REF-SHARP: REFined face and geometry reconstruction of people in loose clothing*

Snehith Goud Routhu
IIIT Hyderabad
India
snehith.goud@research.iiit.ac.in

Sai Sagar Jinka
IIIT Hyderabad
India
jinka.sagar@research.iiit.ac.in

Avinash Sharma
IIIT Hyderabad
India
asharma@iiit.ac.in

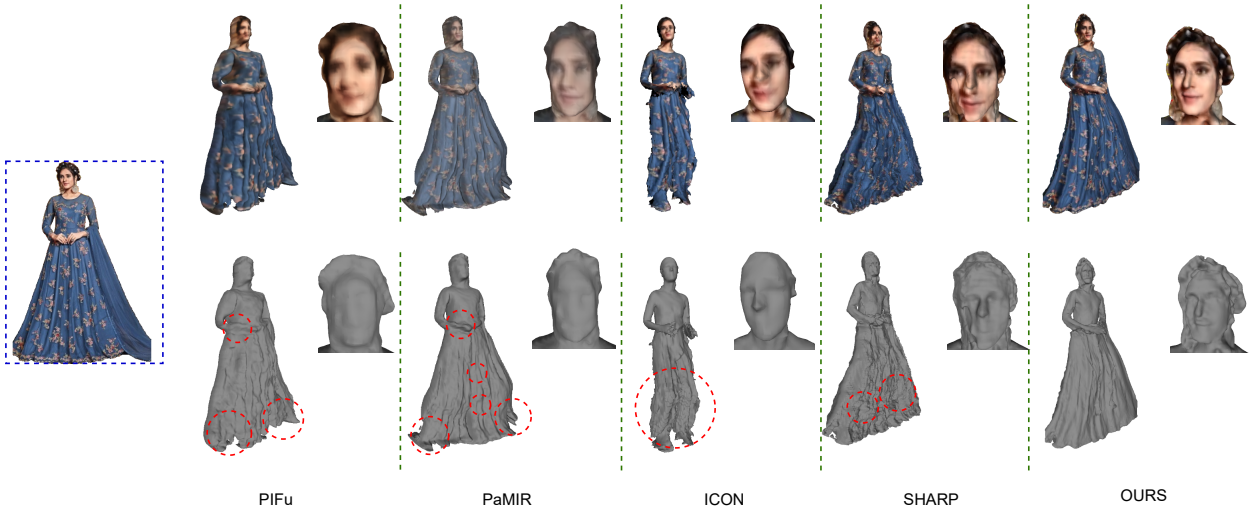

Figure 1: Reconstruction from in-the-wild images using PIFu [39], PaMIR [51], ICON [46] and SHARP [38] and ours. Our method predicts high-fidelity geometry reconstruction along with a consistent appearance in face and loose clothing regions.

ABSTRACT

In this paper, we address the problem of monocular 3D human reconstruction with an acute focus on the challenge of recovering person-specific facial geometry as well as suppressing surface noise, specifically addressing the issue of false geometrical variations caused by textural edges. Most of the existing state-of-the-art methods in this domain fail to address these challenges. More specifically, we propose to integrate facial and wrinkle map priors in a learning-based framework to improve the quality of full-body 3D reconstruction from monocular images. By incorporating facial prior, we recover person-specific identity unlike many of the existing methods which rely on parametric shape models. Similarly, the wrinkle map prior enables our network to alleviate the challenge of false geometrical variations caused by high-frequency textural details present in the input image. We evaluate our method on publicly available datasets & in-the-wild internet images with loose clothing and report superior performance both qualitatively and quantitatively when compared with SOTA methods.

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**.

KEYWORDS

3D Human Reconstruction, Face Reconstruction

ACM Reference Format:
Snehith Goud Routhu, Sai Sagar Jinka, and Avinash Sharma. 2022. REF-SHARP: REFined face and geometry reconstruction of people in loose clothing. In *Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'22), December 8–10, 2022, Gandhinagar, India,* Soma Biswas, Shanmuganathan Raman, and Amit K Roy-Chowdhury (Eds.). ACM, New York, NY, USA, Article 22, 10 pages. https://doi.org/10.1145/3571600.3571622

1 INTRODUCTION

The 3D modeling of humans is interesting and active area of research in computer vision which has tremendous applications in VR/AR, gaming, image, and video editing, tele-presence, virtual try-on, to name a few. With the advent of deep learning, 3D human body reconstruction from monocular RGB images [21, 38–40, 51] is feasible eliminating the requirement of expensive multi-camera

*Produces the permission block, and copyright information

ICVGIP'22, December 8–10, 2022, Gandhinagar, India

ACM ISBN 978-1-4503-9822-0/22/12.
https://doi.org/10.1145/3571600.3571622

calibrated setups [33, 42]. However, the problem is ill-posed in nature because of challenges which include self-occlusion, loose clothing, skewed viewpoints, pose and shape variation, etc. The two key challenges that are largely remain unaddressed by existing literature are fidel reconstruction of high frequency geometrical detail in the facial region as well as complex loose clothing covering large parts of the body. High fidelity reconstruction of faces largely enhances the realism and identity of the digitized human models. However, it appears in a smaller area of the input image making the reconstruction more difficult for existing methods. On the other hand, although clothing covers significant portion of the human body, it is difficult to model large space of garment designs, specifically in case of loose clothing where high frequency geometrical details (owing to folds and curls) prevails in a highly unstructured manner. More importantly, cloths also consist of high frequency textural details, making it difficult to distinguish textural edges with geometrical edges in a monocular reconstruction setup.

