JIFF: Jointly-aligned Implicit Face Function for High Quality Single View Clothed Human Reconstruction

Yukang Cao\textsuperscript{1} Guanying Chen\textsuperscript{2} Kai Han\textsuperscript{1} Wenqi Yang\textsuperscript{1} Kwan-Yee K. Wong\textsuperscript{1}
\textsuperscript{1}The University of Hong Kong \hspace{2em} \textsuperscript{2}The Future Network of Intelligence Institute (FNii), CUHK-Shenzhen

Abstract

This paper addresses the problem of single view 3D human reconstruction. Recent implicit function based methods have shown impressive results, but they fail to recover fine face details in their reconstructions. This largely degrades user experience in applications like 3D telepresence. In this paper, we focus on improving the quality of face in the reconstruction and propose a novel Jointly-aligned Implicit Face Function (JIFF) that combines the merits of the implicit function based approach and model based approach. We employ a 3D morphable face model as our shape prior and compute space-aligned 3D features that capture detailed face geometry information. Such space-aligned 3D features are combined with pixel-aligned 2D features to jointly predict an implicit face function for high quality face reconstruction. We further extend our pipeline and introduce a coarse-to-fine architecture to predict high quality texture for our detailed face model. Extensive evaluations have been carried out on public datasets and our proposed JIFF has demonstrates superior performance (both quantitatively and qualitatively) over existing state-of-the-arts.

1. Introduction

Under the current social distancing measures of the COVID-19 pandemic, video conferencing has become the major form of daily communication. With the increased popularity of 3D hardware like AR goggles, 3D telepresence \cite{42} will soon likely emerge as the next generation communication standard. High quality 3D human reconstruction is at the core of this technology and is one of the current hottest topics. Traditional reconstruction methods depend on expensive capturing hardware and tedious calibration procedure to produce good looking models \cite{15}. This limits their applications to studio settings with expert users and greatly hinders the growth of AR/VR applications. It is highly desirable to develop easy-to-use tools that allow easy creation of high quality 3D human models by home users using commodity RGB cameras.

With the advance in deep learning techniques, recent 3D human reconstruction methods have achieved impressive results using as few as a single image \cite{29, 55, 74}. These methods can be roughly divided into model based methods \cite{3, 4, 6, 7} and model-free methods \cite{11, 12, 18, 26, 49, 55}. Model based methods typically fit a parametric human model (e.g., SMPL \cite{37}) to an image to produce a naked 3D human model. They have difficulties in recovering high-frequency details such as clothing and hair. Model-free methods, on the other hand, solve this problem by predicting the occupancy of a discretized volume space. One very representative model-free method is PIFu \cite{55}, which exploits a Multi Layer Perceptron (MLP) to model an implicit function for predicting the occupancy value of a query point based on pixel-aligned features extracted from an image. PIFu and its variants \cite{34, 56} have achieved state-of-the-art results in free-form full body human reconstruction. However, their reconstructions are often lack of fine face details (see Fig. 1). Considering the ultra-high-definition rendering standards (\textit{i.e.} 4K UHD and 8K UHD) that are common nowadays, their face reconstruction quality is obviously far from satisfactory and largely degrades user experience in AR/VR applications like 3D telepresence.

To achieve high quality human reconstruction with fine face details, we propose a novel Jointly-aligned Implicit
Face Function (aka JIFF) that combines the merits of the implicit function based approach and model based approach. Specifically, we employ the 3D morphable face model (3DMM) [8] as our shape prior and compute space-aligned 3D features to capture detailed face geometry and texture information. There are also recent methods [25, 72] using 3D priors to enhance the implicit function representation by introducing geometric constraints to regularize the reconstruction. For example, [25] and [72] utilize coarse 3D volume features and SMPL body model, respectively, to improve human body reconstruction. To the best of our knowledge, JIFF is the first method focusing on recovering high quality face details in both shape and texture.

