LASOR: Learning Accurate 3D Human Pose and Shape Via Synthetic Occlusion-Aware Data and Neural Mesh Rendering

Kaibing Yang*, Renshu Gu*, Maoyu Wang, Masahiro Toyoura, Gang Xu
*Equal Contribution

Abstract—A key challenge in the task of human pose and shape estimation is occlusion, including self-occlusions, object-occluded cases, and inter-person occlusions. The lack of diverse and accurate pose and shape training data becomes a major bottleneck, especially for scenes with occlusions in the wild. In this paper, we focus on the estimation of human pose and shape in the case of inter-person occlusions, while also handling object-human occlusions and self-occlusion. We propose a novel framework that synthesizes occlusion-aware silhouette and 2D keypoints data and directly regresses to the SMPL pose and shape parameters. A neural 3D mesh renderer is exploited to enable silhouette supervision on the fly, which contributes to great improvements in shape estimation. In addition, keypoints-and-silhouette-driven training data in panoramic viewpoints are synthesized to compensate for the lack of viewpoint diversity in any existing dataset. Experimental results show that we are among the state-of-the-art on the 3DPW and 3DPW-Crowd datasets in terms of pose estimation accuracy. The proposed method evidently outperforms Mesh Transformer, 3DCrowdNet and ROMP in terms of shape estimation. Top performance is also achieved on SSP-3D in terms of shape prediction accuracy. Demo code will be available at https://figame-lab.github.io/LASOR/.

Index Terms—3D human pose and shape estimation, occlusion-aware, neural mesh renderer, silhouette, 2D keypoint.

I. INTRODUCTION

HUMAN pose and shape estimation from a single image is a challenging 3D vision problem. It may promote many promising applications, such as human-computer interaction, computer graphics, virtual fitting room, video understanding, etc. Recently various deep-learning-based methods have been proposed to address this problem [1]–[5]. Such methods provide impressive 3D human pose estimation results from a single RGB image. However, occlusion remains a major challenge for images in the wild, where self-occlusions, person-object occlusions and inter-person occlusions usually appear. It is difficult to achieve good performance when recovering the full 3D human body shape without explicitly taking occlusions into account. In this paper, we mainly focus on estimating the full 3D human pose and shape from a single RGB image under various occlusion cases.

By exploiting a parametric human body model SMPL [6], the task of human pose and shape estimation is usually converted into SMPL parameters estimation. SMPLify [7] fits the SMPL parameters to obtain a mesh consistent with image evidence, i.e., 2D keypoints. Omran et al. [8] regress the SMPL parameters with a Convolutional Neural Network. Kolotouros et al. [9] directly regress the 3D locations of the mesh vertices using a Graph-CNN. With further development, some targeted questions have been raised. To estimate the shape more accurately, a method is proposed by synthesizing the input data with different shapes [10]. For the object-occluded case, a method that utilizes a partial UV map to represent an object-occluded human body is proposed in [11]. Compared with the object-occluded cases, the person-occluded cases involve more complex scenes and yet are quite common in reality. Fig.1 shows some of such tricky cases. There are two main challenges. The first challenge is the lack of training data. Existing datasets do not contain enough samples with occlusions. Moreover, since training data is traditionally collected by motion capture (MoCap) systems, and due to the limitations of existing MoCap systems, data collection of inter-person occlusions can hardly scale up. Another challenge is that severe inter-person occlusions would introduce challenging ambiguities into the network prediction, and thus confuse the
Fig. 2. Overview of our method. (A) is used to synthesize the training input. The input is then passed through the regressor network (B). We use rendering and compare the input with the rendered images in (C) to supervise the network.

full 3D human body shape estimation.

To tackle the obstacles, inspired by [10], we propose to make use of synthetic keypoints-and-silhouette-driven training data and explicitly design an occlusion-aware framework during training. More specifically, silhouettes that reflect various inter-person occlusions are synthesized, and 2D keypoints are masked out if considered occluded. Using keypoints and silhouettes as the intermediate representation, large amount of inter-person occlusion samples can be generated with a low cost. Moreover, supervision of the silhouettes is enforced during training using a neural 3D mesh renderer integrated with the framework. The loss is backpropagated to the SMPL pose and shape parameters, which can enhance the human pose and shape estimation performance. Besides that, in many existing datasets, the cameras are installed in fixed positions. To further facilitate the network training and empower it with better generalization ability, we propose a viewpoint augmentation approach to create keypoints-and-silhouette-driven training data from panoramic views on top of any existing dataset.