Existing monocular 3D human body reconstruction methods can be broadly classified as parametric and non-parametric methods. The first class of methods [21, 24, 32, 34] regress pose and shape parameters of statistical model (SMPL) [28]. Nevertheless, they fail to capture fine geometrical details as the reconstructed geometry is bound within the parametric model space. Note that most of the parametric templates are modeled as naked human body and hence fail to deal with loose clothing scenario.

The other class of non-parametric methods [4, 31, 39, 44, 45] are not constrained by the body prior and use non-parametric representations to infer detailed geometry beyond basic body shape and pose. Recently, deep implicit function learning [15, 39, 40] techniques witnessed increased attention. These methods train multi-layer perceptrons to estimate dense, continuous signed distance fields. from which 3D mesh is reconstructed via Marching Cubes [29]. Another work, [19] proposed to represent the human body by predicting multiple peeled depth maps. These methods fail to ensure the prediction of physically plausible body shapes/poses (like consistent geometry of arms) in the reconstructed mesh as there are no global body shape/pose constraints. Some of the very recent works [16, 17, 38, 51, 52] have addressed this problem by incorporating a body prior i.e. SMPL into the reconstruction framework. Nevertheless, these methods typically yield reconstruction with facial geometrical details incorporated from the input SMPL prior while the person-specific facial details are not retained. This leads to loss of identity as well as misalignment in geometry and appearance of facial features. There are few methods that specifically focus on accurate 3D face reconstruction [10, 12, 14, 27, 50] but does not model the complete human body. Very recently, [8] proposed to integrate facial features from 3DMM model [12] into implicit function learning approaches. However, the method is not able to perform better on the remaining parts of body as there is no explicit body prior. Additionally, all these methods are prone to interpret high frequency textural details present in the appearance space of clothes as false geometrical edges/details thereby yielding inconsistent and noisy geometrical reconstruction specifically in the loose clothing. Although, disambiguate textural edges with geometrical edges is very hard problem but these methods entirely neglect this challenge during reconstruction. Hence, there is an acute need of framework which recover accurate and consistent geometrical/appearance reconstruction of both face and complex in-the-wild loose clothing while modelling 3D human body.

In this paper, we propose REF-SHARP, a novel 3D reconstruction framework for recovering people in loose clothing from a monocular image while recovering fine 3D geometrical details of the face. Additionally, we also attempt to alleviate false geometrical edges caused by textural details in clothes. We improve upon SHARP [38], a SOTA method which uses PeeledHuman representation [19] along with SMPL prior for the prediction of human in loose clothing. However, similar to other methods Figure 1, SHARP also yields low-quality facial geometry as well as false geometrical details misled by textural edges. Thus, we propose to provide high quality pixel-aligned depth prior for face region along with prediction of wrinkle maps capturing geometrical edges to improve upon the aforementioned limitations of SHARP. More specifically, we attempt to recover finer facial details by predicting high resolution pixel-aligned depth in the face region and integrate this with SMPL peeled depth prior. This ensures that person-specific facial geometrical details will be preserved in the final reconstruction after learning-based fusion. Additionally, we propose to learn and predict wrinkle map prior by regressing over geometrical edge maps and concatenate this with face modified SMPL prior to and use suppression and normal loss to ensure suppression of false geometrical edges.

We evaluate our method on publicly available THUman2.0 [49] and 3DHumans [38] datasets and report superior performance over the SOTA methods. To summarize our contributions are: we propose REF-SHARP where we reconstruct high-fidelity geometrical details in the face while recovering full body reconstruction under loose clothing scenarios.

2 RELATED WORK

2.1 Parametric Reconstruction Methods

With the emergence of statistical human models such as SMPL[28] and SCAPE [3] interest has shifted to estimating these models pose and shape from a single image using deep learning methods [7, 11, 20, 21, 25, 32, 35]. [21, 34] tries to optimize the pose and shape parameters of statistical human models for *e.g.*(SMPL) by matching with image features obtained from CNN. The commonly used features are 2D joints [7, 21, 32], 2D joints with silhouettes [11, 25].

Methods proposed in [2, 5] deform the statistical models by adding displacements over the surface to obtain geometric details and clothing to a certain extent. It is to be noted that under this SMPL[28] plus displacement setup only tight clothing can be modeled as the loose garments such as skirts and robes have a surface topology different from the body and are beyond the representation range. [53] models the fine geometric details as free-form 3D deformations applied on the parametric body model. [23] improves the estimation of the prediction by introducing model optimization in the training loop, [30] combines local and global features to estimate fine body poses. Nevertheless parametric models can only capture minimalistic clothing where the clothing is tight fitted with the body and fails to represent humans with loose clothing.

[6, 36] use separate templates for body and garments and bind garment vertices to the parametric body model which becomes

difficult to represent very loose clothing as sarees, robes, and skirts as the topology of the garment is constrained by the binding with the body model.