JIFF exploits space-aligned 3D features extracted from 3DMM as well as pixel-aligned 2D features extracted from image to jointly predict an implicit face function for high quality face reconstruction. In summary, our method first fits the 3DMM to the face in an image and employs two separate encoders to compute 3D shape and texture features, respectively, from the resulting 3D model. Given a 3D query point, we obtain its space-aligned 3D features with trilinear interpolation. Such space-aligned 3D features are combined with pixel-aligned 2D features for predicting the occupancy value of the query point using a MLP. We further extend our pipeline and introduce a coarse-to-fine architecture to predict high quality texture for our detailed face model.

By taking advantages of both the implicit function based approach and model based approach, our method can successfully recover fine face details in both shape and texture (see Fig. 1). Our key contributions are as follows:

- We propose JIFF, a novel implicit face function for high quality single view 3D face reconstruction, which integrates 3D face prior into the implicit function representation for high quality face shape reconstruction.
- We exploit per-vertex color information provided by the 3DMM and introduce a coarse-to-fine architecture for high quality face texture prediction.
- We demonstrate how JIFF can naturally be extended to produce full body human reconstruction by simply appending a “PIFu” head (implemented as a MLP) to its convolutional image encoder.
- We carry out extensive experiments on public benchmarks and demonstrate that JIFF outperforms current state-of-the-arts by a large margin.

2. Related work

Single view human reconstruction Reconstructing 3D human body from a single image is an important and challenging problem which has attracted a considerable amount of attention. Methods have been developed to fit parametric human models such as SCAPE [23], SMPL [9], SMPL-X [46] and STAR [43] to a single image. Human Mesh Recovery (HMR) [33] proposes to regress SMPL parameters from a single image. Labels like body part segmentation [41], silhouette [47], and IUV map [64] have been employed to provide intermediate supervisions for training SMPL prediction models. Although free-form deformation [1, 2, 32, 63] may be applied to the models to partially account for complex shape topology (e.g., clothing and hair), these methods generally have difficulties in reconstructing high quality clothed human models. Model-free methods have been proposed to reconstruct 3D human with arbitrary topology. Voxel-based methods reconstruct 3D human body using a volumetric representation with different intermediate supervisions (e.g., multi-view images [22], 2D pose [59, 60, 73], and 3D pose [28]). However, volumetric representation is memory intensive and is hard to scale to high resolution. Recently, memory efficient implicit function representations [11, 39, 49, 55, 56, 61] have achieved outstanding performance in 3D reconstruction. DeepSDF [44] predicts a signed distance field for surface reconstruction. IF-Net [12] learns multi-scale features for 3D mesh refinement or completion, and is later extended for texture completion [13]. SiCloPe [39] reconstructs a visual hull by predicting 3D pose and 2D silhouettes. PIFu [55] introduces a pixel-aligned implicit function for human reconstruction from a single image. Variants have been proposed for high-resolution reconstruction [56] and real-time rendering [34]. PeeledHuman [30] proposes to encode the human body as a set of peeled depth and RGB maps to handle the self-occlusion problem. Although these methods successfully reconstruct 3D human body geometry and texture from a single image, they fail to deal with issues like self-occlusion and detailed face reconstruction.

3D prior for implicit function representation Efforts have been made to utilize 3D human body prior for implicit model reconstruction. GeoPIFu [25] proposes to use geometry-aligned feature from the predicted 3D feature volume to improve reconstruction quality. PaMIR [72] utilizes SMPL as a 3D prior for implicit function learning. ARCH and ARCH++ [26, 27] extract 3D spatially-aligned features from a canonical SMPL mesh to reconstruct animatable human models. SHARP [31] proposes the peeled SMPL priors for learning. DeepMultiCap [71] uses SMPL to tackle the multi-person reconstruction problem. Although improved results have been reported, the above mentioned methods still cannot recover fine face details in their reconstructions. High quality face reconstruction remains an open problem.

