DOI: 10.1111/cgf.14798

COMPUTER GRAPHICS *forum*
*Volume 42 (2023), number 6, e14798*

# OaIF: Occlusion-Aware Implicit Function for Clothed Human Re-construction

Yudi Tan,[1] Boliang Guan,[2] Fan Zhou[1] and Zhuo Su[1,*]

[1]School of Computer Science and Engineering, National Engineering Research Center of Digital Life, Sun Yat-sen University, Guangzhou, China
suzhuo3@mail.sysu.edu.cn
[2]School of Electronic and Information Engineering, Foshan University, Foshan, China

**Abstract**

*Clothed human re-construction from a monocular image is challenging due to occlusion, depth-ambiguity and variations of body poses. Recently, shape representation based on an implicit function, compared to explicit representation such as mesh and voxel, is more capable with complex topology of clothed human. This is mainly achieved by using pixel-aligned features, facilitating implicit function to capture local details. But such methods utilize an identical feature map for all sampled points to get local features, making their models occlusion-agnostic in the encoding stage. The decoder, as implicit function, only maps features and does not take occlusion into account explicitly. Thus, these methods fail to generalize well in poses with severe self-occlusion. To address this, we present OaIF to encode local features conditioned in visibility of SMPL vertices. OaIF projects SMPL vertices onto image plane to obtain image features masked by visibility. Vertices features integrated with geometry information of mesh are then feed into a GAT network to encode jointly. We query hybrid features and occlusion factors for points through cross attention and learn occupancy fields for clothed human. The experiments demonstrate that OaIF achieves more robust and accurate re-construction than the state of the art on both public datasets and wild images.*

**Keywords:** modelling, image-based modelling, implicit surfaces

**CCS Concepts:** • Computing methodologies → Computer graphics; Shape modelling; Mesh geometry models

## 1. Introduction

Clothed human re-construction is widely applied in human digitization, remote presence and virtual try-on. And re-construction based on the monocular image is a challenging task with lack of information, self-occlusion and depth ambiguity.

Previous methods using parametric model [LMR*15] can reconstruct full human shape from a single RGB image. But the results lose high frequency information such as hair and cloth wrinkle. Therefore, the following works [ZZW*19, APMTM19] take static templates as geometry priors to guide the network to predict displacement fields for such templates. While some others learn latent code in low dimension for pre-defined clothing templates with outfit-specific generative networks [BTTPM19, TBTPM20, CPA*21]. However, these methods, constrained by the topology of the template, are unable to tackle varied styles of human clothes.

The proposed method is based on widely followed implicit surface representation [PFS*19]. This kind of method relies on the property of the neural network, being able to fit any topology in theory. Among shape re-construction approaches with implicit representation, those combining a unified global coding with 3D coordinates have achieved considerable results in the re-construction of general objects [CZ19, KBJM18]. But they are not applicable for clothed humans with lots of local details. Thus pixel-aligned features are proposed to model high-frequency information and show its power in recovering local details [SHN*19, SSSJ20]. These methods usually use high precision, general standing commercial 3D models for training, resulting in their struggling with non-standing poses or wild images with severe occlusion. The generated meshes usually include unexpected deformation and broken body parts. The main reason for this limitation, in addition to the distribution of training data, lies in the depth ambiguity of pixel-aligned

[*] Corresponding author

features. It can be attributed to using a unified image encoder to obtain the image features of all sampling points, regardless of whether they are visible or not on the plane of the image.

We aim to concentrate image encoding on the inference of visible information and make the network aware of the visibilities of sampling points. And then our model can make more clear discriminations in feature representation, which is important for implicit surface learning. The image features are more significant compared to coarse 3D ones when the depth does not need to be integrated explicitly in the image feature encoding stage. Thus, we propose an Occlusion-aware Implicit Function called OaIF, to utilize image feature properly. With a single RGB image and corresponding mask as input, a coarse SMPL model predicted by Pare [KHHB21] is used as body reference and a feature map is generated by hourglass network [NYD16]. Similar to the approach used in Li *et al.* [LWF*21], we embed pixel-aligned features into the topology space of the vertices of the parametric model. Next, the 2D image feature and 3D geometry feature can be jointly encoded on the surface manifold. We also utilize a mask to filter 2D feature, enabling this encoding process to perceive the visibilities of vertices. To query the fused feature and how much it is occluded for any point, a cross attention module is applied. The resulting occlusion factor can be further used to filter pixel-aligned feature for any point in space. Finally, the concatenation of multi-modal features is fed into multi-layer perceptron (MLP) to learn the occupancy field and then a mesh can be extracted with Marching Cubes [LC87]. Since OaIF heavily depends on the accuracy of SMPL model, we apply the body reference optimization in PaMIR [ZYLD21] to align SMPL to the 0.5 level set conditioned in the observed image, which results in a better generalization ability.

We conclude the major contributions as follows:

(1) OaIF is presented, a method based on the implicit function to re-construct clothed human. Compared with PIFu-liked approaches, it is the first to integrate visibility information to reduce the ambiguity of pixel-aligned feature.
(2) A multi-modal feature fusion method is proposed. To use anchor points of SMPL as reference, the feature encoding process can be done in topology space which is more reasonable compared to Euclidean space.
(3) The whole framework realizes a close coupling between implicit function and parametric model. Therefore, it can benefit from the performance of the latest model in human shape estimation.

The rest of this paper is organized as follows. In Section 2, we review the related work in human digitization. In Section 3, we present the detail of the proposed method. In Section 4, we describe implementation details and experimental results and analysis. Finally, we summarize our work in Section 5.

## 2. Related Work

Previous works in human re-construction focus on the prediction of human pose and shape. Though the generated mesh is hardly clothed, the methods also contribute to clothed human reconstruction based on explicit representation. In this section, we classify related works according to the form of 3D model representation.

**Explicit representation**. Statical mesh is widely used in image-based clothed human re-construction [KBJM18, ZCL*19]. These methods constrain the parameter space to be searched by defining the connection between vertices through pre-defined template models such as SMPL. Spin [KPBD19] iteratively optimizes the 2D joint position projected from 3D joints in the training and testing stages to provide more accurate SMPL parameter estimation. GCMR [KPD19] provides global coding from the image for each vertex on the downsampling mesh. It predicts the vertex position in current pose through graph convolution and regresses the SMPL parameters. Lin *et al.* [LWL21] used transformer to capture the long-distance relationship. DC-GNet [ZJCL21] models both the positive and negative dependencies between vertices and introduces mask to tackle a shape completion task. But all the methods are proposed to estimate human pose and shape aligned with the observed image without cloth information. Others aim to recovery local details such as clothes and hair by predicting the relative offset of mesh vertices. Smpl+D [AMB*19] introduces per-vertex offset into SMPL formula to express high frequency details. Tex2Shape [APMTM19] predicts deformation displacement through the convolution of UV map. Li *et al.* [LWF*21] introduced the deformation representation suitable for mesh vertices and uses graph convolution to predict the offset of each vertex, so as to realize clothing modelling.

Related works with point cloud as input also adopt similar methods to learn displacement mapping for each template mesh vertex [BSTPM20a, MSY*21, MYTB21]. These parametric model-based methods can realize animation driving for the generated mesh on the pre-defined skeleton and skin weights. They are also compatible with current graphics tools. But the key limitation is that they can not express complex and varied clothing structures through an offset field of the statical template. Therefore, some works try to learn different network parameters [TBTPM20, MSY*21], latent codes in low dimension [BTTPM19, SMB*20] or embeddings [MYTB21] for different cloth types. The performance of these methods to model clothing deformation depends on the distribution of labelled data and cannot generalize well in a variety of clothing styles.

**Implicit surface**. Different from explicit modelling, the surface of 3D model can be defined as an implicit function, which is parameterized in high-dimensional space, and the implicit representation is learned through the fitting ability of MLP to any function. Based on this idea, for monocular image re-construction of general objects, the target surface is defined on the 0 level-set, and the signed distance field in 3D space is learned to represent the object surface. For any given point, their high-dimensional features are usually encoded as the concatenation of a unified one [PFS*19] or local ones [XWC*19] and 3D coordinates. Then, the signed distance between the query point and the implicit surface is output through MLP.

