Fine Structure-Aware Sampling: A New Sampling Training Scheme for Pixel-Aligned Implicit Models in Single-View Human Reconstruction

Kennard Yanting Chan[1,2], **Fayao Liu**[2], **Guosheng Lin**[1], **Chuan Sheng Foo**[2,3], **Weisi Lin**[1]
[1]Nanyang Technological University
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]Centre for Frontier AI Research, A*STAR, Singapore
kenn0042@e.ntu.edu.sg, liu_fayao@i2r.a-star.edu.sg, gslin@ntu.edu.sg, foo_chuan_sheng@i2r.a-star.edu.sg, wslin@ntu.edu.sg

Abstract

Pixel-aligned implicit models, such as PIFu, PIFuHD, and ICON, are used for single-view clothed human reconstruction. These models need to be trained using a sampling training scheme. Existing sampling training schemes either fail to capture thin surfaces (e.g. ears, fingers) or cause noisy artefacts in reconstructed meshes. To address these problems, we introduce Fine Structured-Aware Sampling (FSS), a new sampling training scheme to train pixel-aligned implicit models for single-view human reconstruction. FSS resolves the aforementioned problems by proactively adapting to the thickness and complexity of surfaces. In addition, unlike existing sampling training schemes, FSS shows how normals of sample points can be capitalized in the training process to improve results. Lastly, to further improve the training process, FSS proposes a mesh thickness loss signal for pixel-aligned implicit models. It becomes computationally feasible to introduce this loss once a slight reworking of the pixel-aligned implicit function framework is carried out. Our results show that our methods significantly outperform SOTA methods qualitatively and quantitatively. Our code is publicly available at https://github.com/kcyt/FSS.

1 Introduction

3D reconstruction of human bodies is an area that has garnered interest due to its potential applications in fields such as virtual reality, 3D printing, and game production. Although it is already possible to accurately reconstruct a human body using high-end, multi-view capturing systems (Collet et al. 2015; Lombardi et al. 2018), such systems are unavailable to typical consumers. This has led to research efforts to develop deep learning models for 3D human reconstruction using sparse inputs like a single RGB image (Alldieck et al. 2019; Natsume et al. 2019; Saito et al. 2020).

An influential class of deep learning methods for single-image clothed human reconstruction is pixel-aligned implicit models (Saito et al. 2019, 2020; Chan et al. 2022b). These methods learn an implicit function that represents the surface of a human body. From the learned implicit function, a mesh of a human body can be extracted using the Marching Cubes algorithm (Lorensen and Cline 1987).

To learn the implicit function, all pixel-aligned implicit models have to be trained using a sampling training scheme.

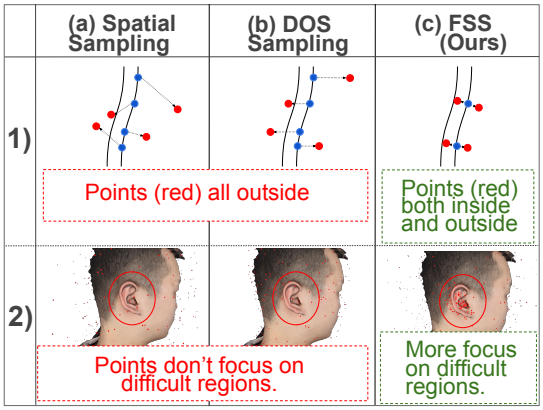
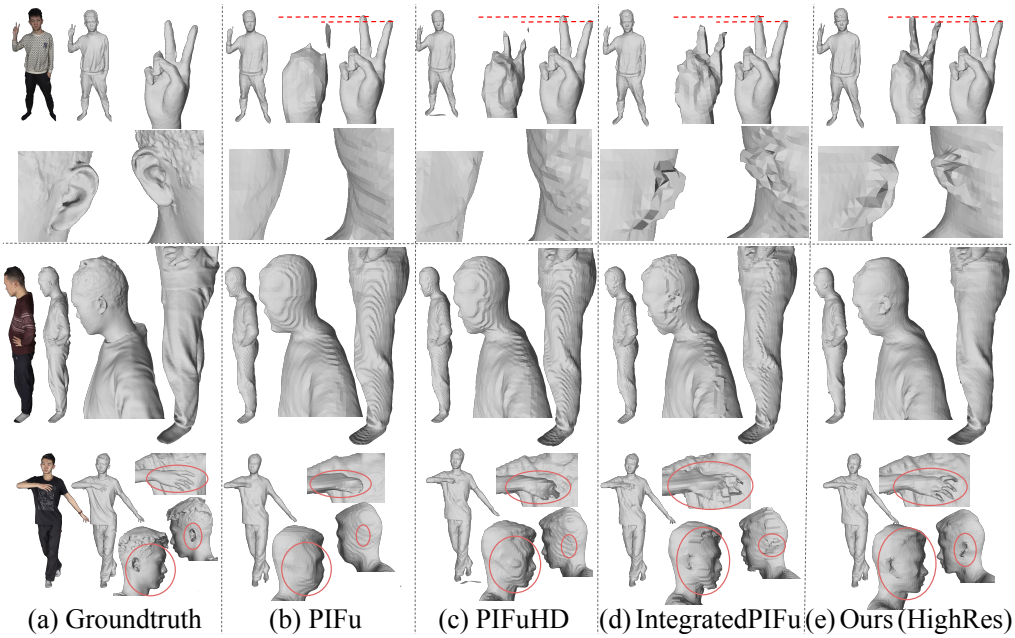


Figure 1: Unlike existing schemes, FSS can: 1. Adapts to thickness of mesh. 2. Prioritize regions that are challenging. Refer to our arXiv version for a more detailed figure.