The overall flowchart of the proposed framework is shown in Fig 2. Module A synthesizes the training data on-the-fly with the view augmentation and inter-person occlusions. The data then passes through the encoder and iterative regressor. The rendering and comparison structure is used to supervise the network. During inference, we just need the keypoint detector [12] and DensePose [13] to obtain the keypoints-and-silhouette data as input.

The main contributions of this work are summarized as follows:

- We propose a robust occlusion-aware 3D pose and shape estimation framework that exploits synthetic keypoints-and-silhouette-driven training data, which overcomes the data scarcity especially for person-occluded scenes in the wild.
- Neural 3D mesh renderer is integrated with the framework to provide silhouette supervision during training, which contributes to much more accurate shape estimation.
- A viewpoint augmentation strategy is introduced to synthesize infinite viewpoint variation, creating keypoints-and-silhouette-driven training data from panoramic views on top of any existing dataset.

II. RELATED WORK

In this section, we discuss recent 3D human pose and shape estimation approaches, as well as the related methods for the occlusion problem.

Optimization-based methods. Optimization-based approaches attempt to fit a parametric body model, like SMPL [6] or SCAPE [14], to 2D observations. SMPLify [7] is the first method to automatically estimate 3D pose and shape from a single unconstrained image by fitting the SMPL to 2D keypoint detections. Other than 2D keypoint detections, different cues like body surface landmarks [15], silhouettes [15] or body part segmentation [3] are also used in the fitting procedure. Recent works try to fit a more expressive model [16], [17] that includes face, body, and hand details. Gu et al. [18] adopt a hierarchical 3D human model that imposes angle and bone length constraints for multi-person pose estimation. These approaches produce reliable results with good 2D observations. However, they are sensitive to initialization and rely heavily on the quality of 2D information.

Regression-based methods. To this end, recent works rely almost exclusively on deep neural networks to regress the 3D human pose and shape parameters. Owing to the lack of
training data with full 3D shape ground truth, 2D annotations including 2D keypoints, silhouettes, or parts segmentation are usually essential. This information can be used as intermediate representation [19], [20], or as supervision [5], [19]. Kolotouros et al. [9] directly regress the 3D location of the mesh vertices from a single image instead of predicting model parameters using a Graph-CNN. Pose2mesh [21] takes a 2D pose as input and avoids the domain gap between training and testing. I2L-MeshNet [22] predicts the per-pixel (line+pixel) likelihood on 1D heatmaps for each mesh vertex coordinate instead of directly regressing the parameters. Fundamentally, regression-based methods are dependent on the sample diversity of the training data. However, since obtaining 3D training data is awfully expensive, most datasets only contain a limited number of human bodies. This is also the reason why the existing datasets are insufficient to support robust and accurate human body shape estimation in the wild.

Methods for occlusion problems. Huang and Yang [23] propose a method to recover 3D pose when a person is partially or severely occluded in the scene from monocular images. However, the occlusions are limited to simple rectangles. A grammar-based model is introduced by [24] with explicit occlusion part templates. To avoid specific design for occlusion patterns, Ghiassi et al. [25] present a method to model occlusion by learning deformable models with many local part mixture templates using large quantities of synthetically generated training data, which is aimed at explicitly learning the appearance and statistics of occlusion patterns. Zhang et al. [11] present a method that utilizes a partial UV map to represent an object-occluded human body, and the full 3D human shape estimation is ultimately converted as an image inpainting problem. The above methods mainly deal with the object-human occlusion problems. Bin et al. [26] present a method that augments images by pasting segmented body parts with various semantic granularity, which has certain generalization ability to shield the human body from each other. The occlusion problem is explored in [27] by using a temporal gated convolution network. Choi et al. [28] present 3DCrowdNet, a 2D human pose-guided 3D crowd pose and shape estimation system for in-the-wild scenes. Sun et al. [29] develop a novel Collision-Aware Representation (CAR) method named ROMP to improve the performance under person-person occlusion. Jiang et al. [30] propose a depth ordering-aware loss to generate a rendering that is consistent with the annotated instance segmentation, but they make use of very few datasets with 3D ground truth. Currently, occlusion is still a major challenge for 3D pose estimation in the wild.