Since the implicit function can represent any surface, PIFu[SHN*19] is proposed, as the first to utilize the pixel-aligned feature to learn unsigned occupancy fields for clothed humans. In the following work, PIFuhd [SSSJ20] enhances details of generated mesh with normal maps and high resolution image as input, but much more parameters need to be optimized. To achieve the limb integrity of the re-constructed model, subsequent works introduce 3D information to constrain the prediction of the occupancy field. Geo-PIFu [HCJS20] uses a voxelization layer to integrate the voxel features generated by coarse human shapes. S3 [YWM*21] introduces point cloud features and additionally learns skin weights as well as skeletons to realize animation driving. StereoPIFu [YWM*21] uses the depth map estimated by binocular images as auxiliary information to re-construct mesh more accurately. However, when such methods deal with poses excluded in synthetic datasets, the generated models are mostly missing limbs or stretched along the *Z*-axis. We attribute this to the dependency of such model on the inference of MLP about 3D information. And using pixel-aligned features without distinguishing visibility results in feature homogeneity and ambiguity.

**Hybrid representation**. Implicit representation is topology-free while human mesh template strongly constraints on the integrity and connection of human parts, thus can provide coarse but enough reasonable 3D information. Hybrid representation takes advantage of both. PaMIR [ZYLD21] voxelizes a predicted SMPL mesh to form corresponding voxel feature with 3D convolution. The hybrid aligned feature facilitates the approach to realize better local detail recovery and re-construction integrity. But the signal conversation from mesh to volume leads to quantization errors especially when the volume is defined in low resolution, thus the prior form SMPL model is not fully utilized. ARCH [HXL*20] learns the occupancy field in the canonical pose space of SMLP model, and then maps back to the original pose space through the deformation field. ARCH++ [HXS*21] learns occupancy fields of two pose spaces simultaneously. Point cloud encoder is used to solve the memory limitation of volume, and occupancy consistency constraint of two spaces is applied to achieve better re-construction in the seam part caused by the interaction of limbs. However, this kind of method needs to map the sampling points to canonical pose space through the deformation field and skin weight, so the non-rigid transformation of some clothes can not be represented accurately. And the point cloud encoder acts in Euclidean space which is different from the manifold space of implicit surface in distance metric. Loopreg [BSTPM20b] also takes the consistency of two pose spaces into account and introduces a semi-supervision framework to perform point cloud registration and mesh re-construction. ICON [XYTB22] only uses local features from the front and back normal maps to achieve better robustness towards poses. But it fails with varied cloth style because of the limited representation ability of completely local geometric feature. Based on hybrid representation, our method makes full use of SMPL topology as a prior of the relationship between limbs and realizes a more effective feature fusion process and neighbourhood perception in topology space.

## 3. Proposed Method

We aim to infer a clothed human model from the monocular RGB image. Given the image and corresponding background mask, we first feed it into a fully convolutional network with feature map as output and predict SMPL parameters through Pare [KHHB21]. Next, we query the pixel-aligned feature for SMPL vertices by projection and interpolation and calculate the visibility to filter it. Based on the pre-defined topology of the human body, we encode the pixel-aligned feature and geometry feature of vertices jointly using a graph attention network. Then the hybrid feature is used to predict occupancy value through MLP. The whole procedure is shown in Figure 1.

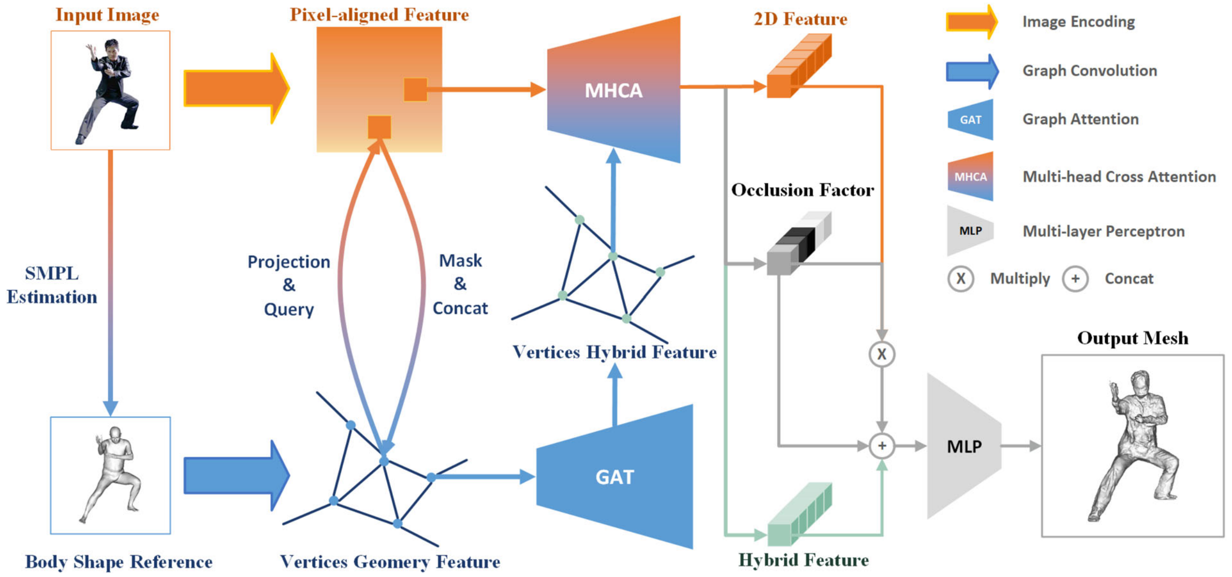

**Figure 1:** *OaIF predicts SMPL body reference with Pare and then obtains the visibility mask from the rendered image of SMPL model in fixed camera parameter. Pixel-aligned feature can be filtered by the mask and will be jointly encoded with geometry feature in the graph attention network. A cross attention is applied to query the hybrid feature and occlusion factor for any point. The colours in figure indicate different feature space. By projection and interpolation, OaIF embeds pixel-aligned feature from 2D pixel grid to topology space.*

### 3.1. Local feature encoding

Local features such as pixel-aligned feature and voxel-aligned feature, which fuse information in a certain receptive field, have been proved by experiments of PIFu-liked approaches to be indispensable in recovering high-frequency details. They ensure limb integrity and improve re-construction fidelity in occupancy field prediction based on the implicit function. This section introduces two local feature extraction methods used in OaIF.

**Pixel-aligned feature**. We follow PIFu [SHN*19, SSSJ20] with stacked hourglass network as the image encoder, which has shown excellent performance in pose estimation and previous human reconstruction tasks. Given any point $p_i \in \mathbb{R}^3$, the pixel-aligned feature can be presented as

$$f_{2D}(p_i, \mathcal{I}) = \mathcal{B}(\psi_\mu(I), \pi(p_i)), \tag{1}$$

where $\psi_\mu$ is the image encoder and $\mathcal{B}$ is bilinear interpolation while $\pi$ is the weak projection from the world coordinate to image plane with fixed camera parameters. These operations are continuously differentiable.

**Vertex hybrid feature**. With a pre-trained network of Pare [KHHB21], we can infer SMPL model from the input image as the body shape and pose reference. Since the parametric model strongly constrains the integrity of human limbs and joint position, and it provides the topological relationship of each vertex on human surface, we fuse pixel-aligned feature and geometric feature of vertex on the basis of the topological structure. Specifically, based on graph convolution, the normal and coordinates are encoded as geometrical feature tensors:

$$X^{l+1} = \sigma(D^{-\frac{1}{2}} \hat{A} D^{-\frac{1}{2}} X^l W^l), \tag{2}$$

in which $X^l \in \mathbb{R}^{N_v \times h_l}$ is the input of $l$th layer and the first one $X^1 = [\mathcal{V}, \mathcal{N}] \in \mathbb{R}^{N_v \times 6}$ is the concatenation of normal and coordinates. $\hat{A} = D - A \in \mathbb{R}^{N_v \times N_v}$ is the Laplacian matrix of adjacency matrix. And $D$ is the diagonal matrix which represents the degree of each vertex.

We find that encoding the image feature on the surface manifold is a more reasonable choice rather than convolution on a regular image plane since it fits the curved body structure more properly. Actually, it is similar to the UV mapping used in Tex2Shape [APMTM19], which flattens surface as a regular grid. However, the partition operation used in UV brings about discontinuity near seams and rearranges the connection of limbs. Thus some priors need to be relearned in convolutional network.