2.2 Non-Parametric Reconstruction Methods

Volumetric regression based methods [44, 45] estimate the occupancy of the voxels in the volumetric space using deep neural networks. These methods are computationally costly for higher resolution as voxel representation is memory intensive, high frequency details are not captured due to lower resolution (typically 128). [39, 47] combines the pixel aligned 2D local features extracted using deep convolutional networks with the implicit representation. Nevertheless PIFu [39] suffers from feature ambiguity problem due to multiple query points mapping to the same 2D image features upon weak perspective projection and lacks global shape robustness. PIFuHD [40] is another variation of PIFu which generates human meshes from high resolution images. GeoPiFu [15] attempted to resolve the feature ambiguity of 3d points projecting to the same image feature by combining U-Net based volumetric features with the pixel aligned features. However, this method is computationally intensive during training and inference. An alternative set of non-parametric approaches attempt to model 3D objects/scenes as sparse layered representation. PeeledHuman [19] proposes a sparse2D representation by posing the problem as an extension to ray tracing, they model the 3D surface by performing ray intersection with the surface and storing them as the peeled depth and rgb maps.

2.3 Prior based non-parametric Methods

DeepHuman [52] uses the SMPL as a prior to reconstruct the clothed body volume and tries to further refine the surface details using image features. However, their method fails to recover high-quality geometric surface details owing to the resolution limitation of the regular occupancy volume. ARCH [17] proposed a Semantic Deformation Fields (SemDF) based approach where the query points are sampled around the body in a canonical space (A-pose), an implicit surface is learned in the canonical space, and deformed using SemDF to match the pose in the input. However, it fails to generate accurate results, especially in the scenarios of loose clothing. PaMIR [51] proposes to condition the implicit field on the SMPL prior, they do this by combining the 2D image features with SMPL volume features while querying. SHARP [38] proposes to model the reconstruction as two tasks (1) deform the SMPL body prior peeled maps in order to obtain fine-grained geometric surface details (2) directly regress the loose clothing as auxiliary peeled depth maps and combine both of them to obtain 3D reconstruction.

2.4 Face reconstruction methods

Similar to SMPL ([28]) for full body, 3D Morphable Models (3DMM) [12] is proposed to model face in parametric representation. Methods proposed in [14, 27, 41, 43] estimate parameters of 3DMM. Implicit functions are combined with 3D morphable models in [37, 48]. Full head models including hair are recovered in [9, 13, 26].

3 BACKGROUND

PEELED-HUMAN representation This is a sparse 2D representation of 3D objects modeled as Peeled Depth and RGB maps. The 3D Human mesh is placed in a virtual world and rays are passed from the camera to intersect with the mesh. The primary set of rays intersecting with the surface is captured as depth map d_1 and RGB map r_1 depicting the visible surface details nearest to the camera. The rays are then further extended beyond the first intersection to hit the next intersecting surface. The corresponding depth and RGB values of the i_{th} layer are represented by d_i and r_i, refer to [19] for further understanding. [19] demonstrate that 4 layers are sufficient to handle self-occlusions and common body poses though the method can be extended to multiple layers.

4 METHOD

Our proposed REF-SHARP is divided into three key modules as shown in Figure 2. The input monocular RGB image is fed to the *"Face Prior Module"* where we predict pixel-aligned high resolution depth map of the cropped face region and overlay it over SMPL peeled prior. In parallel, we predict wrinkle map that models prominent geometrical edges over the surface as outlined in *"Wrinkle-map Prior Module"*. The generated face+SMPL prior and wrinkle maps along with input RGB image are then fed to the *"Reconstruction & Fusion Module"* which reconstructs full body. This module predicts auxiliary and residual depth peel maps, RGB peel maps for the full body, and single layer face residual peel map. These maps are subsequently *fused* to get final peeled maps that are jointly back-projected to get high resolution vertex colored point cloud representing full body. This dense surface point cloud is further converted to mesh representation using Poisson surface reconstruction[22].

4.1 Face Prior Module

Similar to [38], we initially predict SMPL peeled prior $D^i_{smpl} \in D_{smpl}$. We modify the SMPL peeled prior to obtain D^*_{smpl} by replacing the face region of SMPL (which has inconsistency in the face) in the first layer (D^1_{smpl}) with face predicted from [50].

Since SMPL face is not exactly aligned to actual face, we use off-the-shelf face estimation [50] to achieve pixel-to-pixel consistency with the input image. We perform face detection using an off-the-shelf method [1] to detect face in the image and crop it using the detected bounding box M_b which contains the complete head region (hair to chin). We rescale (bicubic interpolation) the cropped region to 512x512 and mask the cropped image using 68 landmarks detected using the face detection library [1] and feed it to [50] for pixel-wise depth prediction. We rescale the depth prediction back to the original resolution of the face in the image. It is important to note that although we re-scale the face region, we are not using any information from any external source. Finally, we replace the SMPL face in first layer of SMPL prior (D_{smpl}) with the rescaled person-specific face prediction to obtain D^*_{smpl}.