Synthetic training data. When the data is too scarce to solve the problem, synthetic data becomes an option. Several papers explore the use of synthetic data in training for human pose and shape estimation. To make up for the scarcity of in-the-wild training data with diverse and accurate body shape labels, Sengupta et al. [10] propose STRAPS, a system that utilizes proxy representations, such as silhouettes and 2D joints, as inputs to a shape and pose regression neural network, which is trained with synthetic training data. Some methods [11], [26] try to obtain novel samples either by adding virtual objects to existing datasets or by pasting segmented body parts

with various semantic granularity. However, while inter-person occlusion is common in 3D human pose estimation scenarios, the inter-person occlusion is rarely synthesized explicitly.

III. TECHNICAL APPROACH

In the following section, we first describe the Skinned Multi-Person Linear model (SMPL) briefly and the basic notations defined in this paper. More details about the data generation are then provided. After that, the regression network and the loss function are presented in detail.

A. SMPL Model

SMPL is a skinned vertex-based model that accurately represents a wide variety of body shapes in natural human poses. SMPL model provides the function \( M(\theta, \beta) \) that takes as input the pose parameters \( \theta \) and the shape parameters \( \beta \) and returns the body mesh \( v \in \mathbb{R}^{N \times 3} \) with \( N = 6890 \) vertices. The 3D joints are calculated by \( j^{3d} = J \cdot e \), where \( J \in \mathbb{R}^{K \times N} \) is a pre-trained linear regressor matrix, and \( K \) is the number of joints.

B. View Augmentation

For the datasets collected in the laboratory, the camera is usually placed in a fixed position. Meanwhile, datasets collected outdoors with 3D ground truth are scarce and of insufficient variations as well. Models trained using these datasets may depend on the camera’s perspective and have poor generalization ability. To this end, this paper adopts a viewpoint augmentation strategy to generate intermediate representations, i.e., keypoints and silhouettes, under different camera parameters. Using these intermediate representations can make full use of this feature and reduce the model’s dependence on a specific camera perspective. In the implementation, since the first three values of the SMPL model pose parameters control the global orientation of the human body model, we alter these parameters to modify the current human body orientation relative to the camera. Modifying the global orientation of the human body model does not modify the 3D posture of the human body, but the projection on the 2D image is quite different. The keypoints and contour maps of the input model in this paper are all two-dimensional data. This greatly enriches the training data when the camera viewpoint in the training data is limited. The specific formula is given as follows:

\[
\hat{\theta}_g = \theta_g + \pi \cdot N(\mu, \sigma^2)
\]

where \( \theta_g \) is the global orientation in the \( \theta \) parameter, and \( N \) is a normal distribution function with mean \( \mu \) and standard deviation \( \sigma \).

C. Synthetic Data Generation

As shown in Fig. 3, two pairs of SMPL parameters \((\theta_1, \beta_1), (\theta_2, \beta_2)\) are sampled from any training dataset. The camera translation vector \( t_1 \) is generate randomly, while the camera intrinsic matrix \( K \) and rotation matrix \( R \) are fixed.
Then, we obtain \( t_2 \) based on \( t_1 \) by adding a shift. In order to get various shape data, we replace \( \beta \) with a new random vector \( \beta' \), generated by sampling \( n \) shape parameters \( \beta_n \sim N(\mu, \sigma^2) \) [10]. We use \( \theta_1 \) and \( \theta_2 \) by replacing the global orientation with the orientation \( \theta_g \) obtained by viewpoint augmentation method instead of the original \( \theta_1 \) and \( \theta_2 \). Then, the 3D vertices \( v_1 \) and \( v_2 \) corresponding to \((\theta_1, \beta) \) and \((\theta_2, \beta') \) are rendered [19] into silhouette \( S_1, S_2 \in [0, 1]^H \times W \). The final silhouette is calculated by (2) as follows,

\[
S = S_1 - S_1 * S_2
\]  

(2)

As shown in Fig. 4 inter-person occlusion samples are generated with different levels of human body overlap.

For 2D joints, the 3D joints \( \hat{j}^{3d} \) corresponding to \((\theta_1, \beta) \) are projected into 2D joints \( j^{2d} \in \mathbb{R}^{K \times 2} \) using prospective projection. \( j^{2d} \) is encoded into 2D Gaussian joint heatmaps, \( G \in \mathbb{R}^{K \times H \times W} \). We obtain the final input \( X \in \mathbb{R}^{(K+1) \times H \times W} \) by concatenating \( S \) and \( G \). Note that the noises in keypoints detection will introduce noises into the 2D joint heatmap encoding. To handle occluded joints explicitly, we randomly remove some body parts from the human body and add occluding boxes to the silhouette in \( S \) during training.