A sampling training scheme provides supervision signals to a pixel-aligned implicit model by generating sample points and their corresponding labels. How the sample points are generated and what labels are computed will drastically affect the models' results. The earliest sampling training scheme, called spatial sampling scheme, was introduced by the first pixel-aligned implicit model - PIFu (Saito et al. 2019). Since then, subsequent pixel-aligned implicit models, such as PIFuHD (Saito et al. 2020), PaMIR (Zheng et al. 2021), ICON (Xiu et al. 2022), S-PIFu (Chan et al. 2022a), and more, simply continue to use the same scheme.

IntegratedPIFu (Chan et al. 2022b) is the first work to contribute a new sampling training scheme (i.e. Depth-Oriented Sampling or DOS) that differs materially from spatial sampling. Unlike spatial sampling, DOS is able to train models to reconstruct thin but important body features like ears and fingers. But DOS suffers from a lack of robustness. Specifically, DOS only helps camera-facing mesh surfaces, and causes wavy, noisy artefacts on non-camera-facing surfaces.

To overcome these issues, we propose **Fine Structure-aware Sampling (FSS)**, a new sampling training scheme that teaches pixel-aligned implicit models to reconstruct thin body features that are not only artefact-free but also structurally accurate. FSS achieves this by proactively adapting



(a) Groundtruth (b) PIFu (c) PIFuHD (d) IntegratedPIFu (e) Ours (HighRes)

Figure 2: Unlike SOTA methods, our method captures thin body features (e.g. fingers, ears) w/o causing noisy, wavy artefacts.

to the thickness and complexity of surfaces. This is shown in Fig. 1, which also compares FSS with the existing schemes.

In addition, we offer two extensions to our FSS scheme. The first of which is the exploitation of the Normals of Sample Points **(NSP)**, which shows how normals of sample points can be capitalized on during training to further improve reconstruction results.

The second extension to FSS is our Mesh Thickness Loss **(MTL)** signal, which allows a pixel-aligned implicit model to directly learn the thickness of different body parts. We strongly encourage readers to refer to our arXiv version for a clearer and more detailed explanation of our work.

2 Related Work

2.1 Single-view Human Reconstruction

Methods that reconstruct a human body mesh from a single image can be broadly classified into two classes: Parametric methods and non-parametric methods.

Parametric methods (Kanazawa et al. 2018; Kolotouros et al. 2019) recover human body shapes by predicting parameters belonging to a chosen human parametric model (e.g. SMPL-X (Pavlakos et al. 2019)). However, these parametric methods can only produce cloth-less and hairless human body meshes. While methods like Multi-garment Net (Bhatnagar et al. 2019) and Bcnet (Jiang et al. 2020) try to predict clothes on top of a human parametric model, these methods are unable to produce accurate clothed meshes.

On the other hand, non-parametric methods do not use a human parametric model. As aforementioned, a subclass of non-parametric methods that has attracted significant attention from the research community is the pixel-aligned implicit models. PIFu (Saito et al. 2019) is the first of such models, and it is able to reconstruct highly accurate clothed human meshes from a single image.

After PIFu, other pixel-aligned implicit models have been proposed. These include PIFuHD (Saito et al. 2020), StereoPIFu (Hong et al. 2021), S-PIFu (Chan et al. 2022a), IntegratedPIFu (Chan et al. 2022b), and more.

2.2 PIFu and IntegratedPIFu

The original PIFu (Saito et al. 2019) uses an encoder-decoder architecture. An illustration of the PIFu architecture is given in Fig. 3 (Not including the predicted normal maps and the magenta arrows, fonts, and boxes). During training and testing, 3D sample points are sampled within the 3D camera space (of the RGB image). The decoder, which is a multi-layer perceptron (MLP), will predict a value from 0 to 1 where value>0.5 means the sample point is 'inside' a groundtruth human body mesh, value<0.5 means the sample point is 'outside', and value=0.5 means the sample point is exactly on a human body surface. During training, PIFu uses what is known as the spatial sampling scheme to generate sample points and assign labels to these sample points. The spatial sampling scheme mainly generates sample points by displacing mesh surface points of a groundtruth human mesh with normally-distributed noise. These sample points can be displaced in any direction and in any magnitude. An illustration of these sample points is shown in Fig. 1a. Sample points in this scheme are given binary labels (either 0 or 1).