However, [50] predicts depth in an orthogonal fashion, in order to adopt this to the perspective setup and compensate for the details lost during the up-scaling we predict per pixel offsets for the bounding box region(M_b) of D^{*1}_{smpl}. In the next module, we use a

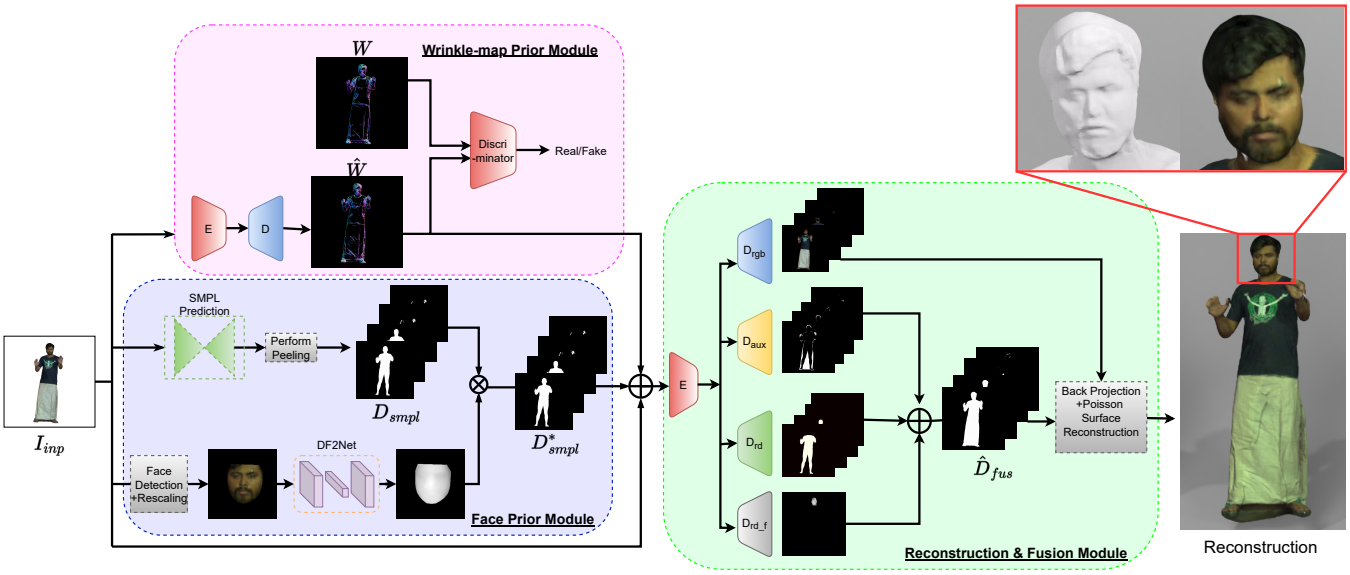

Figure 2: Architecture of the proposed framework.

separate decoder branch for predicting these residual deformations to focus on the region containing the face.

4.2 Wrinkle-map Prior Module

High frequency detail in the input RGB image can be from both geometrical or textural variations. Our wrinkle map representation aims to capture only the high frequency (edges) details that are caused by variation in surface geometry, thereby decoupling these details from textural high frequency details present in the appearance space. However, such prediction can be done reliably only for the first (and visible) peel layer as learning for other layers is hard since the respective body surface is not observed in the input image. We train an image-to-image translation GAN [18] to generate wrinkle map from the input image using ground truth wrinkle map as supervision. We use L1 loss over the predicted wrinkle map w and ground-truth wrinkle map $\hat{w}$ along with the adversarial GAN loss. The final loss function is:

$$loss = w_g * l_{GAN} + w_{L1} * l_{L1} \tag{1}$$

where w_g and w_{L1} are the weighting factors for l_{GAN} and l_{L1}

$$l_{GAN} = E_{I_{inp},\hat{w}}[logD(I_{inp},\hat{w})] + E_{I_{inp},w}[log(1 - D(I_{inp}, w))] \tag{2}$$

$$l_{L1} = |\hat{w} - w|. \tag{3}$$

For generating ground truth wrinkle map $\hat{w}$ we smooth the normal map using a bilateral filter and then calculate the change in normal using Laplacian filter and threshold it to obtain geometric regions which are not smooth.

4.3 Reconstruction & Fusion Module

The generated wrinkle maps ($\hat{w}$) along with D^*_{smpl} are fed as prior to the encoder which predicts auxiliary peel maps $\hat{D}_{aux}$, RGB peel maps $\hat{D}_{rgb}$, residual peel maps $\hat{D}_{rd}$ similar to [38]. We additionally predict residual deformation for face $\hat{D}_{rd\_f}$. The predicted residual deformation $\hat{D}_{rd\_f}$ captures the face region and the hair region is obtained from auxiliary peel maps $\hat{D}_{aux}$. Given the face bounding box obtained from the face detector M_b (obtained in the previous module), we estimate the complete face region peel map $\hat{D}_f$ as a fusion of $\hat{D}_{rd\_f}$, and $\hat{D}^1_{aux}$ as follows:

$$\hat{D}_f = M_b * \hat{D}^1_{aux} + M_b * (\hat{D}_{rd\_f} + D^{*1}_{smpl}). \tag{4}$$