Human face/head reconstruction Parametric model based methods often adopt 3D Morphable Model (3DMM) [8, 19, 20, 24, 36, 57, 76] for face reconstruction and head models (e.g., FLAME [35], DECA [21], LYHM [16, 17], and UHM [50, 51]) for full-head reconstruction. The models adopted by these methods, however, often limit their expressiveness in handling arbitrary shapes. i3DMM [65] combines 3DMM and implicit function to
3. Preliminary

In the following subsections, we are going to give a brief review on PIFu [55] and 3DMM [8, 19] which lay the foundation for JIFF.

3.1. Pixel-aligned implicit function

Recently, implicit functions [58] have been widely adopted for 3D reconstruction [38]. Denoting \( X \in \mathbb{R}^3 \) as a 3D point, a deep implicit function, modeled by a MLP, defines a surface as the level set of the function, e.g., \( f(X) = 0.5 \). The pixel-aligned implicit function \( f_v \) introduced in [55] is written as

\[
f_v(B(\psi(I), \pi(X)), z(X)) \rightarrow [0, 1] \in \mathbb{R},
\]

where \( z(X) \) is the depth of \( X \) and \( B(\psi(I), \pi(X)) \) is the pixel-aligned feature, with \( \psi(I) \) denotes the feature map extracted from the image \( I \) by a convolutional encoder [40], \( \pi(X) \) the 2D projection of \( X \) on \( I \), and \( B(\cdot) \) a bilinear interpolation operation. Despite its simplicity, PIFu has achieved impressive results in full body human reconstruction. However, PIFu is incapable of recovering fine face details in its reconstructions (see Fig. 2).

3.2. 3DMM as 3D face prior

A natural idea to improve PIFu is to enforce some sort of 3D shape prior in learning the implicit function representation. Efforts have been made to enhance the pixel-aligned implicit function with 3D features like coarse 3D volume features [25] and voxelized SMPL mesh features [26, 27, 72]. While improved overall reconstruction results have been reported, these method still cannot recover fine face details in their reconstructions. This is not unexpected as the shape priors adopted by these methods do not actually focus on the face.

An ideal 3D face prior should provide both geometry and texture information, and can be robustly estimated from an image. Parametric face/head models like 3DMM [8] and DECA [21] are potential good candidates as they provide both 3D mesh and texture information to resolve depth ambiguity and improve texture prediction. They can be delineated with a small number of parameters, which can be estimated effectively using existing methods (e.g., [19]). In this work, we choose the widely used 3DMM as our 3D face prior for its simplicity and efficacy. Note that other parametric face models can also be adopted as JIFF does not depend on a specific parametric model.

3DMM models a face as a linear combination of the Principle Component Analysis (PCA) basis vectors. The shape \( \mathbf{S} \) and texture \( \mathbf{T} \) of 3DMM are expressed as

\[
\mathbf{S} = \bar{\mathbf{S}} + \mathbf{B}_{id}\alpha + \mathbf{B}_{exp}\beta,
\]

\[
\mathbf{T} = \bar{\mathbf{T}} + \mathbf{B}_{tx}\delta,
\]

where \( \bar{\mathbf{S}} \) and \( \bar{\mathbf{T}} \) denote the mean shape and texture respectively, \( \mathbf{B}_{id}, \mathbf{B}_{exp}, \) and \( \mathbf{B}_{tx} \) are the PCA bases for face identity, expression, and texture respectively, and \( \alpha, \beta, \) and \( \delta \) are the identity, expression, and texture coefficients respectively. In our implementation, we employ \( \bar{\mathbf{S}}, \bar{\mathbf{T}}, \mathbf{B}_{id}, \) and \( \mathbf{B}_{tx} \) extracted from BFM [48], and \( \mathbf{B}_{exp} \) built from Face-Warehouse [10].