In IntegratedPIFu (Chan et al. 2022b), the authors identified problems with the spatial sampling scheme and proposed the Depth-Oriented Sampling (DOS) training scheme. They showed that the binary labels (as opposed to continuous labels) and unconstrained displacement of mesh sur-

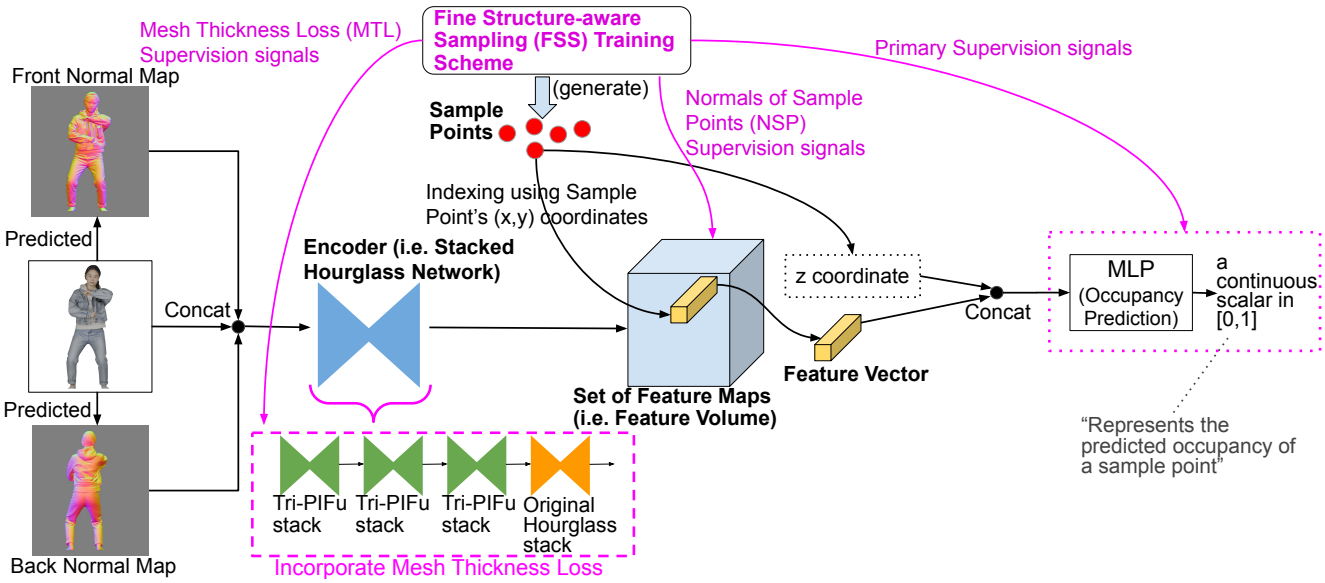


Figure 3: Overview of our FSS sampling training scheme (with NSP and MTL) in a single architecture.

face points in spatial sampling scheme caused thin but obvious body features like the ears and fingers to be missing in the reconstructed meshes. Their proposed DOS scheme introduced continuous labels and constrained displacement of surface points. Continuous labels provide more information and clearer supervision to a pixel-aligned implicit model during training. Constrained displacement of mesh surface points (i.e. surface points are only displaced in the camera-direction) narrows down the meaning of each label, and this makes learning simpler for a pixel-aligned implicit model. These two advantages allow a pixel-aligned implicit model to reconstruct human meshes that retain thin body features like ears and fingers, unlike spatial sampling. However, DOS has serious flaws that will be explained in the next section.

3 Method

Most pixel-aligned implicit models, including PIFuHD, StereoPIFu, S-PIFu, and IntegratedPIFu, use PIFu as a major building block(s) in their own architectures. Thus, we test our FSS scheme in an architecture and pipeline (see Fig. 3) that are identical to those of a PIFu, except for the use of predicted normal maps, FSS scheme, and Tri-PIFu stacks.

First, like in PIFuHD and IntegratedPIFu, we use a pix2pixHD network (Wang et al. 2018) to predict front and back normal maps from an input image. The normal maps and image are concatenated and fed into an encoder, which produces a feature volume. Next, sample points will index the feature volume to retrieve feature vectors that are fed into a MLP, which will predict the sample points' occupancy.

Fig. 3 also shows an overview of our three contributions. First, FSS trains a pixel-aligned implicit model to reconstruct thin body features by giving clearer and more meaningful 'Primary Supervision signals', which are used to guide occupancy prediction. Second, an extension of FSS is to use Normals of Sample Points (NSP) as additional supervision signals. NSP signals guide and mold the representations in the feature volume. Third, another extension of FSS is to introduce Mesh Thickness Loss (MTL) supervision signals, which will train the encoder to account for mesh thickness. We will now elaborate on each of the three.

3.1 Fine Structure-aware Sampling Scheme (FSS)

As explained in Section 2.2, DOS is a sampling training scheme that was recently proposed in IntegratedPIFu (Chan et al. 2022b) to succeed the spatial sampling scheme.

To teach pixel-aligned implicit models to reconstruct thin body features (e.g. ears, fingers), DOS constrains sample points' labels to only the camera direction (see Fig. 1b), making it easier for models to interpret and learn from the labels. For example, if a sample point has a label of 0.7 in DOS, it means that a short distance away (in positive or negative camera-direction) from this sample point, we will find a mesh surface. Conversely, if labels are not constrained to camera-direction only, then a label of 0.7 does not really pinpoint where the mesh surface is (as any direction is likely).