Through Equation (1), we get the pixel-aligned feature of each vertex. Concatenated with the geometrical feature, it is forwarded into graph attention network [VCC*17] to achieve multi-modal fusion. The attention matrix is as follows:

$$\alpha_{ij} = \frac{\exp(\sigma(\vec{a}^T [h_i^l, h_j^l]))}{\sum_{k \in \mathcal{N}_i} \exp(\sigma(\vec{a}^T [h_i^l, h_k^l]))}, \tag{3}$$

in which $H^l = \{h_i^l\}_{i=1}^N$ is hybrid feature in $l$th layer and $H^1 = \{[\mathcal{M}'' \cdot f_{2D}(\mathcal{V}, \mathcal{I}), f_{3D}(\mathcal{V}, \mathcal{N})]\}_i^N$ . $\mathcal{M}''$ is a mask to filter pixel-aligned feature and will be introduced in the next part. Here we choose graph attention instead of the transformer used in Li *et al.*[LWF*21] to model long-distance dependence in the shallow network. This is mainly because of the fact that cloth wraps around the human body rather than scatters around. Therefore, the fusion process demands a constraint of locality supplied by a graph structure.

**Vertex visibility mask**. In the previous works using pixel alignment feature as local coding, queried points with the same coordinates on the image plane share the same feature interpolated feature value, despite their depth values in 3D space being different. It would be reasonable if depth information is implicitly encoded in different channels with a powerful convolutional network. In the experiment, we observe that although this simple assumption provides feature coding for the sampling points invisible on the image plane, it causes severe depth ambiguity with a small amount of data used for training. That is, there are mesh stretching artifacts along the $z$-axis direction especially when the pose is beyond those in the train set. We attribute this to the homogeneity and ambiguity of pixel-aligned feature. Since it is the most informative part of most models proposed in PIFu-liked works, the performance of MLP, which act as an implicit function, is directly influenced by its property. As Figure 2 shows, inspired by Refs. [LWL21, ZJCL21] using vertex mask to enhance the robustness of the model in human shape reconstruction, we model the feature fusion as a feature completion task. Specifically, we render the SMPL model conditioned in fixed camera parameters and image resolution. The generated image corresponds to visible triangle fragments of SMPL mesh, thus we can obtain visibility for each vertex of SMPL. The visibility $\mathcal{M}'$ mixed with a random mask can be used to filter pixel-aligned feature:

$$\mathcal{M}'' = \mathcal{M}' \oplus \mathcal{R}, \tag{4}$$

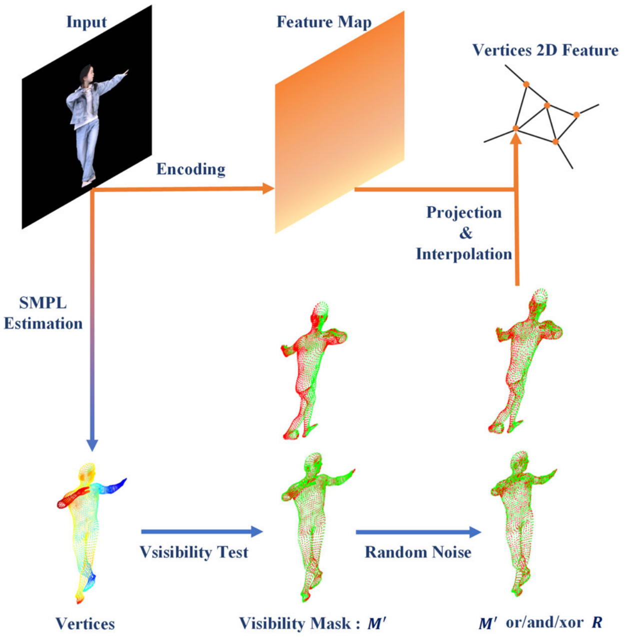

**Figure 2:** *Filter the pixel-aligned feature with the visibility mask. The colour in the lower left indicates depth variation. The red and green of vertex after visibility test means whether to drop pixel.*

in which $\oplus$ is xor operation and the random re-masking operation is to alleviate the influence of misaligned SMPL vertices during training. For each visible vertex, the mask may drop its feature, and for the invisible one, the mask may provide its feature. This masking process makes most invisible points get their image information from the weighting and transmission of the visible part. It is a similar task compared with image inpainting but carried out on the surface manifold rather than a regular pixel grid. With the support of 3D geometric information and graph structure, the neural network can make a more reasonable weighting strategy.

Our key idea is that, for clothed human re-construction, the pixel-aligned feature should be scattered on the visible surface in current view. To infer about the invisible part, the topology and geometry priors of parametric model are indispensable. With the help of graph convolution, this visible information can be spread on the surface manifold constrained by locality.

**Cross attention**. To obtain the hybrid feature of any point in space and retain continuity of calculation on the change of its coordinate, we use cross attention to query feature. Given the pixel-aligned features $f_{2D}(\mathcal{P},\mathcal{I})$ of queried points as query matrix, we assign the hybrid feature of SMPL to key and value matrix:

$$\begin{cases} Q = [F_Q(f^*_{2D}), \gamma(Z)], \\ K = [F_K(H^L), \gamma(\mathcal{V})], \\ V = [F_V(H^L), \gamma(\mathcal{V})], \end{cases} \tag{5}$$

where $F_Q$, $F_K$, $F_V$ are learnable forward networks, and $Att$ is the general attention operation. $H^L$ is the output of the graph attention network and $f^*_{2D}$ is the pixel-aligned feature detached from computation graph. This detach operation is important to decouple the optimization of image encoder parameters from attention computation. Thus stacked hourglass network could concentrate image feature inference on visible part. Additionally, we repeat the feature channels to achieve multihead and the final attention matrix is the mean of heads.

To distinguish the pixel-aligned features of query points with different depths, we only embed $z$-axis coordinates as positional encoding $\gamma(Z)$ since $x$, $y$ values have been implicitly included in the feature. And 3D coordinates of vertices are used for the key and value matrix. Here, the Fourier feature embedder $\gamma$ keeps the same with NeRF [MST*20]. Thus, the consequent attention matrix can be used in a dynamic weighting process to query hybrid feature for arbitary points in $\mathbb{R}^3$:

$$f_h = Att(Q,K,V) \times H^L. \tag{6}$$

Since the attention value indicates how important the vertices are for queried points in embedding space, we assume that the queried points which pay more attention to invisible vertices are more invisible on the image plane. We product the visibility mask with the attention matrix to get a floating occlusion factor $\mathcal{M}$ for queried points:

$$\mathcal{M} = Att(Q,K,V) \times \mathcal{M}'. \tag{7}$$

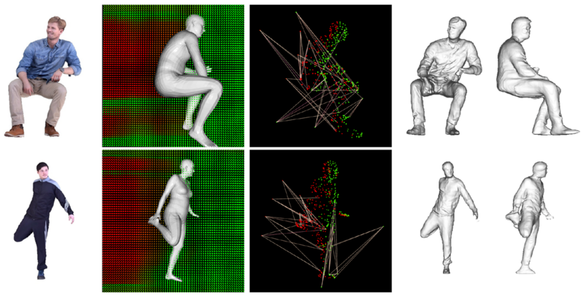

**Figure 3:** *OaIF estimates the occlusion factor for queried points with SMPL vertices as reference. From left to right: input, occlusion factor of point, attention weight of point to anchors, front and side view of re-constructed mesh.*

**Algorithm 1.** Occlusion-aware Feature Encoding.

```
Require: Image feature map $\phi_\mu(\mathcal{I})$; SMPL vertices coordinates $\mathcal{V}$,
    normals $\mathcal{N}$ and adjacency matrix $A$; Camera parameters $\mathcal{K}$; Queried
    points $P$
Ensure: Occlusion-aware hybrid feature $f_h(P)$; Occlusion factor $\mathcal{M}$
$f_{2D}(P) \leftarrow \mathcal{B}(\phi_\mu(\mathcal{I}), \pi(P))$.
$f_{2D}(\mathcal{V}) \leftarrow \mathcal{B}(\phi_\mu(\mathcal{I}), \pi(\mathcal{V}))$.
$f_{3D}(\mathcal{V}) \leftarrow GCN(\mathcal{V}, \mathcal{N}, A)$.
$\mathcal{M}' \leftarrow Render(\mathcal{V}, \mathcal{N}, A, \mathcal{K})$
$\mathcal{M}'' \leftarrow \mathcal{M}'' \oplus \mathcal{R}$
$f_{2D}(\mathcal{V}) \leftarrow \mathcal{M}'' \cdot f_{2D}(\mathcal{V})$ .
$H \leftarrow GAT(f_{3D}(\mathcal{V}), f_{2D}(\mathcal{V}))$.
$Q \leftarrow F_Q(f_{2D}(P), \gamma(Z)), K \leftarrow F_K(H, \gamma(\mathcal{V})), V \leftarrow F_V(H, \gamma(\mathcal{V}))$
$f_h(P) \leftarrow Att(Q,K,V) \times H, \mathcal{M} \leftarrow Att(Q,K,V) \times \mathcal{M}'$
```

Therefore, taking the visibility of SMPL vertices as a reference, we realize the occlusion-awareness for any queried point in space through the attention mechanism as shown in Figure 3 and Algorithm 1.