4. Jointly-aligned implicit face function

In this section, we introduce our novel Jointly-aligned Implicit Face Function (JIFF). JIFF is designed with the goal of incorporating 3D face prior in learning implicit function for high quality single view clothed human reconstruction. As already mentioned in the previous section, we choose 3DMM as our 3D face prior for its simplicity and efficacy. Given an input image, we can obtain the 3DMM parameters using [19]. However, merging the resulting 3DMM mesh with the mesh extracted from PIFu is a non-trivial problem, as they do not share the same geometry and topology (see Fig. 2). Instead of trying to merge the two meshes naïvely in the 3D space, we propose to fuse the information in the feature space, and use both space-aligned 3D features extracted from 3DMM and pixel-aligned 2D features extracted from image to jointly estimate an implicit face function (and hence the name Jointly-aligned Implicit Face Function).

Given a 3DMM mesh \( \mathbf{S} \) fitted to the input image, we employ an encoder to generate a feature volume \( \psi(\mathbf{S}) \) from the mesh (see Sec. 4.2 for details). JIFF can be formulated
Figure 3. Overview of the network architecture for our proposed Jointly-aligned Implicit Face Function (JIFF). JIFF takes a single image as input to reconstruct the high quality 3D human model with fine face details. It seamlessly incorporates 3D face prior into the implicit function representation for both shape and texture reconstruction. By appending two MLP heads, one after the pixel-aligned feature (in parallel with the MLP in yellow) in the top branch and the other after the multi-scale pixel-aligned feature (in parallel with the MLP in yellow) in the bottom branch, JIFF can be naturally extended to produce full body reconstruction. The dashed paths are unique for the face region. The rest are shared for the full body reconstruction.

as

\[ f^g(\tau(\varphi(S), X), B(\psi(I), \pi(X)), z(X)) \rightarrow [0, 1] \in \mathbb{R}, \] (3)

where \( \tau(\varphi(S), X) \) is the space-aligned 3D feature for the query point \( X \) obtained by trilinear interpolation \( \tau \). By employing both pixel-aligned 2D features and space-aligned 3D features, JIFF enjoys the benefits of both the implicit function based approach and model based approach and is capable of recovering fine face details in its reconstructions.

In the following subsections, we are going to give the details for each of the components in our JIFF learning framework. We are also going to describe how JIFF can naturally be extended for full body reconstruction by simply appending a “PIFu” head (implemented as a MLP) to its convolutional image encoder.

4.1. 3DMM prediction and alignment

Given an input image, we first detect the face region \([68]\) and adopt the state-of-the-art 3DMM regression method \([19]\) to predict the 3DMM parameters \((\alpha, \beta, \delta, \gamma, p) \in \mathbb{R}^{257}\), where \(\alpha \in \mathbb{R}^{80}, \beta \in \mathbb{R}^{64}, \delta \in \mathbb{R}^{80}, \gamma \in \mathbb{R}^{27}\), and \(p \in \mathbb{R}^{6}\) represent face identity, expression, texture, illumination, and pose respectively. We can then obtain the 3DMM mesh and texture using Eq. (2), and apply the pose parameters to transform the mesh into the camera coordinate system. Next, we scale the 3DMM mesh to match the original size of the face in the input image\(^1\), and further align the mesh with the back-projected face landmarks. At training time, we carry out the Iterative Closest Point (ICP) \([5]\) algorithm to align the 3DMM mesh with the ground-truth mesh instead. This step is important in computing reliable space-aligned 3D features for training and testing. Details for the alignment processes (both for training and testing) can be found in the supplementary material.

4.2. Point-based 3D face feature encoding

To extract expressive features from the 3DMM mesh, we need to have a proper feature encoder. Inspired by \([49]\), which predicts occupancy from point clouds, we adopt PointNet++ \([52]\) to extract hierarchical 3D point features for each vertex in the 3DMM mesh. These 3D point features are then projected into a 3D feature volume using average pooling. This feature volume is further processed by 3D U-Net \([14]\) to aggregate local and global information. In our implementation, the spatial dimension of our 3D feature volume is set to \(64 \times 64 \times 64\), and the features extracted by 3D U-Net have a dimension of 128. Instead of using a single feature encoder consisting of PointNet++ and 3D U-Net, we propose to employ two separate encoders in our framework: one focusing on shape prediction by taking the plain 3DMM mesh as input, and the other focusing on texture prediction

\(^1\)Scaled and cropped face images are used in the 3DMM regression.
by taking the textured 3DMM mesh as input (see Fig. 3). The 3D features extracted by the geometry and texture encoders are denoted as \( \varphi_g(S) \) and \( \varphi_t(S) \) respectively. Without any color information, \( \varphi_g(S) \) is forced to learn features that depend only on geometry, whereas the additional color information helps learn better texture in \( \varphi_t(S) \).