**D. Regression Network**

Our method is architecture-agnostic. We use the baseline network architecture as in [2], [10], [31], which consists of a convolutional encoder for feature extraction and an iterative regressor that outputs predicted SMPL pose, shape and camera parameters \((\theta, \beta \text{ and } \Pi)\) from the extracted feature. Similarly as in STRAPS [10], the SMPL default discontinuous Euler rotation vectors are replaced by a continuous 6-dimensional rotation representation of \( \theta \) proposed by [32].

The weak-perspective camera model is adopted in this paper, represented by \( \Pi = [s, t] \) where \( s \in \mathbb{R} \) means scale and \( t \in \mathbb{R}^2 \) represents \( x - y \) camera translation. These parameters allow us to generate the mesh corresponding to the regressed parameters, \( \hat{v} = M(\hat{\theta}, \hat{\beta}) \), as well as the 3D joints \( \hat{j}^{3d} = J\hat{v} \) and their 2D reprojection \( \hat{j}^{2d} = \Pi(\hat{j}^{3d}) \).

**Neural 3D Mesh Renderer.** Differentiable rendering is a novel field which allows the gradients of 3D objects to be calculated and propagated through images. We follow the common self-supervision pipeline with differentiable rendering and don’t need correspondence to images. A neural 3D mesh renderer [19] is integrated into our framework to provide silhouette supervision during training. The loss function is defined as

\[
L_S = \frac{1}{H \times W} \times \left\| \hat{S} - S \right\|_2^2
\]

(3)

where \( S \) is the target silhouette image, \( \hat{S} \) is silhouette rendered from \( \hat{v} \) and \( H, W \) are the height and width of the target silhouette \( S \). The reconstruction of 3D shape is an ill-posed problem when a single view is used during training. Additional loss functions are added to provide strong supervision.

The first loss term \( L_v \) performs the supervision between predicted 3D vertices \( \hat{v} \) and ground-truth 3D vertices \( v \), which is,

\[
L_v = \left\| \hat{v} - v \right\|_2^2
\]

(4)

The second loss term \( L_{j^{3d}} \) performs the supervision between predicted 3D joints and ground-truth 3D joints, which is,

\[
L_{j^{3d}} = \left\| \hat{j}^{3d} - j^{3d} \right\|_2^2
\]

(5)

The \( L_\theta \) and \( L_\beta \) losses can provide strong 3D supervision.

\[
L_\theta = \left\| \hat{\theta} - \theta \right\|_2^2
\]

(6)

\[
L_\beta = \left\| \hat{\beta} - \beta \right\|_2^2
\]

(7)
In order to enforce image-model alignment, we add the loss for the projected 2D joints

\[ L_{j_{2d}} = \frac{1}{K} \sum_{i=1}^{K} \omega_i \left \| j_{2d}^i - \hat{j}_{2d}^i \right \|^2 \]

where the value of \( \omega_i \) indicates the visibility (1 if visible, 0 otherwise) for i-th 2D keypoints (K in total). The final loss function can be described as follows:

\[ L = \frac{1}{\sigma_v} L_v + \frac{1}{\sigma_{j_{2d}}} L_{j_{2d}} + \frac{1}{\sigma_\theta} L_\theta + \frac{1}{\sigma_\beta} L_\beta + \frac{1}{\sigma_{j_{2d}}} L_{j_{2d}}^d \]

\[ + \frac{1}{\sigma_S^2} L_S + \log (\sigma_v \sigma_{j_{2d}} \sigma_\theta \sigma_\beta \sigma_{j_{2d}} \sigma_S) \]

where losses are adaptively combined using homoscedastic uncertainty [33]. In the next section, we will analyze the contribution of the silhouette loss \( L_S \) in 3D shape estimation accuracy.

IV. EXPERIMENTS

A. Datasets

This section gives a description of the datasets used for training and evaluation. Following STRAPS [10], we train the network using synthetic data generated from 3DPW [34], AMASS [35], and UP-3D [15]. We report evaluation results on 3DPW, 3DOH50K [11] and SSP-3D [10].