But because the sample points' labels are determined by the shortest distance between a sample point and a mesh surface in the camera direction, rather than the shortest distance in any direction, a DOS-trained pixel-aligned implicit model is trained to only identify surfaces that are front-facing and back-facing. Lateral surfaces become difficult for a DOS-trained model to identify. Thus, DOS suffered from problems such as high-frequency, wavy artefacts on the side-facing surfaces of its reconstructed human meshes. Moreover, badly reconstructed side-facing surfaces also mean thin body features (e.g. ears, fingers) tend to get incorrect shapes.

To overcome the drawbacks of DOS, we propose FSS. FSS solves the above issues with its **5 key features**.

1. Twinned Sample Points In order to have labels that are determined by the shortest distance in any direction (unlike

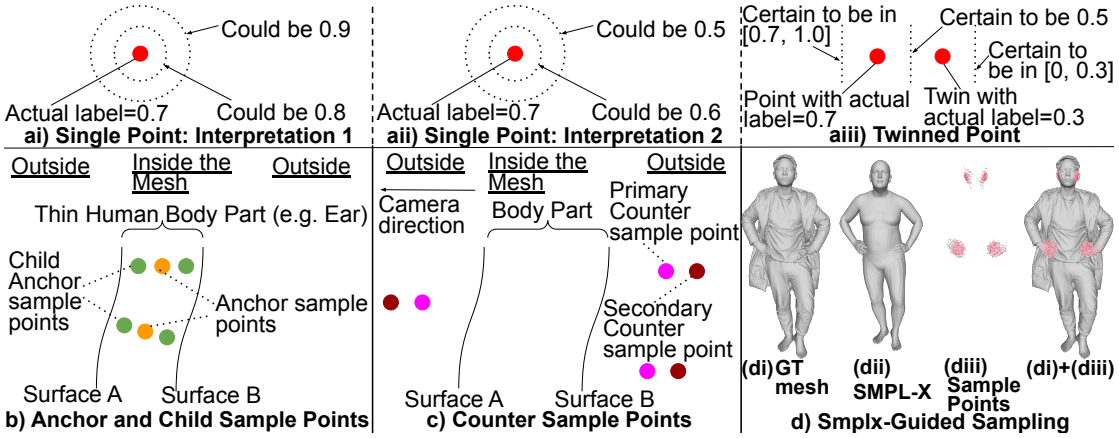


Figure 4: Four of the five key features in FSS. (a) Twinned Sample Points (b) Anchor Sample Points (c) Counter Sample Points (d) Smplx-guided Sampling.

DOS) and yet still makes it easy to pinpoint the actual location of mesh surfaces, FSS proposes twinned sample points.

In both spatial sampling and DOS schemes, the sample points are generated independently from one another. But in FSS, our sample points will each have a corresponding twin (see Fig. 1c). As explained earlier, a sample point with a label determined by shortest distance from any direction cannot unambiguously pinpoint the location of a nearby mesh surface. This is illustrated in Fig. 4ai & Fig. 4aii, which show the same sample point having two different interpretations.

However, this can be overcome if we have a pair of sample points (twins) that are equidistant away from the nearby mesh surface (See Fig. 4aiii). In other words, the midpoint of the pair of twins automatically pinpoints the location of a mesh surface. Being able to precisely pinpoint the mesh surface is pivotal for learning to reconstruct thin body features like ears or fingers. Moreover, as seen in last row of Fig. 1a-b, only a few sample points are near thin surfaces (i.e. ears) each time. It is thus vital to make full use of every point.

2. Proximity-adaptive Displacement Other than twinned sample points, FSS also introduces the concept of a proximity-adaptive displacement. Proximity-adaptive displacement is motivated by the observation that pixel-aligned implicit models trained with spatial sampling scheme rarely encounter any sample point that fall within thin surfaces like ears or fingers. This is because, as shown in Fig. 1, a sample point is generated by displacing a mesh surface point with some random noise. Thus, compared to mesh surface points on thicker surfaces, mesh surface points on thin surfaces are much more likely to be displaced into a region outside of the mesh (See the first and second rows in Fig. 1a and 1b). With the majority of sample points that are near thin surfaces labelled as 'outside' (i.e. label$<$0.5), this naturally encourages the pixel-aligned implicit models to predict 'outside' for any sample points near thin surfaces, resulting in missing ears, fingers, and other thin body parts in reconstructed meshes. Proximity-adaptive displacement overcomes the issue by adjusting how much to displace the mesh surface point based on thickness of the surfaces (see top two rows in Fig. 1c).

3. Anchor Sample Points In addition, Fig. 1c (second row) shows another problem with capturing thin body features when we use continuous labels. In almost all cases, the maximum label of a sample point inside a thin body feature will never be close to 1.0 due to how thin the body feature is. In our experiments, we find that the maximum label tends to be around 0.60. Thus, with thin body features, the range of sample points' labels is distorted from [0, 1] to [0, $\sim$0.60].

