missing body parts when there is no prior information about the human body. 2) Before obtaining the spatial feature, a bounding box is required to first convert the partial scan to a 3D grid. When the object size is varying, the bounding box often fails to match the completed shape.

To solve these issues, we propose to add geometric priors to help regularize the prediction of completed shapes. 1) For recovering human bodies (track 1), we first predict the human pose and convert it into a SMPL model [3] as the generic shape to guide the completion of missing body parts. 2) For recovering of universal objects, we observe that the bounding box acts as a strong prior for the final predicted shape. Therefore, we propose a novel completeness-aware bounding box adaptation module, which learns to adjust the size of the bounding box adaptively to objects of different sizes and partialness. Finally, we further boost the performance by combining the predicted complete shape and the initial partial scan in a shape fusion stage. Thanks to the learned geometric priors, our approach greatly outperforms the baseline implicit-function-based approaches.

2. Method

2.1. Implicit Functions Learning from Surfaces

Implicit methods in the 3D area represent the surface as the level-set of a continuous function. Implicit function learning shows great strong potential to reconstruct objects, for it can model arbitrary resolution or topology. It consists of feature encoding block and feature decoding block (Fig.1a and Fig.1b).

* Equal Contributions

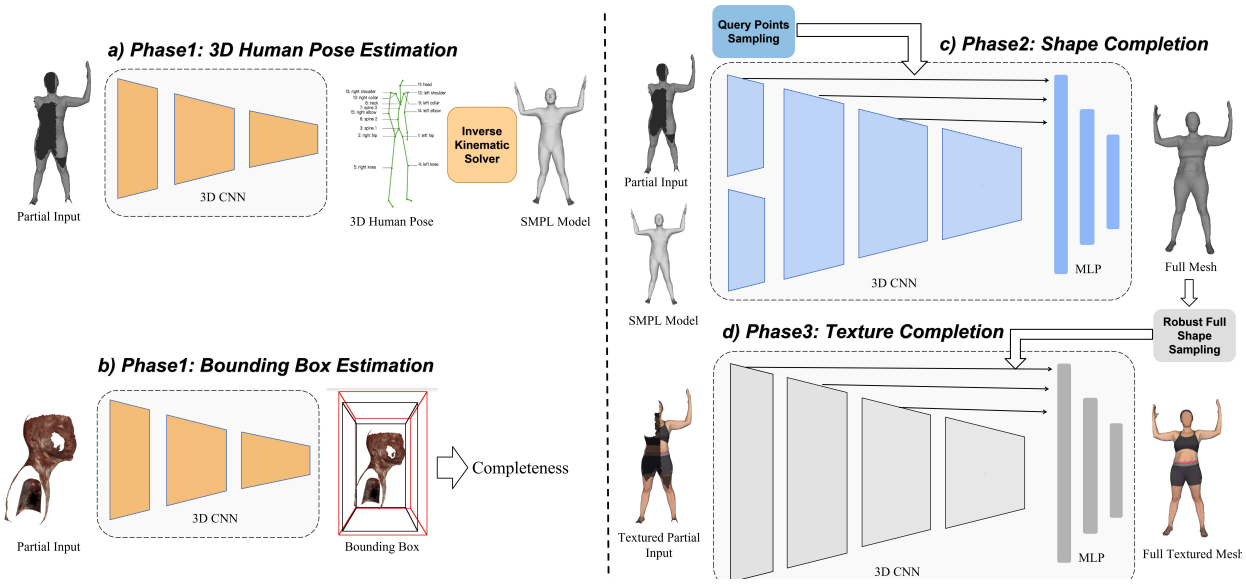

Figure 1. Network Architecture. We use three phases to complete our reconstruction. For the geometric prior, we first apply the 3D CNN module to extract the global features and map them to the 3D human pose and bounding box, respectively. For the phase2 of the human body completion task, features from the SMPL model are aggregated with input partial features to provide additional information. In texture completion, the ground truth shape is used as query points for training and the predicted full shape in the phase2 is used as query points for inference.