The residual deformation maps are added to D^*_{smpl} to get deformation maps $\hat{D}_{def}$. We finally fuse all the peel maps to obtain the final fused peel maps ($\hat{D}_{fus}$). The final depth peel maps are obtained by fusion of auxiliary ($\hat{D}_{aux}$), face peel map ($\hat{D}_f$), and body peel map ($\hat{D}_{rd} + D^*_{smpl}$). The first layer fused peel map ($\hat{D}^1_{fus}$) (which consists of face region) and remaining layer fuse peel maps ($\hat{D}_{fus}$) can be expressed as:

$$\begin{aligned} \hat{D}^1_{fus} &= M^1_s(1 - M_b) * \hat{D}^1_{def} + (1 - M^1_s) * \hat{D}^1_{aux} + \hat{D_f} \\ \hat{D}^i_{fus} &= M^i_s * \hat{D}^i_{def} + (1 - M^i_s) * \hat{D}^i_{aux} \forall i \in 2, 3, 4 \end{aligned} \tag{5}$$

,where M_s is the mask for $SMPL^*$ peeled prior defined as

$$M^i_s = \begin{cases} 1, & \text{if } D^{*i}_{smpl} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

4.4 Loss Functions

We propose to use the following loss functions: face loss (l_{face}), residual deformation loss (l_{rd}), RGB loss (l_{rgb}) and fusion loss (l_{fus}). Additionally, we use smoothness loss (l_{sm}), normal loss (l_{nl}) for

regularization. For normal loss, we back-project the predicted and ground truth depth maps. The overall loss function used is :

$$L = w_{fus} * l_{fus} + w_{rd} * l_{rd} + w_{sm} * l_{sm} + w_{rgb} * l_{rgb} + w_n * l_{nl} + w_{sup} * l_{sup} + l_{face} \quad (7)$$

, where w_{fus}, w_{rd}, w_{sm}, w_{rgb}, w_n and w_{sup} are the respective weights for l_{fus}, l_{rd},l_{sm}, l_{rgb}, l_{nl} and l_{sup}.

The fusion (l_{fus}), RGB (l_{rgb}), residual deformation (l_{rd}) and smoothness (l_{sm}) losses are similar to [38], i.e. L_1 loss between respective ground truth and predicted peel maps. The remaining loss terms are defined as follows:

$$\begin{aligned} l_{face} = w_f * \left| M_b * D^1_{fus} - \hat{D}_f \right| + \\ w_{f\_rd} * \left| \hat{D}_{rd\_f} - D_{rd\_f} \right| + \\ w_{f\_n} * \left| \hat{N}_f - N_f \right|. \end{aligned} \quad (8)$$

We employ L_1 loss over the $\hat{D}_f$ and the ground truth depth in the face region. We calculate L_1 loss between ground truth residual depth D_{rd_f} and predicted $\hat{D}_{rd\_f}$. We use L_1 loss over ground truth and predicted normals N_f and $\hat{N}_f$ respectively.

In order to compensate for the over smoothing we apply the L1 loss between normal maps of predicted and ground truth depth of the first fused layer. Let N be the normals obtained from D^1_{fus} and $\hat{N}$ be the normals obtained from $\hat{D}^1_{fus}$ then the loss is defined as below.

$$l_{nl} = \left| \hat{N} - N \right|. \quad (9)$$

Based on the wrinkle map we penalise the change in gradients of the depth of non-wrinkle regions as regularisation term to ensure that empty regions in the wrinkle map are locally smooth.

$$M_w = \begin{cases} 1, & \text{if } \hat{w} > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

$$l_{sup} = \left| (1 - M_w) \nabla \hat{D}^1_{fus} \right|. \quad (11)$$

We back-project the $\hat{\mathcal{D}}_{fus}$ and $\hat{R}$ to camera co-ordinate frame assuming the projection is weak perspective in order to obtain the 3D point cloud of the reconstruction. The point-cloud is then post-processed, and further meshified using Poisson Surface Reconstruction (PSR)[22] to generate the final 3D body mesh.

5 EXPERIMENTS AND RESULTS

5.1 Datasets

3DHumans [38]: The dataset is a collection of 200 subjects with diverse body shapes and various clothing styles. This dataset consists of relatively loose clothing (South Asian styles), and also tight clothing such as shirts and pants. The dataset consists of around 150 male and 50 unique female subjects with a database of 200 scans.

THuman2.0 [49]: The dataset is a collection of 500 subjects with high quality 3D scans captured with a DSLR rig. Each subject has around 3-4 poses hence providing us with various poses in the dataset. However, the dataset lacks very loose clothing like long skirts etc.

5.2 Implementation Details

For wrinkle map generation we employ a GAN [18] with our generator being a ResNet generator with a set of down convolution blocks followed by 18 residual blocks then a set of up convolution blocks and our discriminator is a patch-based discriminator. The peeled depth estimation network is an encoder-decoder network. The input to the network is a concatenation of RGB image, SMPL peeled prior and generated wrinkle map. The shared encoder consists of an initial convolution layer of 64 filters of size 7×7 followed by a couple of down sampling layers of filter size 3×3 with stride 2 and respective filters of 128 and 256, in each layer. The down sampled output of $256 \times 128 \times 128$ is then passed through a series of (18) residual blocks. The encoded shared features are then passed to the decoders which predict different outcomes based on the task. The decoders D_{aux}, D_{rd} ,D_{rgb} and D_{rd_f}, consists of 2 upsampling layers of filter sizes 3×3 and channels 128 and 64, respectively. This is followed by a convolutional layer of filter size 7×7. Sigmoid activation is used in D_{aux} ,D_{rd} andD_{rd_f} decoder branches, whereas a Tanh activation is used for the D_{rgb} decoder branch. The D_{rd} output values are scaled to a range of [- 1, 0.5] and D_{rd_f} to [-0.025,0.025] which can be found empirically.