There is thus a bias towards predicting values less than 0.5 (i.e. 'outside') for any sample point close to thin surfaces. To solve this, we can naively increase the number of sample points that are inside thin surfaces. However, that would introduce a large number of additional sample points. Thus, in order to correct the bias efficiently, we propose the idea of anchor sample points (illustrated in Fig. 4b). Anchor sample points are sample points that are at the deepest location inside a thin body part. These points will have the highest label (e.g. 0.60) that is possible in the thin body part. From an anchor sample point, we generate a number of **child anchor sample points**. Child anchor sample points are sample points that are in between an anchor sample point and the nearest mesh surface (See Fig. 4b). In addition to correcting the bias, anchor sample points and child anchor sample points are also important for indicating to the pixel-implicit aligned model where the max label value would be reached, and that the model should start predicting a label value that is lower than the max label value for any sample point located in between an anchor sample point and a child anchor sample point (i.e. clearer supervision signals for the model).

4. Counter Sample Points FSS also includes sample points that deter floating artefacts. Floating artefacts often appears either in front or behind a reconstructed mesh, where 'front' or 'behind' is determined by the camera-direction. Thus, we propose counter sample points. Counter sample points are points that are either in front or behind the mesh. They are always outside of the mesh and are used to discourage a pixel-aligned implicit model from predicting floating artefacts in regions that are actually empty. We further enhance this with the concept of a "twinned" counter

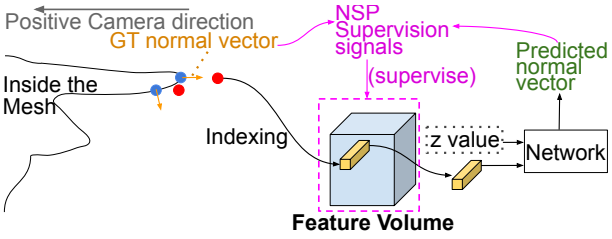


Figure 5: Illustration of NSP, an extension to FSS.

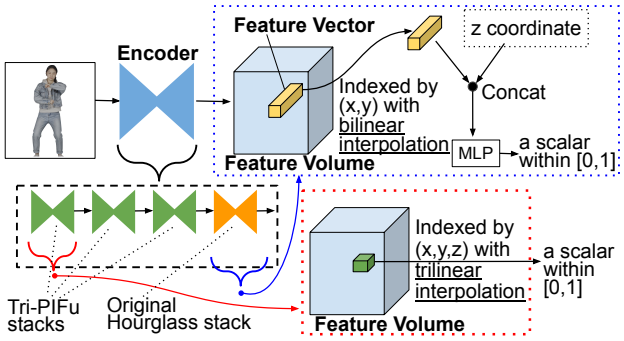


Figure 6: Illustration of our Tri-PIFu architecture

points. We pair a counter point (i.e. a primary counter point) with a secondary counter point (see Fig. 4c). The secondary counter point is located further away (in camera direction) from the mesh surface compared to the primary counter point, and thus the secondary counter point will have a lower label than primary counter point. Twinned counter points indicate the direction of where the labels should be decreasing in, thereby providing clearer supervision signals.

5. Smplx-guided Sampling Finally, the human body is a complex and convoluted structure. To pixel-aligned implicit models, some body parts will be easier to reconstruct than others. Existing sampling training schemes (i.e. spatial sampling and DOS) do not discriminate between these different body parts and do not assign less sample points to body parts that are easier to reconstruct. Examples of such body parts include the neck, lower leg, and chest, which are made of mostly flat surfaces and are easy for models to reconstruct.

Thus, FSS introduces Smplx-guided sampling, which allows us to select which human body parts we want to focus on (and assign more sample points there). In our context, we have FSS to focus on the important thin human body features (e.g. ears and fingers) that we are interested in reconstructing. Smplx-guided sampling requires the use of a groundtruth SMPL-X mesh (Pavlakos et al. 2019) (only during training). Using the SMPL-X mesh, we can identify the location of ears and fingers of the groundtruth clothed human mesh (See Fig. 4d). We then generate a higher concentration of sample points from those locations so as to help a pixel-aligned implicit model focus on these body parts. As illustrated in the third row of Fig. 1, Smplx-guided sampling reduces the density of sample points on easy-to-reconstruct parts like the human neck and increases the density of sample points on hard-to-reconstruct parts like the human ears.

3.2 Exploiting Normals of Sample Points (NSP)

The sample points' normals, which can be computed, are regrettably not used in existing sampling training schemes. The closest related work that tried is PHORHUM (Alldieck, Zanfir, and Sminchisescu 2022). PHORHUM slightly modifies the spatial sampling scheme such that normals of sample points that lie **exactly on the mesh surface** are used to regularize the training process. In contrast, our proposed idea will use the normals of **all** our sample points.

In reality, the normals of sample points are very useful for inferring the underlying structure of a human body mesh. Hence, as an extension to our FSS scheme, we exploit the normals of sample points to further enhance the training process of pixel-aligned implicit models.

We define the normal of a sample point as the normal of the mesh surface point that is nearest to that sample point in the camera direction.