AMASS is a large and varied database of human motion that unifies 15 different optical marker-based mocap datasets by representing them within a common framework and parameterization. Some of the SMPL parameters are selected to train our network.

UP-3D combines two LSP datasets and the single-person part of MPII-HumanPose dataset, which is used to train our network.

3DPW is captured via IMUs and contains both indoor and outdoor scenes. It provides abundant 2D/3D annotations, such as 2D pose, 3D pose, SMPL parameters, human 3D mesh, etc. It includes inter-person occlusion, object-human occlusion, and non-occluded/truncated cases. Being a rare in-the-wild 3D human pose and shape dataset, 3DPW is the most suitable for testing our work. 3DPW is used for training and evaluation with the same split as in [10].

3DPW-Crowd is a subset of the 3DPW validation set and the selected sequences are courtyard_hug_00 and courtyard_dancing_00. 3DPW-Crowd is deemed a suitable benchmark for evaluating performance for inter-person occlusion scenes in [36], as it has much higher bounding box IoU and CrowdIndex [37], which measures crowding level.

3DOH50K contains 311 in-the-wild images of 62 tightly clothed sports persons (selected from the Sport-1M video dataset) with a diverse range of body shapes, along with the corresponding pseudo-ground-truth SMPL shape and pose labels.

3DOH50K is the first 3D human pose dataset that introduces object-human occlusions on purpose when recording. It contains 1290 test images. This dataset is used for evaluation.

Dataset Statistics. Our work mainly focuses on human pose and shape estimation with inter-person occlusions. We compare different datasets to give insights about the types of occlusion and the number of people in Table I.

| Dataset       | Object-Occluded | Inter-person Occlusion | Num of People |
|---------------|-----------------|------------------------|---------------|
| 3DPW          | ++              | ✓                      | 18            |
| 3DPW-Crowd    | -               | ✓                      | 2             |
| SSP-3D        | -               | ✓                      | 62            |
| 3DOH50K       | +++             | ✓                      |               |

B. Implementation Details

Network Architecture. For a fair comparison with previous methods, we use fine-tuned ResNet-18 [38] as the default encoder and ResNet-50 as the additional encoder. The encoder produces a feature vector \( \phi \in \mathbb{R}^{512} \). The iterative regression network consists of two fully connected layers. During inference, we use darkpose [39] to get 2D keypoints of 3DPW and use Mask-RCNN [12] to get 2D keypoint predictions of other datasets. We use DensePose [13] to get silhouette predictions. All implementations are done in PyTorch [40].

Setting Details. The input silhouettes are cropped and resized to 256 × 256, while keeping the aspect ratio of 1.2. COCO keypoint order is used and the number of keypoint is 17. The input tensor is 18 × 256 × 256. The loss weights are set to be \( \sigma_v = 1.0, \sigma_{j_{2d}} = 1.0, \sigma_\theta = 0.1, \sigma_\beta = 0.1, \sigma_{j_{2d}} = 0.1, \sigma_S = 0.1 \). We use the Adam optimizer [41] with a batch size of 140 and an initial learning rate of 0.0001 to train our encoder and regressor. The mean values in iterative regression are the same as SPIN [2]. The training step takes 10 days with 100 epochs on a single 3090 GPU.

Training Datasets. For a fair comparison with the previous method STRAPS [10], we use the training data provided by the STRAPS [10]. The training data are collected from AMASS, 3DPW and UP-3D. In the training step, we use the ground-truth SMPL parameters and do not use any pair of image and 3D label.

Evaluation Metrics. We adopt Procrustes-aligned [42] mean per joint position error (MPJPE_PA) for evaluating the 3D pose accuracy. To evaluate the 3D shape error, we employ Procrustes-aligned per-vertex error (PVE_PA). Besides, to evaluate the shape estimation, we calculate the scale-corrected per-vertex euclidean error in a neutral pose (PVE-T_SC) and mean intersection-over-union (mIOU) on SSP-3D. In all the tables, ↑ means the larger the better, otherwise it is the lower the better.

C. Comparisons to the State-of-the-Art:

To demonstrate the effectiveness of our method, we perform quantitative evaluations on 3DPW, 3DOH50K and SSP-3D. In Table I the methods in the upper rows do not use any paired images and 3D labels. Separated by the horizontal line, the methods in the bottom rows use the paired ground-truth 3D labels and images during training. Our method does not directly use any real images in the training step.
Fig. 5. **Qualitative comparison with the state-of-the-art methods.** Some samples with different levels of inter-person occlusions from 3DPW dataset are shown. Compared with Mesh Transformer [28], ROMP [29] and STRAPS [10], our method gives more accurate pose and shape predictions for the person-occluded cases.