### 3.2. Loss function

Given the mixed and pixel-aligned features, we use MLP to regress the occupancy field as an implicit surface:

$$\mathcal{L}_o = |\hat{o} - \mathrm{MLP}([\mathcal{M} \cdot f_{2D}, f_h, \mathcal{M}, \gamma(\mathcal{P})])|^2, \tag{8}$$

in which $\hat{o}$ is the ground truth occupancy value. Though the feature is redundant, we find that while pixel-aligned feature recovers details, the neighbourhood and global awareness of hybrid features in topological space can ensure limb integrity and guide the details to be more reasonable. Due to using the sparsity of the SMPL vertices, it is easy to bring about relatively smooth hybrid features. We use sine as activation in MLP, and introduce corresponding modulator in skip connection to ensure the implicit function more sensitive with slight input changes. Additionally, we introduce the real visibility of sampled points in the rendered image as a regularization. The loss function can be expressed as follows:

$$\mathcal{L}_{\mathcal{M}} = |\mathcal{M} - \min\{\exp[\alpha(\mathcal{Z}_{\mathcal{I}} - \mathcal{Z})], 1\}|. \tag{9}$$

It should be pointed out that although it is feasible to directly predict the visibility of sampled points supervised by corresponding labels, the learning process is difficult without 3D information as prior. Moreover, this is incompatible with the sampling method used by PIFu, which pre-samples near the surface of the mesh with Gaussian distribution of offset. A large number of sampled points for training are located in a small area near the surface inside the model, which is invisible, but close to the surface. Obviously, the prediction of such points' occupancy needs to rely partly on the image information on the surface. Therefore, we infer the visibility of sampled points by referring to vertices on the human body surface and take the visibility loss as an extra guiding condition. As presented in Equation (9), it is designed as a form of depth awareness, in which $\mathcal{Z}_{\mathcal{I}}$ is the depth buffer value of the pixel in the rendered image and $\mathcal{Z}$ is the depth of sampled points. We use mesh rasterizer and point rasterizer, respectively, with the same camera parameter to get the ground truth fragments. And $\alpha$ controls the influence of the difference in depth to visibility. Using the float value of visibility rather than the binary one as supervision can smooth the filtering operation. Hence the visibility mask here can be interpreted as an occlusion factor to restrain pixel-aligned feature of points deviated from the visible surface. The total loss is

$$\mathcal{L} = \mathcal{L}_o + \lambda \mathcal{L}_{\mathcal{M}}. \tag{10}$$

## 4. Experiments

In this section, we provide the experiment settings and comparisons with state-of-the-art methods. Then we conduct several ablation studies on the proposed method.

### 4.1. Implementation details

We use the open-source model of Pare [KHHB21] fine-tuned in Thuman2.0 as a pre-trained model to estimate the SMPL parameters. And the loss of silhouette after rendering is applied additionally. For the training of OaIF, we set the batch size to 4 and the epoch to 10. Adam is used as the optimizer with an initial learning rate of $2e-4$ and decays every 20,000 iterations with a rate of 0.1. To obtain the ground truth of vertex visibility in the training process, we use Pytorch3d to render the SMPL model under weak perspective projection with the predicted camera parameters and query the visibility of triangular fragments as the mask of corresponding vertices. In the inference stage, we use the body reference optimization in PaMIR [ZYLD21] to make the current SMPL prediction aligned with the input image and introduce the 2D joint position estimated by OpenPose and the silhouette difference based on differentiable rendering as the loss terms of optimization.

In the network structure, we follow PIFu, using the stacked hourglass network with two stacks without intermediate supervision to reduce the number of parameters. The only difference is that we use 256×256×256 as the dimension of the output feature map to get better details. Limited by memory size, we use the downsampled 432 SMPL vertices as reference points, and the early graph convolution outputs 32-dimensional geometric features with the coordinates and normal vector of each vertex as input. Thus, the channel number for the hybrid vertex feature is 288. For positional encoding, we apply the $\gamma(\cdot)$ function in NeRF to get 63 channels. The number of neurons in MLP hidden layers is (608, 1024, 512, 256, 256, 1), in which the input part includes pixel-aligned features, hybrid features obtained through cross attention, occlusion factor and positional encoding. The activation function is set as sine [SMB*20] and the modulars [MGB*21] are used in hidden layer 3, 4 to replace tensor concatenation in skip connection. We set $\alpha = 5$ in visibility loss and $\lambda = 0.2$ in total loss to achieve better balance. For the ratio of random mask, an experimental number of 0.2 works well.

Because of memory size, we only use a shallow GAT network with three residual modules and set the neighbourhood of the adjacency matrix to 2 to achieve larger receptive field. And the GNN to encode geometry feature of SMPL keeps a vanilla number of vertices of SMPL rather a downsampling one. For faster convergence in the training stage, we integrate the voxel feature of SMPL model as the controller of frequency and amplitude in the modulator. This is also used in PaMIR[ZYLD21] but as an additional global feature for MLP.

### 4.2. Data pre-processing

To get the ground truth difference between depth of sampling points and z-buffer of rendered images as needed supervision, we use points rasterizer in Pytorch3D to render point clouds and obtain their depth in current view. There is a maximum number of points that can be detected in a ray. Thus, a few points may not be assigned a depth value. We just simply set the same with z-buffer of images which means they are on the surface and can get the pixel-aligned feature without filtering. This hardly influences the training process because we use random mask to achieve better robustness.

For ground truth SMPL parameters, we just select the one predicted by Pare[KHHB21] with the lowest Chamfer and P2S loss in different views and fine-tune its shape to fit the scan. It is not the standard registration procedure and may lead to misaligned parts between the SMPL model and scan just as Figure 4 shows. Most misaligned parts are hands and fingers, so this problem is partly because of the limited expression ability of SMPL about fingers and may

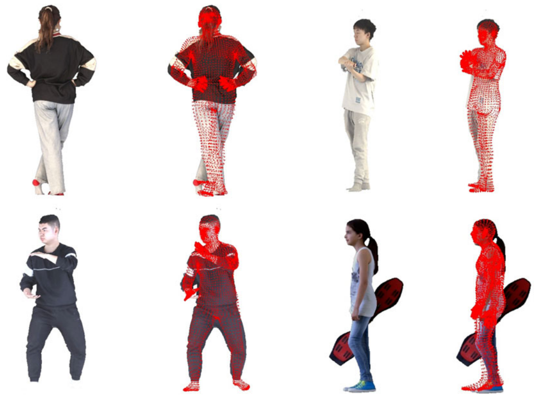

**Figure 4:** *Some misaligned cases in our data pre-processing. We simply apply shape estimation methods based on the rendered image and fine-tune shape parameters to get the ground truth SMPL model. Thus, it is not completely aligned with scan.*

**Table 1:** *Quantity evaluation on THuman2.0 and 3DPeople. We normalize the coordinates to* [0, 1]*, thus, there is no unit in Chamfer (e-3) and P2S (e-3). But they also indicate relative performance.*

| | THuman2.0 | | | 3DPeople | | |
|---|---|---|---|---|---|---|
| Method | Chamfer | P2S | Normal | Chamfer | P2S | Normal |
| Tex2Shape[APMTM19] | 2.132 | 2.075 | 0.170 | 3.583 | 3.031 | 0.214 |
| PIFu[SHN*19] | 1.748 | 1.695 | 0.185 | 3.628 | 2.763 | 0.223 |
| PIFuHD[SSSJ20] | 1.421 | 1.335 | 0.171 | 3.513 | 2.720 | 0.219 |
| PaMIR w gtSMPL[ZYLD21] | 0.961 | 0.891 | 0.169 | 3.368 | 2.309 | 0.215 |
| **Ours** | 1.372 | 1.297 | 0.180 | 3.603 | 2.696 | 0.241 |
| **Ours w/o mask** | 0.789 | 0.693 | 0.154 | 3.129 | 2.046 | 0.183 |
| **Ours w gtSMPL** | **0.726** | **0.659** | **0.147** | **2.925** | **1.814** | **0.165** |

cause inaccurate and uncertain hands re-construction as our qualitative evaluation shows. In the previous experiments, we present that OaIF heavily depends on the accuracy of predicted SMPL parameters. So the performance may be further improved if registered SMPL or SMPL-X model [PCG*19] can be obtained through external tools.