We use the normals of sample points to supervise and enhance the representations in the Feature Volume in Fig. 3. To do so, we use the sample points to index the Feature Volume (refer to Fig. 5) and obtain a set of feature vectors. The feature vector indexed by a sample point, along with the z coordinate of this sample point, will be fed into a small neural network that will attempt to predict the normal vector of this sample point. The mean squared error between the predicted and groundtruth normal vectors is computed, and the derived supervision signals are used to train the representations in the Feature Volume. The small neural network only acts as a dummy as it is only used during training. The impact of NSP is supported by our ablation studies that will be shown later. Refer to our arXiv version for more details.

3.3 Mesh Thickness Loss (MTL) Supervision

Works, such as (Hong et al. 2021; Peng et al. 2021; Chan et al. 2022b), observed that pixel-aligned implicit models tend to produce reconstructed meshes with implausible thickness. An intuitive solution to this problem is to teach the pixel-aligned implicit model to directly learn the thickness of different body parts (at different orientations). Thus, FSS proposes a mesh thickness loss (MTL) signal to do that.

It is, however, not trivial to introduce this loss. This is because predicted meshes (and their thicknesses) are not available during train time. During training, for each image, a pixel-aligned implicit model will process and predict the labels of (typically) only 8000 sample points. During testing, to predict a human mesh, the same model would need to process and predict for (typically) 256^3=16,777,216 sample points. Thus, the time cost alone makes it infeasible to compute the mesh thickness of predicted meshes at train time. Moreover, since the marching cube algorithm used to produce the predicted mesh is not differentiable, it is unclear how we could backpropagate the error between predicted meshes' thickness and groundtruth meshes' thickness.

To solve the issue, we first introduce an architectural modification of the PIFu framework. In Fig. 3, we see that the Encoder is a Stacked Hourglass Network consisting of 4 stacks.

In PIFu, unlike what is shown in Fig. 3, 4 Hourglass stacks (in orange) are used. Tri-PIFu stack (in green) is a new type of stack that we proposed to replace 3 of the 4 Hourglass stacks. Each of the 4 stacks, regardless of their types, will produce a separate Feature Volume (see Fig. 6).

As shown by the blue dotted box in Fig. 6, an Hourglass stack would produce a Feature Volume that is indexed via bilinear interpolation to retrieve feature vectors. Feature vectors are concatenated with z coordinates and fed into a MLP that predicts occupancy values that range from 0 to 1.

In contrast, as shown by the red dotted box in Fig. 6, a Tri-PIFu stack does not use the MLP, and we do not concatenate the feature vector with the z coordinate. Instead, we use the sample point's (x,y,z) coordinates to index the Feature Volume via **trilinear interpolation** to directly retrieve an occupancy value that ranges from 0 to 1.

The Feature Volume produced by a Tri-PIFu stack is interpreted as a 3D space with shape of (D, H, W), where D, H, W represents its depth, height, and width respectively. Tri-PIFu's aim is to model an implicit function of a human mesh surface **inside** this 3D space. To do so, we applied sigmoid activation function on the Tri-PIFu stack's outputs, thereby ensuring every value or element in the Feature Volume falls within [0,1]. This ensures that any sample point that index the Feature Volume via trilinear interpolation will always obtain a value within [0,1]. Once trained, the stack will model an implicit function (of a predicted human mesh surface) inside its Feature Volume. Aside: The predicted surface is at the 0.5 (not 0) level-set of the implicit function.

With the implicit function in the Feature Volume, it is easy to get a metric of mesh thickness at any (x,y) position. We simply sum up the Feature Volume in the z (or D) dimension to obtain a 2D plane of shape (H, W). Each value or element in this 2D plane would be a consistent approximation of the mesh thickness at that (x,y) position. We refer to this 2D plane as the mesh thickness plane.

We can easily compute the groundtruth mesh thickness plane using the groundtruth mesh. The mean squared error between the groundtruth and predicted mesh thickness planes is then computed during training. This loss signal is our MTL signal. MTL signals are an extension to our FSS scheme as the signals provide additional supervision to the training process. MTL will be ablated later. More on MTL (e.g. why we still use 1 Hourglass stack) in Supp. Mat.

4 Experiments

4.1 Datasets

In our experiments, we use the THuman2.0 dataset (Yu et al. 2021) as the training set for both our models and other competing models. THuman2.0 dataset contains 526 high-quality scans (or meshes) of ethnic Chinese human subjects. We use a 80-20 train-test split of these meshes. For each training mesh, we first render a RGB image of the mesh's front view using a weak-perspective camera. We then render 10 other images by evenly fanning out (i.e. changing the yaw) from the first RGB image, in both clockwise and counter-clockwise directions. This set-up is similar to the ones used in IntegratedPIFu (Chan et al. 2022b) and S-PIFu (Chan et al. 2022a), which are benchmarks that we aim to compete against in our experiments.