**3DPW.** Firstly, we evaluate the test dataset in 3DPW. As shown in Table 11, our method achieves a 12 percent improvement compared with [10]. Although our method does not use paired images and 3D labels, it shows competitive performance with the state-of-the-art that requires training data comprised of paired images and 3D labels. Imposing less restrictions on the training data means our method can scale up easier without relying on paired images and labels. Our method also shows excellent shape estimation performance. Note that while Mesh Transformer [28], 3DCrowdNet [36] and ROMP [29] report higher MPJPE-PA accuracy on 3DPW, the proposed method clearly shows better performance on the silhouette for human with various shapes, as shown in Fig. 6.

**3DPW-Crowd.** We also evaluate the proposed method for 3DPW-Crowd. As shown in Table 11, our method achieves the best performance among the methods that do not need paired training data (top rows), including STRAPS [10] and Pose2Mesh [21].

**SSP-3D.** The SSP-3D dataset contains 62 subjects. The proposed method shows the top performance on shape estimation in terms of PVE-T-SC.

**3DOH50K.** In 3DOH50K dataset, the samples have severe occlusions. As shown in Table 11, better performance is achieved compared with [10] on handling object-human occlusions. The
D. Ablation Study

TABLE II
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON 3DPW, 3DPW-CROWD, 3DOH50K AND SSP-3D. NUMBER MARKED WITH * WERE EVALUATED USING THE OPEN-SOURCED OFFICIAL CODE. NUMBER MARKED WITH - WERE PROVIDED BY [10], ALL OTHER NUMBERS ARE REPORTED BY THE RESPECTIVE PAPERS. IN ORDER TO BE MORE CONCISE, THE BEST PERFORMANCE GIVEN BY METHODS WITHOUT USING THE PAIRED GROUND-TRUE 3D LABELS AND IMAGES DURING TRAINING ARE MARKED IN BOLD BLACK, AND OTHER METHODS ARE MARKED IN LIGHT GREEN.

| Method         | 3DPW MPJPE-PA | 3DPW-Crowd MPJPE-PA | 3DOH50K MPJPE-PA | SSP-3D PVE-T-SC | mIOU↑ |
|----------------|----------------|---------------------|-----------------|----------------|-------|
| STRAPS [10]    | 66.8           | 70.5                | 73.5            | 15.9           | 0.8   |
| HMR(unpaired)  | -              | -                   | 20.8            | 0.61           | -     |
| Pose2Mesh [23] | 58.3           | 79.8                | -               | -              | -     |
| Ours           | 57.9           | 67.6                | 72.5            | 14.5           | 0.67  |
| HMR [51]       | 76.7           | -                   | 92.9            | 22.9↑          | 0.69↑ |
| SPIN [27]      | 59.2           | 69.9                | 73.6            | 22.2↑          | 0.70↑ |
| ROMP [29]      | 54.9           | 43.9                | -               | -              | -     |
| 3D-CrowdNet [28] | 57.7     | 73.5                | -               | -              | -     |
| Mesh Transformer [28] | 52.2 | 56.8               | -               | -              | -     |
| OOH [11]       | 47.9           | -                   | 58.5            | -              | -     |

Silhouette Loss. To study the effect of the neural-mesh-rendering-enabled silhouette loss $L_S$ in [3], we compare two models trained (i) without the silhouette loss, and (ii) with the silhouette loss. The results are shown in Table III. The qualitative comparison with the baseline and silhouette loss are shown in Fig. 7. The silhouette loss improves both the network convergence and the final performance. For the model with silhouette supervision, training converges within 100 epochs, whereas the baseline model takes 250 epochs. Note that the baseline uses the Human3.6m dataset while our method does not use it. Our method still gives better shape estimation in terms of PVE-T-SC.

Silhouette As Intermediate Representation. To study the impact of having silhouette as intermediate representation, we compare two models trained with (i) keypoint only, and (ii) keypoint and silhouette, both with the process of synthetic data generation. The results are shown in Table IV. The method with silhouette as intermediate representation consistently shows a higher performance. This proves that the combination of keypoints and silhouette representations is necessary.