We use Adam optimiser with an initial learning rate of 0.0005. Our network takes around 30 hrs to train for 30 epochs on 4 Nvidia GTX 1080Ti GPUs with a batch size of 4 and w_{fus}, w_{rd}, w_{sm}, w_{rgb}, w_n and w_{sup} are set to $1, 1, 0.1, 0.001, 0.2 and 0.001$ respectively.

5.3 Qualitative & Quantitative Evaluation

We evaluate our model by comparing it with following SOTA methods: PIFu[39], PaMIR[51], ICON[46] and SHARP[38].

Qualitative Results: We present detailed qualitative results generated by our method and comparisons.First, Figure 3 we refer to the results generated by our method on the subjects from 3DHumans and THUman2.0 in Figure 3. Here, the first two columns are on 3DHumans dataset and last two columns are on THUman2.0 dataset. It can be observed that our model can deal with various styles of clothing while predicting high geometric details in the face. Next, we also test the generalization of our model on unseen internet images and report them in Figure 4. It can be observed that our wrinkle map enhances the results qualitatively. It is also observed that our method produces highly detailed faces which can be observed in the third column of Figure 4.

Finally, we qualitatively compare the aforementioned SOTA methods in Figure 5, Figure 1. For each input image, we show full body reconstruction along with face geometry and texture. Accurate face prediction enhances the quality of reconstruction after back projecting the texture. It is observed that in all the SOTA methods, when the texture is back-projected, misalignment is clearly visible. Our method reconstructs face to actual geometry thereby preserving back-projected texture and enhancing the realism.

Quantitative Results: We also evaluate our model quantitatively by comparing it with the aforementioned SOTA methods. In Table 1, we show quantitative results on head region where we trained the models on both 3DHUmans and THUman2 .0 datasets using the same train and test split. We use Chamfer distance (lower the better) and point to surface distance (P2S)(lower the better) as evaluation metrics. We can infer that our method consistently outperform all

	3DHumans [38]		THUman2.0 [49]	
Method	Chamfer($*10^{-5}$)	P2S	Chamfer($*10^{-5}$)	P2S
PIFu	13.9	0.0071	18.1	0.0074
PaMIR	5.9	0.0068	3.6	0.0033
ICON	4.1	0.0045	3.1	0.0039
SHARP	3.1	0.0033	2.7	0.0031
OURS	**2.5**	**0.0028**	**2.4**	**0.0027**

Table 1: Performance of our method in head-only region.

Figure 3: Qualitative results of our method on 3DHumans (columns 1 and 2) and THuman2 (columns 3 and 4) datasets. Top row: input image, 2nd and 4th rows: full-body and head region reconstruction of our method, 3rd and 5th rows: ground truth full body and head only region scans.

the SOTA methods. Our framework which integrates face prior reconstructs accurate facial geometry close to ground truth face

Figure 4: Qualitative results of our method on in-the-wild internet images.

Figure 5: Qualitative comparison on 3DHumans (top row), THuman 2.0 (middle row) and in-the-wild (bottom row) images.

while achieving consistent fusion with face and head region. We also perform quantitative evaluation of full body region in Table 2. In full body reconstruction we are close to SHARP. Nevertheless, we can observe in Figure 1, Figure 4, and Figure 5 (bottom row) our method yield qualitatively far superior results, while generalizing to in-the-wild internet images.

5.4 Ablation

Our model predicts accurate geometry compared to directly overlaying the face prior with the body, as shown in Figure 6. It is to be noted that directly overlaying face prior produces artifacts in the head region. Our proposed residual deformation on the face (D_{rd_f}) seamlessly fuse the face prior with the head region (e.g., near front hairline), as illustrated. In Table 3, we provide a study on the impact of suppression loss (l_{sup}) and normal loss (l_{nl}) on the

	3DHumans [38]	THUman2.0 [49]
Method	P2S	P2S
PIFu	0.00826	0.0091
ICON	0.00822	0.0064
PaMIR	0.00714	0.0049
SHARP	0.00514	0.0055
OURS	**0.00508**	**0.00546**

Table 2: Performance of our method in full body region.

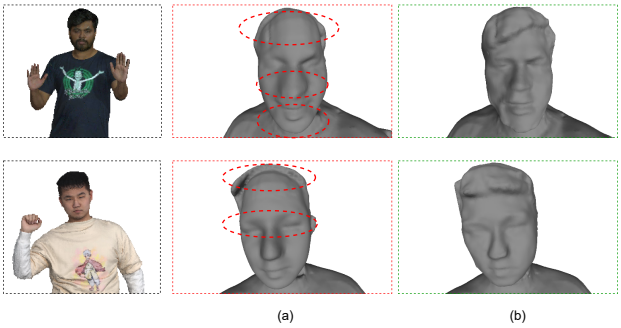

Figure 6: Effect of our networks refinement over face. (a) Directly overlaying fae prior; (b) Our fused reconstruction.