### 4.3. Occupancy prediction

Referring to Eq. (3), JIFF takes both space-aligned 3D geometry features \( \tau(\varphi_g(S), X) \) and pixel-aligned 2D features \( B(\psi_g(I), \pi(X)) \) to predict the occupancy value of a query point \( X \). Following [55], we adopt stack-hourglass encoder [40] for image feature extraction. The image encoder \( \psi_g \) takes image \( I \) \( \in \mathbb{R}^{512 \times 512 \times 3} \) as input, with background masked out by the segmentation mask and produces a feature map \( \psi_g(I) \) \( \in \mathbb{R}^{128 \times 128 \times 256} \). The pixel-aligned 2D feature for \( X \) is then obtained via bilinear interpolation on \( \psi_g(I) \). Similarly, the space-aligned 3D geometry feature for \( X \) is obtained via trilinear interpolation on \( \varphi_g(S) \).

Inspired by Geo-PIFu [25], instead of directly concatenating the pixel-aligned 2D feature and space-aligned 3D geometry feature, we first apply two separate MLPs to transform these features independently. The transformed features are then concatenated and fed into an MLP for occupancy prediction (see Fig. 3). We train our shape prediction network by minimizing the mean squared error between the predicted and ground-truth occupancy values of the query points. The ground-truth occupancy value is 1 if \( X \) is inside the surface, and 0 otherwise.

### 4.4. Face texture prediction

Apart from recovering geometric details for face reconstruction, JIFF is also capable of improving face texture prediction. Similar to shape prediction, the deep implicit function for face texture prediction can be formulated as

\[
\mathbb{R}^3 = f^t(\tau(\varphi_t(S), X), B(\psi_t(I), \pi(X)), z(X)) \in \mathbb{R}^3,
\]

where \( \tau(\varphi_t(S), X) \) is the space-aligned 3D texture feature and \( B(\psi_t(I), \pi(X)) \) is the pixel-aligned 2D feature. The space-aligned 3D texture feature extracted by the texture feature encoder \( \varphi_t(S) \) embeds space-aware texture information from the textured 3DMM, which is particularly helpful for face texture prediction. Note that none of the existing methods (e.g. PAMIR [72]) that employ parametric models to improve human reconstruction consider such space-aware texture information due to the absence of texture information in their parametric models.

#### Coarse-to-fine texture prediction

Thanks to the space-aligned 3D features extracted from the textured 3DMM mesh, JIFF can produce better face texture than PIFu. To further improve the quality of the predicted texture, we, inspired by H3D-Net [55], introduce a coarse-to-fine architecture for texture prediction that exploits pixel-aligned features extracted from a higher resolution image (see Fig. 3). Our proposed architecture is composed of a coarse branch and a fine branch. The coarse branch is identical to Tex-PIFu [55] which takes the concatenation of the pixel-aligned feature from the shape prediction network and pixel-aligned feature from the texture prediction network as input to a MLP to predict coarse texture. The fine branch takes an upsampled image of size \( 1024 \times 1024 \times 3 \) as input to the image encoder to produce a feature map of size \( 512 \times 512 \times 256 \) (4 times the width and height of the feature map in the coarse branch). Pixel-aligned feature computed from this fine feature map is concatenated with the output from the penultimate layer (denoted as \( \Omega \) in Fig. 3) of the MLP in the coarse branch to form a multi-scale pixel-aligned feature.