View Augmentation and Inter-person Occlusion Augmentation. We also investigate the effects of view augmentation (VA) and inter-person occlusion augmentation (IPOA). As shown in Table V our method can improve the accuracy under both evaluation standards. Some qualitative results of adding the VA and IPOA modules are visualized and compared in Fig. 8. The model with both VA and IPOA can handle unusual viewpoints and severe occlusions well, which yields the best performance.

TABLE III
EVALUATION ON THE SSP-3D DATASET.

| Method                  | PVE-T-SC | mIOU↑ | convergence time (epochs) |
|-------------------------|----------|-------|----------------------------|
| W/O silhouette loss     | 14.5     | 0.67  | 100                        |
| W/ silhouette loss      | 17.2     | 0.67  | 250                        |

TABLE IV
ABLATION STUDY ON SILHOUETTE AS THE INTERMEDIATE REPRESENTATION.

| Method                   | 3DPW PVE_PA | 3DPW MPJPE_PA | SSP-3D PVE_T-SC | mIOU↑ |
|--------------------------|-------------|---------------|-----------------|-------|
| Keypoint                 | 78.8        | 61.9          | 18.3            | 0.57  |
| Keypoint+Silhouette      | 72.8        | 57.9          | 14.5            | 0.67  |

TABLE V
ABLATION STUDY OF VA AND IPOA MODULES ON 3DPW. THE BASELINE WITH * MEANS THAT RESNET-50 IS THE ENCODER.

| Method                  | 3DPW PVE_PA | 3DPW MPJPE_PA |
|-------------------------|-------------|---------------|
| Baseline                | 77.8        | 62.1          |
| Baseline+VA             | 75          | 59.4          |
| Baseline+VA+IPOA        | 73.2        | 58.6          |
| Baseline*+VA+IPOA       | 72.8        | 57.9          |

TABLE VI
EVALUATION ON THE 3DPW DATASET USING GROUND-TRUE KEYPOINTS AND SILHOUETTE.

| Method                  | 3DPW PVE_PA | 3DPW MPJPE_PA |
|-------------------------|-------------|---------------|
| Baseline                | 44.3        | 35.0          |
| Baseline+VA             | 40.8        | 32.1          |
| Baseline+VA+IPOA        | 38.1        | 30.4          |

Experiments with Ground Truth 2D Inputs. We also report the performance of the proposed method using ground-truth keypoints and silhouettes. The results are shown in Table VI. Compared with noisy 2D keypoint detection and silhouette, the input of 2D ground-truth improves the accuracy by a large margin. This indicates that if the proposed framework is equipped with a better 2D keypoint detector and silhouette segmentation method, much better results will be obtained.
Fig. 6. Qualitative comparison for different body shapes with the state-of-the-art methods 3DCrowdNet [36], Mesh Transformer [28], ROMP [29] and STRAPS [10]. The samples are from the SSP-3D dataset. The results of ROMP [29], STRAPS [10], and Mesh Transformer [28] are generated from their corresponding official codes, and the results of 3DCrowdNet [36] are provided by the author. Our method is able to accurately predict a diverse range of body shapes.

| TABLE VII | EVALUATION OF THE 3DPW DATASET. THE OPERATION OF SYNTHESIZING THE SILHOUETTE CAN GENERATE DIFFERENT LEVELS OF INTER-PERSON OCCLUSIONS. PERCENT MEANS THE PROBABILITY OF PERFORMING THE OPERATION. |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Percentage | PVE_PA | MPJE_PA |
| 40%        | 76.9   | 62.1    |
| 80%        | 74.8   | 60.3    |
| 100%       | 73.2   | 58.6    |

| TABLE VIII | ABLATION STUDY ON ADJUSTING THE CONFIDENCE THRESHOLD OF THE KEYPOINT DETECTION FOR THE 3DPW DATASET. |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Threshold  | PVE_PA | MPJE_PA |
| 0          | 73.9   | 59.3    |
| 0.1        | 73.7   | 59.1    |
| 0.2        | 73.5   | 58.8    |
| 0.3        | 73.3   | 58.6    |
| 0.4        | 73.2   | 58.7    |
| 0.5        | 73.3   | 58.7    |
| 0.6        | 73.9   | 59.3    |

We also evaluate the effects of using different percentages of data for inter-person occlusion synthesis. 40, 80 and 100 percentage of the data are processed respectively, and the corresponding results are compared in Table VII. Using 100
percentage of training data to synthesize occlusions, the best accuracy is achieved on the 3DPW dataset. This reasons well since occlusions are extremely common in 3DPW.