loss terms	P2S
$w/o\ l_{sup}$ & $w/o\ l_{nl}$	0.0055
$w\ l_{sup}$ & $w/o\ l_{nl}$	0.0054
$w/o\ l_{sup}$ & $w\ l_{nl}$	00518
$w\ l_{sup}$ & $w\ l_{nl}$	**0.00508**

Table 3: Ablation study of l_{nl} and l_{sup}.

model's performance. We show P2S estimation on full body (which includes face) on 3DHumans [38] dataset. We observe when we train our model without l_{sup} and l_{nl} losses, we obtain inferior P2S values. Subsequently, we add the suppression loss (l_{sup}) and get slightly improved performance over the previous setup. Further, only adding normal loss (l_{nl}) also significantly boosts the performance where we achieve P2S value of 0.00518. Finally, when trained the model by using both normal and suppression loss to achieve further improvement in P2S, i.e, 0.00508. Hence, the proposed losses contribute to improved performance of our framework. Additionally, we quantitatively evaluate the performance of our method by removing the wrinkle map prior in Table 4. As we can observe, the performance deteriorates in absence of this prior. Finally, we also perform ablative study on the face prior in Table 5 where we can infer that both the P2S error as well as Chamfer distance around head region increases in absence of face prior. We also show qualitatively that there is significant misalignment between the predicted geometry and texture of the face while using SMPL face as prior in our framework whereas our proposed method with pixel-aligned depth face prior achieve much superior texture to geometry alignment as shown in Figure 7.

	P2S
Without wrinkle map prior	0.00517
With wrinkle map prior	**0.00508**

Table 4: Effect of wrinkle map prior.

Prior	Chamfer($*10^{-5}$)	P2S
w/o face prior	2.8	0.0031
w face prior	**2.5**	**0.0028**

Table 5: Ablation study of facial prior on head region.

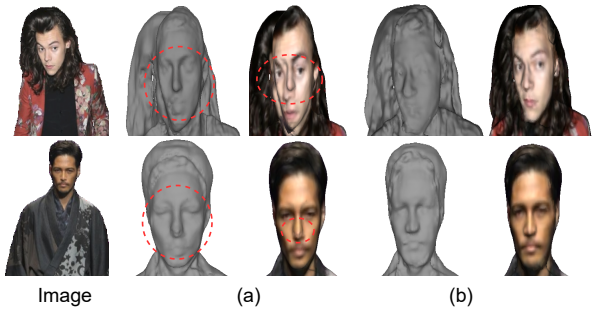

Figure 7: Qualitative ablative analysis of facial prior on internet images. (a) with SMPL face prior (b) with our face prior.

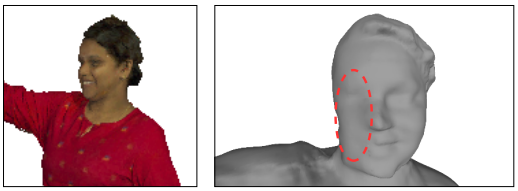

Figure 8: Limitation of our method.

5.5 Limitations

Majority of existing pixel-aligned face depth prediction methods predicts only the frontal faces reliably. Thus, our method can't recover good geometrical reconstruction where the face has skewed pose with large portion of face self-occluded. In this scenario, our proposed model predicts a smooth face in the occluded region as illustrated in the Figure 8. This can be resolved partially if we use a full head parametric model which provides prior for the occluded regions of the face. Nevertheless, full head models also suffers from recovering accurate person-specific face in extreme self-occlusion cases. Alternatively, we can also model the occluded regions in a generative fashion to recover hidden region which can be explored as part of the future work.

It is important to note that in monocular reconstruction setup (with fixed illumination), it is an ill-posed problem to absolutely discriminate between geometrical and textural edges. Hence, our proposed wrinkle map formulation is also susceptible to failure as it is largely dependent on training data distribution. Thus, it will be interesting to explore a solution in the multi-view (varying

illumination) or temporal learning setup where it can be easy to differentiate between these geometrical and textural edges.

6 CONCLUSION

Predicting accurate 3D face and body largely enhances the realism of 3D human body models. In this paper, we proposed a novel reconstruction framework where we incorporate facial prior and wrinkle map prior to recover detailed geometry of face and body in people wearing loose clothing. We demonstrated results on in-the-wild settings by training our model with publicly available datasets.

REFERENCES

[1] [n.d.]. Face Recognition. https://github.com/ageitgey/face_recognition.

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. 2019. Tex2Shape: Detailed Full Human Body Geometry from a Single Image. In *IEEE International Conference on Computer Vision (ICCV)*.

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: shape completion and animation of people. *ACM Trans. Graph.* 24 (2005), 408–416.

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2020. LoopReg: Self-supervised Learning of Implicit Surface Correspondences, Pose and Shape for 3D Human Mesh Registration. In *Neural Information Processing Systems (NeurIPS)*.

[5] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. 2019. Multi-Garment Net: Learning to Dress 3D People from Images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE.

[6] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. 2019. Multi-Garment Net: Learning to Dress 3D People From Images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 5419–5429.

[7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. 2016. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *ECCV*.

[8] Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K. Wong. 2022. JIFF: Jointly-aligned Implicit Face Function for High Quality Single View Clothed Human Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Hang Dai, Nick Pears, William AP Smith, and Christian Duncan. 2017. A 3d morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE international conference on computer vision*. 3085–3093.