Similar to the shape prediction network, this multi-scale pixel-aligned feature and the space-aligned texture feature are transformed independently by two separate MLPs before they are being concatenated and fed to the final MLP for fine texture prediction.

We first train the coarse branch, and then train the fine branch with the parameters of coarse branch being frozen. \( L_1 \) loss is used to train both the coarse and fine branches. At training time, we randomly perturb \( X \) with an offset \( \epsilon \in \mathcal{N}(0, d) \) along its unit surface normal \( N \), i.e. \( X' = X + \epsilon \cdot N \), for point sampling [55]. This strategy allows the color of a surface point to be defined in a 3D space around its exact location which can stabilize the training process.

### 4.5. Full body reconstruction

Up till now, our discussion has been focused on face reconstruction. JIFF can actually be naturally extended for full body reconstruction by simply appending a “PIFu” head (implemented as a MLP) to its convolutional image encoder in the shape prediction network. Given a query point \( X \), we apply JIFF to predict its occupancy value if it is projected to the face region in the input image, otherwise we predict its occupancy value using the “PIFu” head. Adopting other PIFu variants for improved reconstruction is also trivial. Similarly, to predict the texture for a non-face query point, we append a “non-face texture” head (also implemented as a MLP) to the texture prediction network which takes the multi-scale pixel-aligned feature as input to predict the texture color. Though without the space-aligned 3D features, our coarse-to-fine design can also notably improve the texture prediction for non-face regions.

### 5. Experiments

In this section, we evaluate our proposed JIFF on public datasets and compare it with other state-of-the-art methods.

#### Implementation details

We apply the stack-hourglass image encoder [40] to extract image feature for shape prediction. Following the design of Peng et al. [49], our 3D
AXYZ datasets

Many recent methods use RenderPeople datasets. [https://yukangcao.github.io/JIFF]

maining the surface normal direction, and uniformly sample the re-
et al, [53] to uniformly sample 15/16 of the points on the
mesh surface followed by a Gaussian perturbation along
the surface normal direction, and uniformly sample the re-
main ing 1/16 of the points within the bounding box of the
mesh. This strategy is helpful to eliminate isolated outliers
in the reconstruction. We implement our method using Py-
Torch [45] and carry out our experiments on three NVIDIA
2080Ti GPUs. Our code will be made publicly available at
https://yukangcao.github.io/JIFF.

Datasets

Many recent methods use RenderPeople\(^2\) and AXYZ datasets\(^3\) to train their models. However, these two datasets are commercial data and not publicly available. Instead, we use the public THuman2.0 dataset [66] as our main testbed, which contains high-quality human scans with different clothes and poses, captured by a dense DSLR rig. 3D mesh and the corresponding texture map are provided for each subject. We randomly split the data into 465 subjects for training and 61 subjects for testing. For each subject, we follow Saito et al. [55] to render 360° images by rotating the camera around the mesh and varying the illumination. As our main focus is to improve face details in the reconstruction, we omit images in which no human face can be detected. In addition, we use BUFF dataset [67], which contains 5 subjects, as an additional testing dataset to evaluate our method. Besides, we evaluate JIFF on 2 free models from RenderPeople.

5.1. Comparison with state-of-the-arts

We compare our method with recent state-of-the-art methods, including PIFu [55], PIFuHD [56], PaMIR [72], ARCH [27], and ARCH++. For a fair comparison, we retrain PIFu, PIFuHD, and PaMIR on the THuman2.0, as they are originally trained on either non-public commercial data or a different version of THuman data. As the codes for ARCH and ARCH++ are not publicly available, their authors help to provide the evaluation results on our data upon our request.