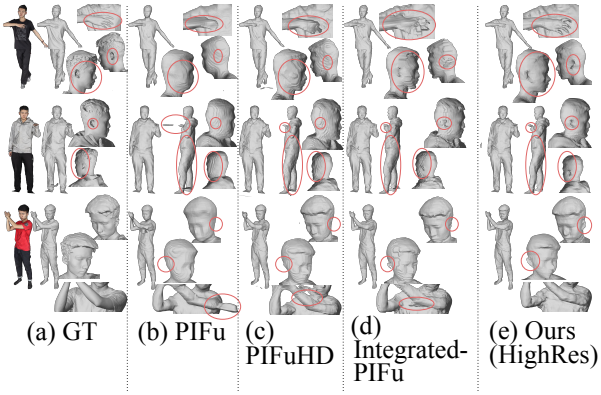


Figure 7: Evaluation with SOTA models. (Please zoom in)

In addition, we made use of the BUFF dataset (Zhang et al. 2017) to evaluate all the models. None of the models are trained using the BUFF dataset. Like what the authors did in IntegratedPIFu, we conducted systematic sampling (based on sequence number) on the BUFF dataset, giving us 101 human meshes to be used for evaluating the models. By using systematic sampling, we avoided getting meshes that have both the same human subject and the same pose.

4.2 Comparison with State-of-the-art

We trained two models. The first uses the architecture in Fig. 3. The second is same as the first but uses the HRI component proposed by IntegratedPIFu to incorporate high-res images. We compare our models against existing models on single-view clothed human reconstruction. The existing models include PIFu (Saito et al. 2019), PIFuHD (Saito et al. 2020), and IntegratedPIFu (Chan et al. 2022b). To be fair to all models, predicted normal maps are used in all models, including PIFu. Following (Saito et al. 2019, 2020; Chan et al. 2022b), we use Chamfer distance (CD), Point-to-Surface (P2S), and Normal reprojection error (NML) as the metrics in our quantitative evaluation. In addition, to compare our methods against methods that use SMPL-X meshes as priors, we added S-PIFu (Chan et al. 2022a) in our quantitative evaluation. Due to space constraint, S-PIFu is compared qualitatively with ours in our Supp. Mat. We also compared against ICON (Xiu et al. 2022) in our Supp. Mat.

Qualitative Evaluation We evaluate the models qualitatively in Fig. 2 and Fig. 7. The figures show our method reconstructs fine, thin features like fingers, hands, and ears correctly. Unlike existing models, our method does not produce high-frequency wavy artefacts or unnatural protrusions (i.e. implausible mesh thickness) on reconstructed meshes.

Do refer to our arXiv version for a qualitative evaluation based on real Internet images sourced from Shutterstock.

Quantitative Evaluation We also evaluate our models quantitatively in Tab. 1. From the table, we can see that our low-resolution model, 'FSS (Ours)', is able to outperform the existing models (first five rows) in all except one column.

Methods	H	THuman2.0 Test Set			BUFF		
		CD (10^{-4})	P2S (10^{-4})	NML (10^{-2})	CD (10^{3})	P2S (10^{3})	NML (10^{-2})
PIFu	×	5.314	4.933	8.149	2.089	1.977	6.243
PIFu + DOS	×	5.943	5.794	8.224	2.214	2.077	6.393
S-PIFu	×	5.000	4.728	8.079	2.140	2.011	6.178
ICON	×	6.862	7.808	12.96	2.757	3.137	8.569
PIFuHD	✓	5.267	4.688	7.667	2.177	2.072	5.967
IntegratedPIFu	✓	5.172	4.276	7.620	2.061	1.778	5.935
FSS w/o NSP, MTL	×	5.004	3.965	8.052	2.001	1.737	6.056
FSS w/o MTL	×	4.931	3.916	7.780	1.947	1.652	5.838
FSS w/o NSP	×	4.923	3.969	7.995	1.957	1.678	5.941
FSS	×	**4.833**	**3.854**	7.800	**1.943**	**1.576**	5.812
HRI + FSS	✓	4.896	3.905	**7.615**	1.945	1.611	**5.715**

Table 1: Our models (last 5 rows) vs SOTA methods. ('H' indicates if a 1024x1024 RGB image is required and used. By default, a 512x512 image is used. HRI=High-Resolution Integrator proposed by IntegratedPIFu, FSS=Fine Structure-aware Sampling, NSP=Trained with Normals of Sample Points, MTL=Trained with Mesh Thickness Loss)

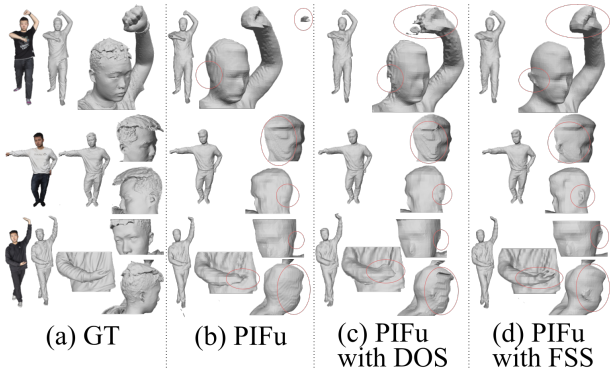


Figure 8: Evaluating our FSS scheme (w/o NSP w/o MTL)

The failure to outperform PIFuHD and IntegratedPIFu in the 'NML' metric for the THuman2.0 dataset can be attributed to PIFuHD's and IntegratedPIFu's use of a higher resolution input RGB image. Indeed, with our high-resolution model, 'HRI + FSS (Ours)', we are able to significantly outperform the existing models in all metrics for both datasets.