[10] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[11] Endri Dibra, Himanshu Prakash Jain, A. Cengiz Öztireli, Remo Ziegler, and Markus H. Gross. 2017. Human Shape from Silhouettes Using Generative HKS Descriptors and Cross-Modal Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 5504–5514.

[12] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 2020. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)* 39, 5 (2020), 1–38.

[13] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–13.

[14] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*. Springer, 152–168.

[15] Tong He, John P. Collomosse, Hailin Jin, and Stefano Soatto. 2020. Geo-PIFu: Geometry and Pixel Aligned Implicit Functions for Single-view Human Reconstruction. *ArXiv* abs/2006.08072 (2020).

[16] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. 2021. ARCH++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11046–11056.

[17] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. 2020. ARCH: Animatable Reconstruction of Clothed Humans. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3090–3099. https://doi.org/10.1109/CVPR42600.2020.00316

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR* (2017).

[19] S. Jinka, R. Chacko, A. Sharma, and P. Narayanan. 2020. PeeledHuman: Robust Shape Representation for Textured 3D Human Body Reconstruction. In *2020 International Conference on 3D Vision (3DV)*. IEEE Computer Society, 879–888. https://doi.org/10.1109/3DV50981.2020.00098

[20] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 8320–8329.

[21] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-end Recovery of Human Shape and Pose. In *Computer Vision and Pattern Recognition (CVPR)*.

[22] Michael M. Kazhdan and Hugues Hoppe. 2013. Screened poisson surface reconstruction. *ACM Trans. Graph.* 32, 3 (2013), 29:1–29:13.

[23] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 2252–2261.

[24] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. 2021. Probabilistic Modeling for Human Mesh Recovery. In *ICCV*.

[25] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. 2017. Unite the People: Closing the Loop Between 3D and 2D Human Representations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 4704–4713.

[26] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.

[27] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. 2020. Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 5891–5900.

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[29] William E. Lorensen and Harvey E. Cline. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *COMPUTER GRAPHICS* 21, 4 (1987), 163–169.

[30] Gyeongsik Moon and Kyoung Mu Lee. 2020. Pose2Pose: 3D Positional Pose-Guided 3D Rotational Pose Prediction for Expressive 3D Human Pose and Mesh Estimation. *ArXiv* abs/2011.11534 (2020).

[31] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. 2019. SiCloPe: Silhouette-Based Clothed People.. In *CVPR*. Computer Vision Foundation / IEEE, 4480–4490.

[32] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. 2018. Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation. (2018), 484–494. http://doi.ieeecomputersociety.org/10.1109/3DV.2018.00062

[33] X. Yu M. Whalen G. Harvey S. Orts-Escolano R. Pandey J. Dourgarian et al. K. Guo P. Lincoln. P. Davidson, J. Busch. 2019. The relightables: volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics(ToG)* 34 (2019).

[34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. 2018. Learning to Estimate 3D Human Pose and Shape from a Single Color Image. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 459–468.

[36] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. 2017. ClothCap: seamless 4D clothing capture and retargeting. *ACM Trans. Graph.* 36 (2017), 73:1–73:15.

[37] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. 2021. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5620–5629.

[38] Sai Sagar Jinka, Astitva Srivastava, Chandradeep Pokhariya, Avinash Sharma, and P. J. Narayanan. 2022. SHARP: Shape-Aware Reconstruction of People in Loose Clothing. *arXiv e-prints*, Article arXiv:2205.11948 (May 2022), arXiv:2205.11948 pages. arXiv:2205.11948 [cs.CV]

[39] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2304–2314. https://doi.org/10.1109/ICCV.2019.00239

[40] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *CVPR*.

[41] Evangelos Sariyanidi, Casey J Zampella, Robert T Schultz, and Birkan Tunc. 2020. Inequality-constrained and robust 3D face model fitting. In *European Conference on Computer Vision*. Springer, 433–449.

[42] T. Simon, S. Lombardi, J. Saragih, and Y. Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics(ToG)* 6 (2018).

[43] Ji-Hey Song and Hyun-Joon Shin. 2009. On Parameterizing of Human Expression Using ICA. *Journal of the Korea Computer Graphics Society* 15, 1 (2009), 7–15.

[44] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. 2018. BodyNet: Volumetric Inference of 3D Human Body Shapes. In *ECCV*.

[45] Abbhinav Venkat, Sai Sagar Jinka, and Avinash Sharma. 2018. Deep Textured 3D Reconstruction of Human Bodies. In *BMVC*.

[46] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. 2022. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13296–13306.

[47] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomír Mech, and Ulrich Neumann. 2019. DISN: Deep Implicit Surface Network for High-quality Single-view 3D Reconstruction. In *NeurIPS*.

[48] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12803–12813.

[49] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*.

[50] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. 2019. DF2Net: A Dense-Fine-Finer Network for Detailed 3D Face Reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*. 2315–2324.

[51] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. 2021. PaMIR: Parametric Model-Conditioned Implicit Representation for Image-based Human Reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. https://doi.org/10.1109/TPAMI.2021.3050505

[52] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. 2019. DeepHuman: 3D Human Reconstruction From a Single Image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 7738–7748. https://doi.org/10.1109/ICCV.2019.00783

[53] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. 2019. Detailed Human Shape Estimation From a Single Image by Hierarchical Mesh Deformation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 4486–4495.