### 4.3. Evaluation

For THuman2.0[1] and 3DPeople[2] datasets, we take 450 scans as training data, respectively, and render the image with spherical harmonic illumination [VRM*17] in 60 uniformly distributed view angles. We pre-sample 50K points with the same sampling strategy as PIFu, and randomly select 5000 points during training. For the visibility annotation of sampling points, we use orthogonal projection to render the point cloud and mesh model in fixed camera parameters. Then we obtain the depth value of sampling points at their corresponding pixels and the depth buffer of mesh with point cloud and mesh rasterizer, respectively.

We conduct the quantity evaluation on Thuman2.0 and 3DPeople with 75 and 30 scans, respectively. The models are rendered in front view and we pre-compute the 2D keypoints using OpenPose to supervise the body reference optimization during inference. We compare OaIF against the state of the art methods based on the implicit function. For PaMIR, we use the open-source model, which is trained in 360 views with about 1000 scans and uses the ground truth SMPL as coarse body references. We use Pare and GCMR, respectively, in the inference stage and choose the better one as results of PaMIR. PIFu and PIFuHD are trained as usual. To align the world coordinates with ground truth scan, for methods to be compared with, we optimize the scale factor and transformation vector until the convergence of Chamfer and point to surface distance before metrics calculation. For the proposed method, we evaluate the model trained with the prediction of Pare as input. As shown in Table 1, it is slightly worse than PaMIR in all metrics, which can be attributed to the heavy dependence on the accuracy of SMPL parameters estimation mentioned before. Embedding pixel-aligned feature into SMPL vertex couples our method with the backbone we used. While when the proposed method is trained with ground truth SMPL as initial body reference, OaIF outperforms all the state-of-the-art methods in all metrics. The Chamfer and P2S indicate better global integrity and accuracy in re-construction. Though our method filters pixel-aligned feature with predicted mask, it also performs slightly better in Normal term which indicates local details. Last, it results performance decline slightly when visibility mask and Equation (9) is not explicitly introduced.

We show some test cases of the 3DPeople dataset in Figure 5. Since our method aims to be occlusion-aware, we mainly focus on non-standing poses such as sitting (rows 4 and 5) and crouching (row 6). Tex2Shape[APMTM19], as a representative method of explicit re-construction, only predicts human shape and cloth displacement, thus we repose the results with SMPL pseudo labels. As the figure shows, it cannot tackle non-standing poses and tend to output human with obesity when the cloth is loose. Without geometry prior, PIFu outputs some unnecessary floating lumps whose corresponding pixel-aligned feature is ambiguous and fails to give reasonable results when some limbs are occluded in the image. PaMIR uses low-resolution voxel feature as additional information which is vague when the body huddles up. We attribute this to the inductive bias of the 3D convolutional network. For voxels of different body parts, they may be closed in Euclidean space while far away on the human body surface. So in rows 2 and 5, there are unexpected connection structures between legs. Thus our graph attention is more powerful to deal with this since the multi-modal features are fused in topology space. Moreover, benefitting from our depth-aware occlusion factor and pixel-aligned feature filtering operation, we reduce the stretching artifacts along the $z$-axis, especially near hands and feet.

Next, we evaluate the proposed method with images in the wild to test the generalization ability in Figure 6. Extra background masks and 2D key points are also pre-computed for PaMIR and OaIF. PIFu and PIFuhd do not use SMPL prior and cannot generalize well in wild images. Thus we only present the result of PaMIR, ICON and OaIF. The PaMIR model is pre-trained with CMR, so we choose a better result with CMR and Pare as SMPL estimation methods, respectively, for a fair comparison. As for ICON [XYTB22], we use the public pre-trained model for evaluation. The results show that ICON leans to re-construct less cloth and there are block artifacts in

[1] THuman2.0: https://github.com/ytrock/THuman2.0-Dataset
[2] 3DPeople: https://3dpeople.com/en

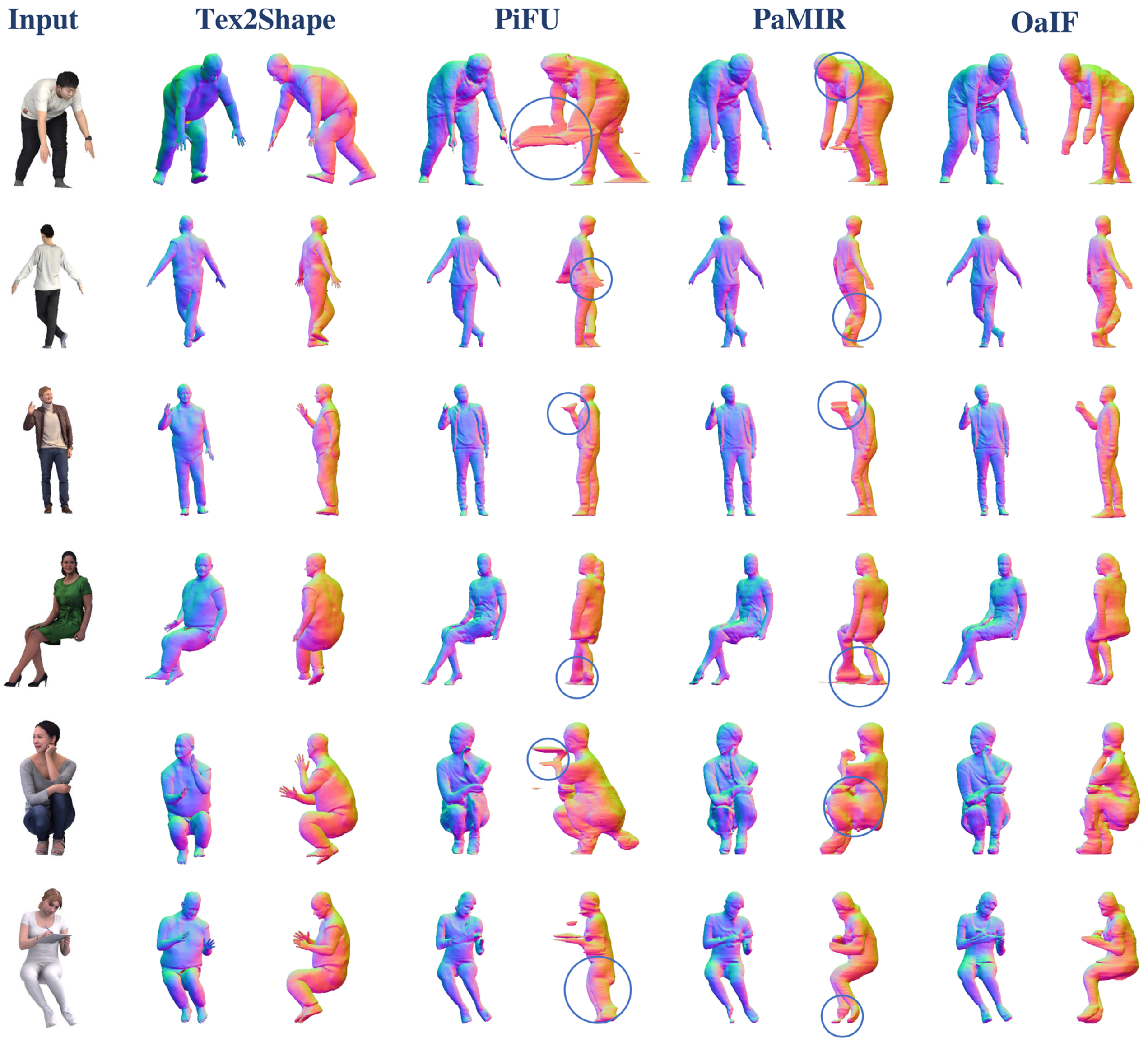

**Figure 5:** *Qualitative comparisons with the state-of-the-art methods based on implicit function. For each column, we show both front (left) and side (right) views to exemplify the stretching artifacts along the z-axis, especially in non-standing poses. From left to right: input images, results of Tex2Shape[APMTM19], PIFu [SHN*19], PaMIR [ZYLD21] and OaIF (ours).*

some cases due to the limited representation ability of its completely local feature. Please note that we do not quantitatively evaluate ICON since they use thousands of scans from AGORA [PHT*21] for training which may contain our test cases. It can be seen from our methods how the details in the front view spread to the side with encoding in topology space. And we effectively reduce the stretching artifacts.