4.3 Ablation Studies

Evaluation of FSS scheme (w/o NSP w/o MTL) We compare a PIFu trained with spatial sampling scheme, a PIFu trained with DOS scheme, and a PIFu trained with our FSS (w/o NSP w/o MTL). We present the qualitative results in Fig. 8. We find that with our FSS (w/o NSP w/o MTL), a PIFu does not produce wavy, noisy artefacts and is much better at reconstructing thin features (e.g. ears, fingers).

Quantitatively, we can compare the rows of 'PIFu', 'PIFu + DOS', and 'FSS w/o NSP w/o MTL (Ours)' in Tab. 1. These rows show that a PIFu trained with FSS greatly outperforms a PIFu trained with either spatial sampling or DOS.

Do refer to our arXiv version for a further ablation study exploring the effects of including and excluding each of the five key features of FSS mentioned in Section 3.1.

Evaluation of NSP To evaluate the usefulness of capitalizing on Normals of Sample Points (NSP) during the training process, we compare a PIFu trained without NSP with a PIFu that is trained with NSP. See qualitative results in Fig. 9. In our arXiv version, we also showed that NSP improves the quantitative results for all metrics.

Evaluation of MTL To evaluate our Mesh Thickness Loss (MTL), we compared a PIFu with a PIFu modified to use MTL. Qualitatively, the results in Fig. 10 show that MTL improves structural accuracy of outputs. Similar improvement is also observed quantitatively (see our arXiv version).

5 Conclusion

We have proposed Fine Structured-Aware Sampling (FSS), a novel sampling training scheme to train pixel-aligned implicit models for clothed human reconstruction. FSS also shows how Normals of Sample Points (NSP) and a Mesh Thickness Loss signal (MTL) can be capitalized to further improve results. See our arXiv version for more details.

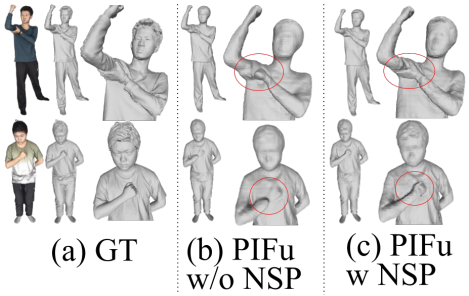


Figure 9: Effect of training with sample points' normals

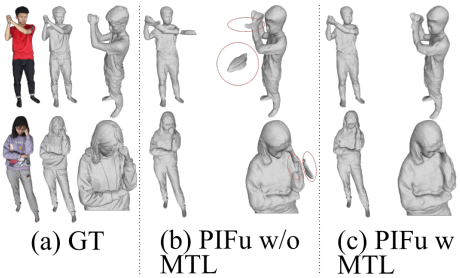


Figure 10: Evaluation of our mesh thickness loss signal

Acknowledgements

This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

References

Alldieck, T.; Magnor, M.; Bhatnagar, B. L.; Theobalt, C.; and Pons-Moll, G. 2019. Learning to reconstruct people in clothing from a single RGB camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1175–1186.

Alldieck, T.; Zanfir, M.; and Sminchisescu, C. 2022. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1506–1515.

Bhatnagar, B. L.; Tiwari, G.; Theobalt, C.; and Pons-Moll, G. 2019. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5420–5430.

Chan, K.; Lin, G.; Zhao, H.; and Lin, W. 2022a. S-PIFu: Integrating Parametric Human Models with PIFu for Single-view Clothed Human Reconstruction. In *Advances in Neural Information Processing Systems*.

Chan, K. Y.; Lin, G.; Zhao, H.; and Lin, W. 2022b. Integratedpifu: Integrated pixel aligned implicit function for single-view human reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, 328–344. Springer.

Collet, A.; Chuang, M.; Sweeney, P.; Gillett, D.; Evseev, D.; Calabrese, D.; Hoppe, H.; Kirk, A.; and Sullivan, S. 2015. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4): 1–13.

Hong, Y.; Zhang, J.; Jiang, B.; Guo, Y.; Liu, L.; and Bao, H. 2021. Stereopifu: Depth aware clothed human digitization via stereo vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 535–545.

Jiang, B.; Zhang, J.; Hong, Y.; Luo, J.; Liu, L.; and Bao, H. 2020. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, 18–35. Springer.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2252–2261.

Lombardi, S.; Saragih, J.; Simon, T.; and Sheikh, Y. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4): 1–13.

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

Natsume, R.; Saito, S.; Huang, Z.; Chen, W.; Ma, C.; Li, H.; and Morishima, S. 2019. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4480–4490.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 10975–10985.

Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9054–9063.

Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2304–2314.

Saito, S.; Simon, T.; Saragih, J.; and Joo, H. 2020. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 84–93.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Xiu, Y.; Yang, J.; Tzionas, D.; and Black, M. J. 2022. ICON: implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13286–13296. IEEE.

Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; and Liu, Y. 2021. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*.

Zhang, C.; Pujades, S.; Black, M. J.; and Pons-Moll, G. 2017. Detailed, Accurate, Human Shape Estimation From Clothed 3D Scan Sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zheng, Z.; Yu, T.; Liu, Y.; and Dai, Q. 2021. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*.