Encoding from Occupancy Surfaces: The encoder takes voxelized data as input. It is mainly composed of 3D CNN modules and the maxpooling layers followed by instance normalization [7] and ReLU activation. Features are extracted hierarchically with increased receptive fields to obtain local details and global context. Following the design in IF-Net [1], the multi-scale feature grids $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n$ are obtained by utilizing 3D CNN and the maxpooling recursively to downscale the features.

$$\begin{aligned} \mathcal{X}_l &= \text{Maxpolling}(\mathcal{X'}_l) \\ \mathcal{X'}_l &= \text{IN}(\epsilon(\text{3DCNN}(\mathcal{X}_{l-1}))) \end{aligned} \tag{1}$$

where IN denotes instance normalization, and ϵ denotes ReLU activation function. From the encoder, we concatenate multi-scale feature grids to obtain $\mathcal{X}_{0:n}$:

$$\mathcal{X}_{0:n} = [\mathcal{X}_0, \mathcal{X}_1, \cdots, \mathcal{X}_n] \tag{2}$$

where $[,]$ denotes the concatenation operator.

Decoding from the Feature Space: From the multi-scale features $\mathcal{X}_{0:n}$, we generate the corresponding interpolated features $\mathcal{X}_{0:n}(q)$ with the query points $\mathbf{q} \in \mathbb{R}^{N\times 3}$. The decoder module $\mathcal{F}$ predicts whether there is a surface point at the query point $\mathbf{q}_i$ coordinate.

Typically, neural networks tend to struggle with modeling high-frequency information. High-frequency information in the spatial domain greatly reduces the complexity and frequency the neural network needs to model [4]. By adding positional embedding in Fourier domain, we can effectively map our features to higher-dimensional space. A shared Multi-layer Perceptron (MLP) is used to decode the features:

$$\mathcal{F}(\mathcal{X}_{0:n}, \mathbf{q}_i) = \text{MLP}([\mathcal{X}_{0:n}(q_i), PE(q_i)]) \tag{3}$$

where PE denotes the positional embedding.

2.2. Geometric Priors

3D Human Pose Estimation: The complex human poses and non-rigid deformations make human body reconstruction very challenging, especially when our observations are very vestigial. Some simple geometric priors, such as the SMPL model [3], can guide shape reconstruction with correct 3D human pose and deformation. The parameters of the SMPL model generally include 3D poses and shapes. We focus only on learning 3D poses, since it is more important to infer the body poses for the vestigial inputs, and reduce the complexity of the geometric prior learning task. In 3D human pose estimation, a simple skeleton system is usually used to model the vital parts of a human. We describe the human body as a stick model consisting of 25 3D joint coordinates. Following our encoder design, we extract the global feature from the final feature grid $\mathcal{X}_n$ with the largest receptive field (Fig.1a):

$$\mathcal{G}^p(\mathcal{X}_n) = \text{Maxpolling}(\text{MLP}(\mathcal{X}_n)) \tag{4}$$

The global feature $\mathcal{G}^p(\mathcal{X}_n)$ is mapped to the 3D human pose directly. We then optimize the 3D human pose to the SMPL model based on the inverse kinematic solver with basic shape parameters.

Completeness-aware Bounding Box Prediction: Due to the large-scale variance of universal objects, we cannot use a unified bounding box for all objects like human bodies. We design a bounding box regression network to predict a proper bounding box from partial input. The partial input object is voxelized according to its tight bounding box (coordinate range) $\mathbf{b}_t = [x_t, y_t, z_t, l_t, w_t, h_t]$. The network shares a similar structure with the human pose estimation network:

$$\mathcal{G}^b(\mathcal{X}_n) = \text{Maxpolling}(\text{MLP}(\mathcal{X}_n)) \tag{5}$$

The global feature $\mathcal{G}^b(\mathcal{X}_n)$ is mapped to the relative coordinate of 6D bounding box w.r.t the tight bounding box $\tilde{\mathbf{b}}_r = [x_r, y_r, z_r, l_r, w_r, h_r]$, which can be converted to to the absolute coordinate $\mathbf{b}_a = [x_a, y_a, z_a, l_a, w_a, h_a]$ for further shape and texture completion task. Besides, the size of predicted bounding box can work as a indicator for the incomplete level. We generate the completed mesh only if the bounding box meet either of two criterion:

$$max\{\mathbf{b}_{a,size}\}/min\{\mathbf{b}_{a,size}\} > t_1 \tag{6}$$

$$max\{\mathbf{b}_{r,size}\}/min\{\mathbf{b}_{r,size}\} > t_2 \tag{7}$$

where $\mathbf{b}_{size} = [l, w, h]$ and t_1, t_2 are two hyperparameter for completeness.