## 4.4. Ablation

**Feature fusion**. We first evaluate the validity of our cross-attention module. Using vertices or points as a carrier of 3D prior has been proved feasible in ARCH++ [HXS*21]. It utilizes PointNet++ [QYSG17] to encode point clouds to get geometry features from anchor points and use K-nearest neighbours weighting to query the feature for any points. Thus we replace cross attention with the same weighting strategy and also set $K = 3$. The results are shown in Figure 7. The re-construction model is filled with jitters and holes. We attribute this to the discontinuity of argmax operation when coordinates of queried points change continuously. And it gets worse near high frequency details. It is strictly required in implicit representation that encoding should be continuously differentiable. Therefore, cross attention as a smooth weighting method is the better solution for the approaches that use anchor points as a reference. And we can apply any constraints, we want to the attention matrix. For example, to realize locality in $K$-nearest,

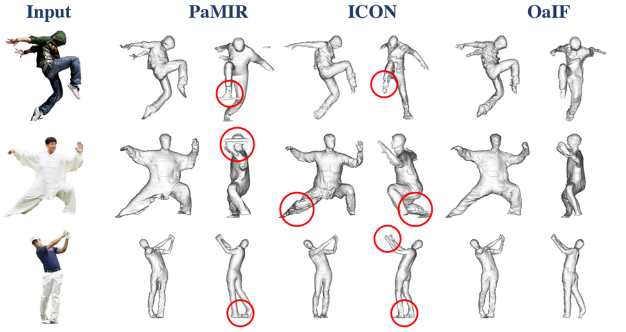

**Figure 6:** *Evaluation of images in the wild. With occlusion-aware encoding for queried points, our method reduces the stretching artifacts in feet and hands. From left to right: input masked images in the wild, results of PaMIR [ZYLD21], ICON [XYTB22] and OaIF (ours).*

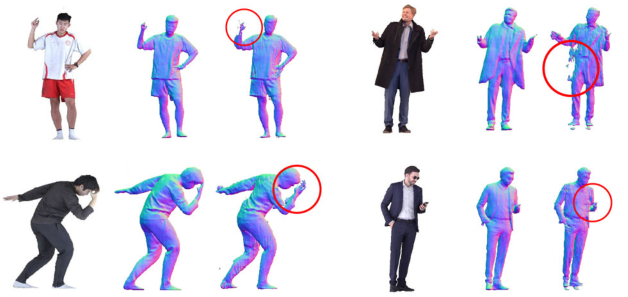

**Figure 7:** *Replace cross attention (left) with K-nearest neighbour weighting (right). Because the argmax operation is not continuously differentiable, there are jitters and parts missing on the surface which correspond to the input jumping of MLP.*

**Table 2:** *Quantitative results of the ablations of multi-modal fusion and SMPL estimation.*

| | THuman2.0 | | | 3DPeople | | |
|---|---|---|---|---|---|---|
| Method | Chamfer | P2S | Normal | Chamfer | P2S | Normal |
| W/o fusion | 0.855 | 0.807 | 0.166 | 3.547 | 2.239 | 0.208 |
| Ours w CMR | 1.081 | 0.929 | 0.173 | 3.403 | 2.657 | 0.210 |

a regularizer in the loss function is enough. We leave this for future work.

**SMPL estimation**. As we introduce before, the performance of OaIF is heavily dependent on the SMPL parameter estimation method. We evaluate OaIF with CMR [KPD19] used in PaMIR and Pare [KHHB21] as pose estimator, respectively, on non-standing pose. The quantitative results of model using CMR are shown in Table 2. Since Pare is more robust to varied poses, it outputs more reasonable parameters. As Figure 8 shows, CMR gives wrong prediction on the left leg, while Pare outputs a more correct one. Thus, the re-constructed geometries using two human shape estimation methods differ a lot, especially in limbs integrity. For this, each vertex of SMPL has its own semantics in training and corresponds to the pixel feature. If miss-aligned pixel features are embedded to vertices, the error will spread on the surface and further influence the inference of the hybrid feature and occlusion factor of queried points. We regard this as a limitation of the proposed method. But clothed human re-construction from a monocular image, as mentioned at the start, is very challenging. Shape prior of the parametric model would help a lot if a powerful enough shape estimation method is available while this is easier without consideration of cloth and other details.

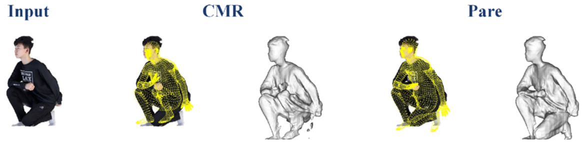

**Figure 8:** *Ablation on SMPL estimation method. Our method heavily depends on the semantics alignment between SMPL and the input image. From left to right: input images, results when CMR [KPD19] and Pare [KHHB21] are used as SMPL regressor, respectively.*

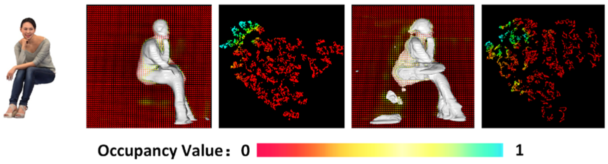

**Figure 9:** *Ablation on multi-modal feature fusion. From left to right: input image, a cross-section with multi-modal fusion, the tSNE visualization with multi-modal fusion, a cross-section without multi-modal fusion and the tSNE visualization without multi-modal fusion.*

**Hybrid encoding**. We then consider the influence of hybrid coding of 3D geometric features and pixel alignment features on the re-construction effect. Assuming that there is no interdependence between the two modal features, we use group convolution as the feature mapping process in graph attention, which can divide the feature channels into intervals to avoid their interaction. It performs worse slightly in all metrics as presented in Table 2. As shown in Figure 9, the joint encoding of multi-modal features makes the vertex features for arbitrary points more reasonable, which is less ambiguous for the implicit function. We show the output of last second hidden layer with the tSNE visualization. There are more obvious and regular boundaries among the sampling points with different occupancy. The feature from encoders directly affects the prediction accuracy of occupancy field, mainly because the encoding process can improve the feature difference and make the implicit function based on MLP easier to learn. Please note that we do not present the quantity evaluation results of ablations on feature fusion and SMPL estimation method since the accuracy gap is obvious and easy to be captured in listed rendering images.

Although the previous work based on hybrid representation also uses multi-modal features, they isolated the features of different modes from each other in the encoding stage. The features are further mapped in the MLP, but for the sampling points, there is no neighbourhood or global information perception at the same time

**Table 3:** *Evaluation with different fusion methods. Constrained by memory size, we use 6890, 1732 and 432 SMPL vertices for GCN, MHSA and GAT, respectively, which represent different downsampling levels.*

| | THuman2.0 | | | 3DPeople | | |
|---|---|---|---|---|---|---|
| Method | Chamfer | P2S | Normal | Chamfer | P2S | Normal |
| OaIF w GCN | 0.855 | 0.767 | **0.149** | 3.228 | 2.061 | **0.176** |
| OaIF w MSA | 1.216 | 1.173 | 0.184 | 3.689 | 2.672 | 0.227 |
| OaIF w GAT | **0.832** | **0.714** | 0.154 | **3.159** | **2.033** | 0.178 |

as feature mapping. This is mainly limited by the fact that the implicit function is defined in the feature space of a single sampled point rather than the product space of several sampled points. Otherwise, it is required a considerable amount of parameters for the implicit function to learn. Therefore, to realize the multi-modal feature fusion while preserving the neighbourhood and global information perception, we propose to use the coding method in topology space with template or anchors as a reference, and then diffuse the hybrid feature to the whole space through continuously differentiable cross attention.