In the 2D keypoint detector, each joint prediction is associated with a confidence score. When the predicted confidence score is above a threshold, the joint is considered detected correctly. The keypoints with confidence lower than the threshold are removed. We test the effects of adjusting the confidence threshold. The results on 3DPW are shown in Table VIII. For samples with severe occlusions, by removing the poorly detected 2D keypoints properly, better results are achieved. The line chart about the MPJPE_PA is shown in Fig. 9. Empirically, setting the threshold to be 0.4 yields the best performance.

![Fig. 7. Ablation study of the silhouette loss enabled by neural 3D mesh rendering.](image1)

![Fig. 8. Effects of view augmentation (VA) and inter-person occlusion augmentation (IPOA). The model with both VA and IPOA yields the best performance.](image2)

V. CONCLUSIONS

In this paper, we propose a new framework that synthesizes occlusion-aware silhouette and 2D keypoints representation to overcome data scarcity when regressing the human pose and shape, which proves to work well on inter-person occlusions as well as other occlusion cases. Differentiable rendering is integrated to enable silhouette supervision on the fly, contributing to higher accuracy in the pose and shape estimation. A new method for synthesizing keypoints-and-silhouette-driven data in panoramic viewpoints is presented to increase viewpoint diversity on top of existing dataset. Thanks to these strategies, the experimental results demonstrate that the proposed framework is among the start-of-the-art methods on the 3DPW and 3DPW-Crowd datasets in terms of pose accuracy. Although it is unfair to compare with these methods using paired training data, the proposed method evidently outperforms Mesh Transformer [28], 3DCrowdNet [36] and ROMP [29] in terms of shape estimation. Top performance is also achieved on SSP-3D in terms of human shape estimation accuracy. The occlusion-aware synthetic data generation strategy, neural-mesh-renderer-enabled silhouette supervision and viewpoint augmentation strategy can be also applied to other approaches for 3D human pose and shape estimation to improve the overall performance of the community.

**Future Work.** Our occlusion-aware method can be extended to the crowded people case. Video information can also be considered to obtain more consistent and accurate human pose and shape. Additionally, compared with other methods, limited training data is used for now. In the next step, more datasets like Human3.6M and COCO can be added in the training procedure to train the model.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for providing constructive suggestions and comments, that contributed to highly improve this work. The authors would like to thank the authors of 3DCrowdNet, Dr. Hongsuk Choi et al., for providing the experimental results on the SSP-3D dataset in Fig. 6.
Kaibing Yang received the Bachelor’s degree in Electronic Information Engineering in 2017 from Anhui Polytechnic University, P.R. China. He is currently a graduate student in the College of Computer Science and Technology, Hangzhou Dianzi University. His research interests are on geometric modeling and computer vision.

Renshu Gu received the B.Sc. degree from Nanjing University in 2011, and the Master’s and Ph.D. degrees from University of Washington in 2020, advised by Professor Jenq-neng Hwang. She is now with the College of Computer Science and Technology, Hangzhou Dianzi University, P.R. China. Her research interests include image/video analytics, computer vision and multimedia. Dr. Gu is a member of IEEE.

Maoyu Wang graduated from Harbin Institute of Technology in 2016 with a bachelor’s degree in Internet of Things Engineering. He is currently a graduate student in Hangzhou Dianzi University ITMO Joint Institute. His research interests include geometric modeling and 3D reconstruction.

Masahiro Toyoura received the B.Sc. degree in Engineering, M.Sc. and Ph.D. degrees in Informatics from Kyoto University in 2003, 2005 and 2008 respectively. He is currently an Associate Professor at Department of Computer Science and Engineering, University of Yamanashi, Japan. His research interests are digital fabrication, computer and human vision. He is a member of IEEE and ACM.

Gang Xu received the B.Sc. degree in Mathematics from Shandong University in 2003, Ph.D. degrees in Mathematics from Zhejiang University in 2008. He is currently a full Professor at College of Computer Science and Technology, Hangzhou Dianzi University, P.R. China. His research interests are geometric modeling and simulation, computer graphics and 3D visual computing. Dr. Xu is a distinguished member of China Computer Federation.