2.3. Shape Completion

For phase2, to voxelized the input data, we start by normalizing each 3D mesh to a unit box and densely sampling spatial points. Then, we generate occupancy inputs $\mathcal{X}_0 \in \{0,1\}^{K\times K\times K}$ by assigning sampled spatial points to the nearest neighbor grids. For the query points sampling, we directly sample sufficient points $\mathbf{q}$ near the surfaces to capture more shape details, and convert them into occupancy representations. Besides, for the textured human body completion, we extract features from partial inputs and our estimated SMPL models, and concatenate them at an early stage. For the decoder, a MLP is used to decode the features into occupancy probabilities:

$$\mathcal{F}^s(\mathcal{X}_{0:n}, \mathbf{q}_i) = \sigma(\text{MLP}([\mathcal{X}_{0:n}(q_i), PE(q_i)])) \mapsto [0,1] \tag{8}$$

where σ denotes sigmoid activation. Finally, the predicted occupancy outputs are converted to meshes using marching cubes, and recovered into the original scale.

Since texture completion is based on shape completion, improving the accuracy of shape completion becomes a critical issue. Therefore, we apply our completion results to the missing part only by fusing with the partial inputs to improve the accuracy of shape completion. For the scanned regions that exist in original inputs, we retain their original shapes.

2.4. Texture Completion

The ground truth shape is used for training and the predicted complete shape from phase2 is used for inference. The encoder takes 4-channel (RGB and occupancy) voxelized data as input. The step to voxelized inputs is the same as the shape completion. For query points, we densely sample on the mesh and add noise in the normal direction to compensate for the error in the shape completion phase. For the decoder, given a point p_i from the untextured full shape, it is decoded to obtain the color of the surface:

$$\mathcal{F}^t(\mathcal{X}_{0:n}, \mathbf{q}_i) = \text{MLP}([\mathcal{X}_{0:n}(q_i), PE(q_i)]) \mapsto [0,255]^3 \tag{9}$$

2.5. Loss Functions

For different phases, we train different networks separately. Therefore, the loss functions of each phase are also independent.

ℓ_1 **Loss:** For 3D human pose estimation, we use the average ℓ_1 loss between the ground truth joint coordinates and the predicted ones:

$$L_{\text{pose}} = \sum_{i=1}^{N_{\text{joint}}} \left\|\hat{t}_i - t_i\right\|_1 \tag{10}$$

where $\hat{t}_i$ is the predicted i^{th} joint coordinate and t_i is the corresponding ground truth coordinate:

$$L_{\text{bbox}} = \left\|\hat{\mathbf{b}} - \mathbf{b}\right\|_1 \tag{11}$$

where $\hat{\mathbf{b}}$ is the predicted bounding box parameters and $\mathbf{b}$ is the corresponding ground truth one. For texture completion, RGB values are directly used to compute loss:

$$L_{\text{texture}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{\mathbf{c}}_i - \mathbf{c}_i\|_1 \tag{12}$$

where $\hat{\mathbf{c}}_i$ is the predicted i^{th} RGB value and $\mathbf{c}_i$ is the corresponding ground truth RGB value.

Balanced BCE Loss: For shape completion, most of the querying points are non-occupied. We introduce a balanced BCE loss to solve the imbalance in the querying occupancy.

$$L_{\text{shape}} = -\frac{1}{N}\sum_{i=1}^{N} w_p \cdot o_i \cdot \log(\hat{o}_i) + w_n \cdot (1-o_i) \cdot \log(1-\hat{o}_i) \tag{13}$$

where w_p and w_n denote the proportion of voxels with states 1 and 0 in the ground truth, respectively. $\hat{o}_i$ is the predicted i^{th} occupancy probability and o_i is the corresponding ground truth occupancy.