We next test OaIF with different multi-modal feature fusion methods for ablation study. For graph attention network (GAT) and graph convolution network (GCN), we set the receptive field to 16 vertices, thus four residual blocks are stacked with normalized $\hat{A}^2$ as the adjacency matrix. We follow the general transformer settings for multi-head attention (MSA) without positional encoding since coordinates information has been encoded by early GCN.

The quantitative results are shown in Table 3, where the best ones of different matrics are marked in bold. OaIF with MSA underperforms in all metrics with large gap than the others constrained by topology prior. We observe over-smooth attention matrix in training and inference because of the large number of sequence elements. This indicates difficulty of modelling long-distance dependency in our network architecture.

Since the dataset for clothed human re-construction is always limited in poses and cloth style, we still need the locality as an inductive bias to constrain the learning process. For GAT and GCN, because the receptive field is enough large to cover back side of human body, the scale of the used graph does not matter a lot. While a dynamic weighting process in GAT is more powerful to fuse features of different modalities. Thus GAT performs slightly better than general GCN in most metrics. As for GCN using adjacency matrix with more vertices, it is efficient to capture local details. Generally speaking, when enough accurate SMPL parameters can be obtained, a multi-scale framework should be most suitable for hybrid feature fusion since it can learn global and local features simultaneously. This requires changing the 2D encoder to a multi-scale one correspondingly and applying an additional attention module to integrate features from all scales. We leave this for future work.

### 4.5. Analysis about SMPL prior

We observe that most stretching artifacts are near body boundary which is near to background. Since most PIFu-liked methods do not take background into account and simply mask it with white, surrounding pixel-aligned features are prone to be featureless because of large receptive field in 2D pixel grid. While OaIF restrains the influence from background by embedding 2D features to vertices on SMPL model, this can also reduce z-stretching. We present more results from two datasets in Figure 10 which can prove our idea.

For those methods without SMPL prior, they always fail to give reasonable re-construction when the pose is non-standing. That is because of the inherent ambiguity of pixel-aligned features obtained from either the image feature map or the predicted normal map. And the reason why they perform well in standing poses (such as the third row) is that the limbs are usually orthogonal with view direction. Most limbs of the human body are cylinder-liked, thus the front side and backside are very similar in feature space when they are located in the plane orthogonal with the $z$-axis. Consequently, when the limbs rotate around the $x$-axis or $y$-axis, it becomes hard for PIFu and PIFuHD to predict the occupancy of points accumulated along the $z$-axis because of lack of information. This results in their failure in sitting poses in Figure 10.

As for PaMIR, the voxel feature is integrated as global information which is obtained through 3D convolution. Thus, the SMPL prior is not fully utilized because of the low resolution of volume. More importantly, the distance metrics of 3D and 2D feature spaces are defined in Euclidean space. This is unstable for information transmission and neighbourhood feature weighting as poses change. While the graph convolution used in OaIF makes the distance between different body parts constant with varied poses which is important for methods to generalize well and unify the data distribution of training and inference.

## 5. Limitations and Future Work

Although OaIF can improve the feature representation and quality of re-construction compared to PIFu-liked methods, there are still several limitations conditioned in current data and method. The occluded parts of re-constructed clothed human are still over-smooth and lack of details. Then, the proposed method usually fails to reconstruct accurately for cases out-of-distribution of training sets. Last, since we introduce SMPL as a prior, the robust of used human pose and shape estimation method influences a lot. Based on these limitations, multi-view consistency as a prior in training may be explored to better recovery occluded geometry in the future work. A self-supervision framework is also worth being well studied to make training data sets more complete.

## 6. Conclusion

We propose an approach to perceive the visibility information of query points in 3D re-construction based on PIFu-liked methods. We mainly provide hybrid features and occlusion factor for any point based on the SMPL reference vertices, which also can be called anchor points. And the reference point can realize the perception of neighbourhood and long-range information on the topological manifold, which is more reasonable compared to the pixel grid. The hybrid coding of multi-modal features can also avoid the homogenization caused by the depth ambiguity of pixel-aligned fea-

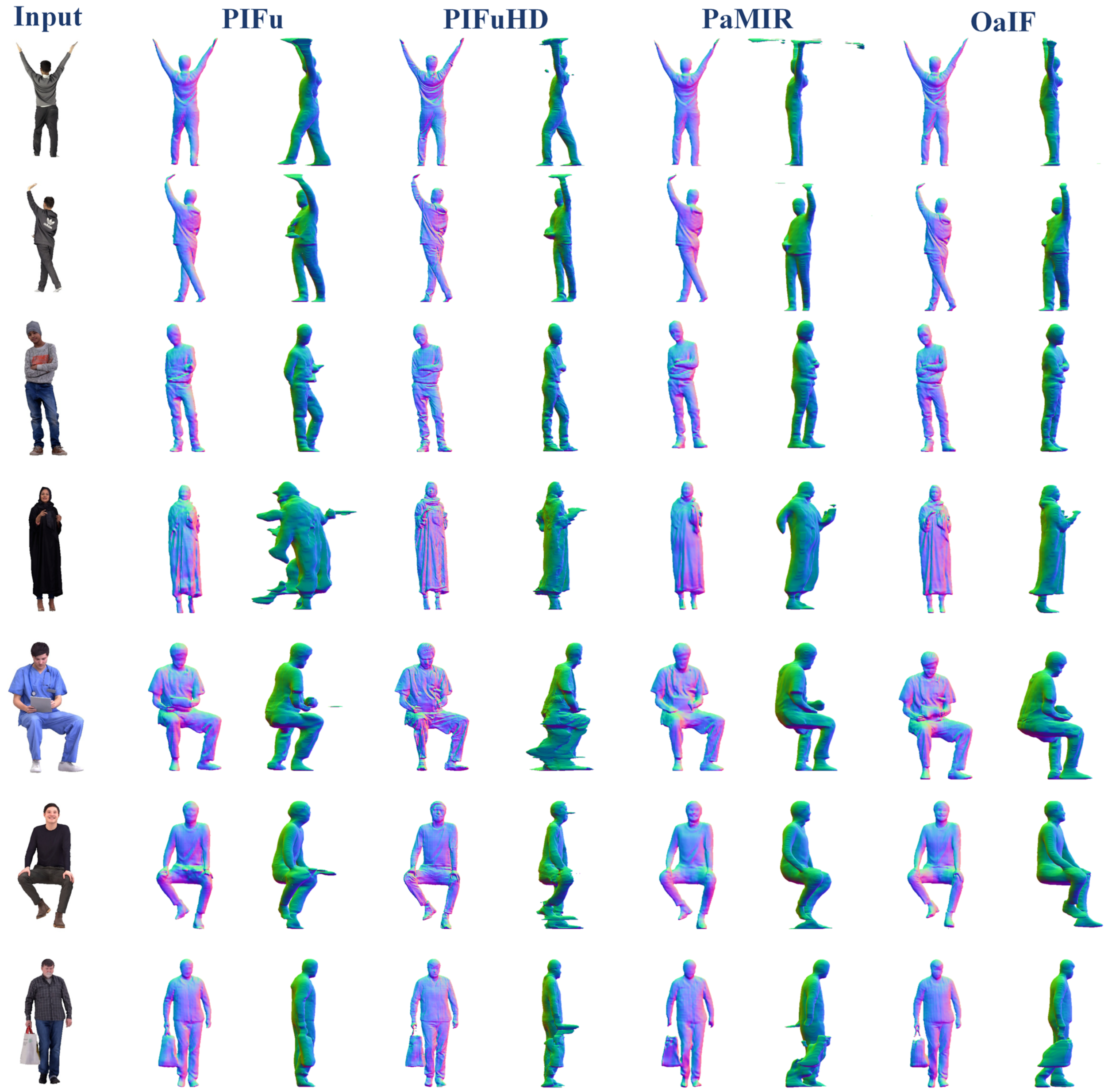

**Figure 10:** *More qualitative results on THuman2.0 and 3DPeople. Though PIFuHD does not require background removed in training and inference, we still do that for it and it is obvious unexpected structures near body boundary still exist. From left to right: input images, results of PIFu [SHN*19],PIFuHD [SSSJ20], PaMIR [ZYLD21] and OaIF (ours).*

tures, and closely couple the image feature with the geometric information of the human body surface. Experiments show that with occlusion perception, the proposed method generalizes better than previous methods in pose and perspective.