Quantitative comparison

We quantitatively evaluate on the test split of THuman2.0 dataset and the BUFF dataset [67]. As our main focus is to recover faces with fine details, we first compare our method with others on the head region. The results are shown in Table 1. We measure both Point-to-Surface distance (P2S) and Chamfer mean distance in the head region and L2 face mean distance [53] in the face region by back-projecting 2D image points to 3D face surface. It can be observed that our method significantly outperforms all the others on both datasets, confirming that our method is capable of recovering fine face details. We locate the head region by applying an off-the-shelf human parsing method [70]. It is interesting to note that PIFuHD performs the worst for the head/face region among the methods under comparison. We further show body-only reconstruction results in Table 1 and report the P2S and Chamfer distances. We can observe that our JIFF slightly improves the body-part quality from PIFu, although there is no 3DMM-like extra information adopted for training the body. We think the improvement originates from the fact that JIFF is jointly trained on face and body. The shared network backbone is updated by gradients from both head and body branches and the improved head branch also affects the body branch in a positive way, resulting in a stronger backbone for both head and body reconstruction.

Qualitative comparison

We qualitatively compare our method with the state-of-the-arts in Fig. 4 for both face shape and texture (the corresponding full body reconstructions can be found in the supplementary material). Our

<table>
<thead>
<tr>
<th>Method</th>
<th>THuman2.0</th>
<th>BUFF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Head/F</td>
<td>Head/F</td>
</tr>
<tr>
<td>PIFu [55]</td>
<td>0.427</td>
<td>0.761</td>
</tr>
<tr>
<td>PIFuHD [56]</td>
<td>0.650</td>
<td>0.855</td>
</tr>
<tr>
<td>PaMIR [72]</td>
<td>0.403</td>
<td>0.693</td>
</tr>
<tr>
<td>Ours</td>
<td>0.141</td>
<td>0.291</td>
</tr>
</tbody>
</table>

**Table 1. Quantitative comparison on head/face and body-only reconstructions. Results are measured in cm (the lower the better).**
Figure 4. Qualitative comparison with other state-of-the-arts. Here we focus on comparing the face geometry and texture. Please refer to the supplementary material for full body reconstruction results.

Figure 5. Error map of the face region based on P2S distance. Our proposed method achieves much better results compared with other state-of-the-arts.

Table 2. Effectiveness of our 2D-3D feature fusion. (∆ denotes the concatenation operator.)

<table>
<thead>
<tr>
<th></th>
<th>Face L2 distance ↓</th>
<th>Head P2S ↓</th>
<th>Head Chamfer ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) 2D only</td>
<td>0.427</td>
<td>0.761</td>
<td>0.756</td>
</tr>
<tr>
<td>(b) 3D only</td>
<td>0.279</td>
<td>0.485</td>
<td>0.492</td>
</tr>
<tr>
<td>(c) 2D ∆ 3D</td>
<td>0.171</td>
<td>0.310</td>
<td>0.332</td>
</tr>
<tr>
<td>(d) MLP(2D) ∆ MLP(3D)</td>
<td>0.141</td>
<td>0.291</td>
<td>0.308</td>
</tr>
</tbody>
</table>

Our reconstruction is also closest to the ground truth. Other methods struggle in these aspects. We show the error map in Fig. 5 which demonstrates that our method outperforms others for the majority part of the face region. We also evaluate our method on two free models from RenderPeople in Fig. 6. Again, our method is capable of recovering fine face details, which is consistent with the ground truth.

5.2. Ablation study

2D-3D feature fusion We validate the effectiveness of the fusion of pixel-aligned 2D feature with space-aligned 3D feature in our framework by comparing the following variants: (a) using pixel-aligned 2D feature alone; (b) using space-aligned 3D feature alone; (c) simply concatenating the 2D and 3D features; and (d) applying transformation by additional MLPs before concatenation (i.e., implementation of JIFF). The intuition of applying the additional MLPs before concatenation is to learn a proper embedding space for the two different types of features to fuse better. The results are reported in Table 2. It can be observed that JIFF significantly outperforms the other variants.

Dual 3D feature encoders In our framework, we introduce two separate 3D feature encoders focusing on improv-
Figure 7. Reconstruction results for different 3D feature encoder designs. From left to right: (a) Single encoder with plain 3DMM. (b) Single encoder with textured 3DMM. (c) Dual encoders with plain 3DMM and textured 3DMM.