Figure 2. Visualization of some results in track 1. The left is partial input and the right is the reconstruction result. Via SMPL model fusion, we can recover error-free human poses for the shape score. And the texture details can also be generated.

3. Experiments and Results

3.1. Implementation Details

We implement our network using Pytorch [5]. We take feature grids with $n = 6$ scales. We sample 100,000 points for the voxelization preprocessing and the resolution of the grid is 128^3. We use the Adam optimizer [2] with learning rate 0.0001. Our models are trained for 40 epochs on 4 Nivida TITAN Xp GPUs. The batch size on each GPU is set to 1.

3.2. Recovering Textured Human Body Scans

In this track, we train and evaluate our model on the 3DBodyTex.v2 dataset. The challenge is to recover complete textured human body scans from partial bodies of different poses. Considering the potential geometry prior provided by the SMPL model, our model can regenerate the full colored mesh with high accuracy. Some qualitative demos are shown in Fig. 2 , in which the missing parts are replenished appropriately.

3.3. Recovering Textured Object Scans

The dataset consists of different generic objects with considerable scale variation in track 2. Compared to the model using the consistent bounding box, our method to predict possible bounding box as geometry prior for the specific object can generate more reasonable results. Some representative results are shown in Fig. 3.

3.4. Ablation Study

IF-Net	SMPL-fusion	BL	SF	final
✓				82.29
✓	✓			83.26
✓	✓	✓		83.58
✓	✓	✓	✓	**84.68**

Table 1. Ablation study of SHARP Challenge 1 Track 1. BL denotes balanced BCE loss and SF denotes shape fusion.

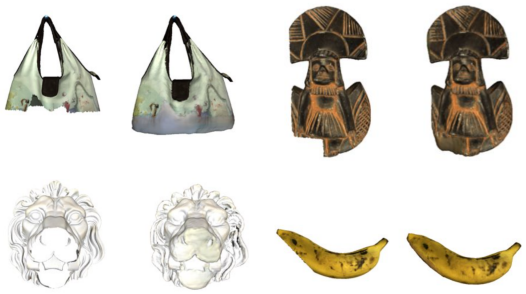

Figure 3. Visualization of results in track 2. The left is partial input and the right is the reconstruction result. Our model can achieve great performance to recover the partial scans. Using the predicted bounding box as prior, the completion is coherent with shape and texture.

IF-Net	Bbox	Post-processing	final
✓			62.34
✓	✓		64.14
✓	✓	✓	**69.56**

Table 2. Ablation study of SHARP Challenge 1 Track 2. Bbox denotes the completeness-aware bounding box prediction. Post-processing denotes shape fusion and completeness-based filter

In this part, we remove some of the modules of the model to have a better understanding of how exactly our strategy improves the model performance. In track 1, Table 1 shows the effect of SMPL model fusion, balanced loss and post-processing.The fusion of geometry prior from SMPL model can make great use of the human pose to obtain more accurate shape. The balanced loss can optimize the training procedure to reason about the missing parts. The post-processing can select better completion from output and input. The result of Table 2 demonstrates that the predicted bounding box can address the variation of the scale effectively to get more accurate geometry. The post-processing strategy also improves the quality of the full mesh.

4. Conclusions

Our solution to the 3rd SHARP challenge improves existing approaches by incorporating learned geometric priors. To handle partial inputs with too many missing parts, we add priors of the general shape by a fusion module that combines a SMPL model generated from pose prediction with the input. To address varying appearances and sizes of the objects, we introduce completeness-aware bounding box adaptation to learn an appropriate prior for the magnitude of the recovered shape. Our future work includes improving the heuristic to reject bad predictions and finding better ways to incorporate the SMPL model.

References

[1] J. Chibane, T. Alldieck, and G. Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6970–6981, 2020. 1, 2

[2] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[3] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1, 2

[4] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 2

[5] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 4

[6] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020. 1

[7] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 2