## Acknowledgements

This research is supported by the National Key R&D Program of China (No. 2022YFF0903100), and the National Natural Science Foundation of China (No. 61872394).

## References

[AMB*19] ALLDIECK T., MAGNOR M., BHATNAGAR B. L., THEOBALT C., PONS-MOLL G.: Learning to reconstruct people in clothing from a single RGB camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA, USA, 2019), IEEE/CVF, pp. 1175–1186.

[APMTM19] ALLDIECK T., PONS-MOLL G., THEOBALT C., MAGNOR M.: Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International*

*Conference on Computer Vision* (Seoul, South Korea, 2019), IEEE/CVF, pp. 2293–2303.

[BSTPM20a] Bhatnagar B. L., Sminchisescu C., Theobalt C., Pons-Moll G.: Combining implicit function learning and parametric models for 3D human reconstruction. In *Proceedings of the European Conference on Computer Vision* (Glasgow, Scotland, 2020), Springer, pp. 311–329.

[BSTPM20b] Bhatnagar B. L., Sminchisescu C., Theobalt C., Pons-Moll G.: LoopReg: Self-supervised learning of implicit surface correspondences, pose and shape for 3D human mesh registration. In *NeurIPS: Proceedings of the Advances in Neural Information Processing Systems* (Vancouver, Canada, 2020), vol. *33*, pp. 12909–12922.

[BTTPM19] Bhatnagar B. L., Tiwari G., Theobalt C., Pons-Moll G.: Multi-garment Net: Learning to dress 3D people from images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul, South Korea, 2019), IEEE/CVF, pp. 5420–5430.

[CPA*21] Corona E., Pumarola A., Alenya G., Pons-Moll G., Moreno-Noguer F.: SMPLicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN, USA, 2021), IEEE/CVF, pp. 11875–11885.

[CZ19] Chen Z., Zhang H.: Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA, USA, 2019), IEEE/CVF, pp. 5939–5948.

[HCJS20] He T., Collomosse J., Jin H., Soatto S.: Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *NeurIPS: Proceedings of the Advances in Neural Information Processing Systems* (Vancouver, Canada, 2020), vol. *33*, pp. 9276–9287.

[HXL*20] Huang Z., Xu Y., Lassner C., Li H., Tung T.: ARCH: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA, USA, 2020), IEEE/CVF, pp. 3093–3102.

[HXS*21] He T., Xu Y., Saito S., Soatto S., Tung T.: Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), IEEE/CVF, pp. 11046–11056.

[KBJM18] Kanazawa A., Black M. J., Jacobs D. W., Malik J.: End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT, USA, 2018), IEEE/CVF, pp. 7122–7131.

[KHHB21] Kocabas M., Huang C.-H. P., Hilliges O., Black M. J.: PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), IEEE/CVF, pp. 11127–11137.

[KPBD19] Kolotouros N., Pavlakos G., Black M. J., Daniilidis K.: 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* (Seoul, South Korea, 2019), IEEE/CVF, pp. 2252–2261.

[KPD19] Kolotouros N., Pavlakos G., Daniilidis K.: Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA, USA, 2019), IEEE/CVF, pp. 4501–4510.

[LC87] Lorensen W. E., Cline H. E.: Marching cubes: A high resolution 3D surface construction algorithm. In *ACM Siggraph Computer Graphics* (New York, NY, USA, 1987), vol. *21*, ACM, pp. 163–169.

[LMR*15] Loper M., Mahmood N., Romero J., Pons-Moll G., Black M. J.: SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG) 34*, 6 (2015), 1–16.

[LWF*21] Li K., Wen H., Feng Q., Zhang Y., Li X., Huang J., Yuan C., Lai Y.-K., Liu Y.: Image-guided human reconstruction via multi-scale graph transformation networks. *IEEE Transactions on Image Processing 30* (2021), 5239–5251.

[LWL21] Lin K., Wang L., Liu Z.: End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN, USA, 2021), IEEE/CVF, pp. 1954–1963.

[MGB*21] Mehta I., Gharbi M., Barnes C., Shechtman E., Ramamoorthi R., Chandraker M.: Modulated periodic activations for generalizable local functional representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), IEEE/CVF, pp. 14214–14223.

[MST*20] Mildenhall B., Srinivasan P. P., Tancik M., Barron J. T., Ramamoorthi R., Ng R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision* (Glasgow, Scotland, 2020), Springer, pp. 405–421.

[MSY*21] Ma Q., Saito S., Yang J., Tang S., Black M. J.: Scale: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN, USA, 2021), IEEE/CVF, pp. 16082–16093.

[MYTB21] Ma Q., Yang J., Tang S., Black M. J.: The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), IEEE/CVF, pp. 10974–10984.

[NYD16] Newell A., Yang K., Deng J.: Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision* (Amsterdam, Netherlands, 2016), Springer, pp. 483–499.

[PCG*19] PAVLAKOS G., CHOUTAS V., GHORBANI N., BOLKART T., OSMAN A. A., TZIONAS D., BLACK M. J.: 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* (Long Beach, CA, USA, 2019), IEEE/CVF, pp. 10975–10985.

[PFS*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA, USA, 2019), IEEE/CVF, pp. 165–174.

[PHT*21] PATEL P., HUANG C.-H. P., TESCH J., HOFFMANN D. T., TRIPATHI S., BLACK M. J.: AGORA: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN, USA, 2021), IEEE/CVF, pp. 13468–13478.

[QYSG17] QI C. R., YI L., SU H., GUIBAS L. J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the Advances in Neural Information Processing Systems* (Long Beach, CA, USA, 2017), vol. *30*, Curran Associates, Inc.

[SHN*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., KANAZAWA A., LI H.: PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul, South Korea, 2019), IEEE/CVF, pp. 2304–2314.

[SMB*20] SITZMANN V., MARTEL J., BERGMAN A., LINDELL D., WETZSTEIN G.: Implicit neural representations with periodic activation functions. In *NeurIPS: Proceedings of the Advances in Neural Information Processing Systems* (Vancouver, Canada, 2020), vol. *33*, pp. 7462–7473.

[SSSJ20] SAITO S., SIMON T., SARAGIH J., JOO H.: PIFuHD: Multilevel pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA, USA, 2020), IEEE/CVF, pp. 84–93.

[TBTPM20] TIWARI G., BHATNAGAR B. L., TUNG T., PONS-MOLL G.: SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16* (Glasgow, UK, 2020), Springer, pp. 1–18.

[VCC*17] VELIČKOVIĆ P., CUCURULL G., CASANOVA A., ROMERO A., LIO P., BENGIO Y.: Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017). Cornel University. Ithaca, NY.

[VRM*17] VAROL G., ROMERO J., MARTIN X., MAHMOOD N., BLACK M. J., LAPTEV I., SCHMID C.: Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI, USA, 2017), IEEE/CVF, pp. 109–117.

[XWC*19] XU Q., WANG W., CEYLAN D., MECH R., NEUMANN U.: DISN: Deep implicit surface network for high-quality single-view 3D reconstruction. In *NeurIPS: Proceedings of the Advances in Neural Information Processing Systems* (San Diego, CA, USA, 2019), vol. *32*.

[XYTB22] XIU Y., YANG J., TZIONAS D., BLACK M. J.: ICON: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA, USA, 2022), IEEE/CVF, pp. 13296–13306.

[YWM*21] YANG Z., WANG S., MANIVASAGAM S., HUANG Z., MA W.-C., YAN X., YUMER E., URTASUN R.: S3: Neural shape, skeleton, and skinning fields for 3D human modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN, USA, 2021), IEEE/CVF, pp. 13284–13293.

[ZCL*19] ZHANG H., CAO J., LU G., OUYANG W., SUN Z.: DaNet: Decompose-and-aggregate network for 3D human shape and pose estimation. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France, 2019), ACM, pp. 935–944.

[ZJCL21] ZHOU S., JIANG M., CAI S., LEI Y.: DC-GNet: Deep mesh relation capturing graph convolution network for 3D human shape reconstruction. In *Proceedings of the 29th ACM International Conference on Multimedia* (Chengdu, China, 2021), ACM, pp. 171–180.

[ZYLD21] ZHENG Z., YU T., LIU Y., DAI Q.: PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 6 (2021), 3170–3184.

[ZZW*19] ZHU H., ZUO X., WANG S., CAO X., YANG R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA, USA, 2019), IEEE/CVF, pp. 4491–4500.