Figure 8. Ablation study on different architectures for texture prediction. From left to right: (a) coarse resolution only (original PIFu); (b) fine resolution only; (c) fine resolution w/ 3D face prior; (d) coarse-to-fine w/ 3D face prior (ours).

Table 3. Effectiveness of dual 3D feature encoders.

<table>
<thead>
<tr>
<th></th>
<th>Face SSIM ↑</th>
<th>Face LPIPS ↓</th>
<th>Head L1 error ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) single encoder + plain 3DMM</td>
<td>0.7631</td>
<td>0.1143</td>
<td>0.1137</td>
</tr>
<tr>
<td>(b) single encoder + textured 3DMM</td>
<td>0.7589</td>
<td>0.1222</td>
<td>0.1090</td>
</tr>
<tr>
<td>(c) dual encoders + plain/textured 3DMM</td>
<td>0.7649</td>
<td>0.1107</td>
<td>0.1051</td>
</tr>
</tbody>
</table>

Table 4. Effectiveness of our coarse-to-fine architecture for texture prediction.

<table>
<thead>
<tr>
<th></th>
<th>Face SSIM ↑</th>
<th>Face LPIPS ↓</th>
<th>Head L1 error ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Coarse</td>
<td>0.7126</td>
<td>0.1281</td>
<td>0.1346</td>
</tr>
<tr>
<td>(b) Fine</td>
<td>0.7264</td>
<td>0.1254</td>
<td>0.1285</td>
</tr>
<tr>
<td>(c) Fine w/ 3D</td>
<td>0.7580</td>
<td>0.1126</td>
<td>0.1097</td>
</tr>
<tr>
<td>(d) Coarse-to-Fine w/ 3D</td>
<td>0.7649</td>
<td>0.1107</td>
<td>0.1051</td>
</tr>
</tbody>
</table>

variants would result in blurry face texture or poor backside texture. Moreover, our dual encoder successfully help improve the texture of the backside of the head.

Coarse-to-fine texture prediction In Table 4 and Fig. 8, we compare our coarse-to-fine architecture for texture prediction in our framework with the following variants: (a) coarse resolution only; (b) fine resolution only; (c) fine resolution jointly with 3D face prior; and (d) coarse-to-fine setting jointly with 3D face prior. From both qualitative and quantitative results, fine resolution and 3D face prior could indeed help gain high-resolution texture, and face texture details respectively. Also, it can be observed that our coarse-to-fine architecture can notably improve the texture prediction over its single resolution counterparts.

6. Conclusion

We have presented JIFF, a novel Jointly-aligned Implicit Face Function for high quality single view 3D human reconstruction, by incorporating the 3D face prior, in the form of 3DMM, into the implicit representation. We further introduced a coarse-to-fine architecture for high quality face texture reconstruction. By simply appending two MLPs, one for shape and the other for texture, to JIFF, we successfully extended it to produce full body human reconstruction. We thoroughly evaluated JIFF on public benchmarks, establishing the new state-of-the-art for face details reconstruction. As JIFF is simple and effective, we believe it can be easily coupled with other human reconstruction approaches to improve their reconstruction quality for faces. Though JIFF achieves better performance than existing methods, subtle face details like eyelids still cannot be accurately reconstructed. One future research direction therefore is improving the quality for subtle details, as well as extending the idea to other body parts, e.g., hands and feet.

Acknowledgements This work was partially supported by Hong Kong RGC GRF grant (project# 17203119), the National Key R&D Program of China (No.2018YFB1800800), and the Basic Research Project No. HZQB-KCZYJ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone. We thank Yuanlu Xu for ARCH and ARCH++ results.
References


[27] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In IEEE Conference on Computer Vision and Pattern Recognition, June 2020. 2, 3, 6

[28] Aaron Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via
volumetric regression. In *European Conference on Computer Vision Workshops*, pages 0–0, 2018. 2


[53] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitiza-


