# Pose2UV: Single-shot Multi-person Mesh Recovery with Deep UV Prior

Buzhen Huang, Tianshu Zhang, and Yangang Wang, *Member, IEEE*

***Abstract*—In this work, we focus on the task of multi-person mesh recovery from a single color image, where the key issue is to tackle the pixel-level ambiguities caused by inter-person occlusions. Overall, there are two main technical challenges when addressing the ambiguities: how to extract valid target features under occlusions and how to reconstruct reasonable human meshes with only a handful of body cues? To deal with these problems, our key idea is to utilize the predicted 2D poses to locate and separate the target person, and reconstruct them with a novel learning-based UV prior. Specifically, we propose a visible pose-mask module to help extract valid target features, then train a dense body mesh prior to promote reconstructing natural mesh represented by the UV position map. To evaluate the performance of our proposed method under occlusions, we further build an in-the-wild 3D multi-person benchmark named as 3DMPB. Experimental results demonstrate that our method achieves state-of-the-art compared with previous methods. The dataset, codes are publicly available on our website.**

***Index Terms*—Multi-person mesh recovery, human prior, visible pose and mask estimation.**

## I. Introduction

MULTI-PERSON mesh recovery from color images may promote a broad spectrum of applications as diverse as sports science, human behavioral modeling, holographic communication, *etc*. Conventionally, high-quality multiple human mesh recovery relies on multi-view systems [1]–[3], which are expensive and complicated. Recent years have witnessed frontier progress in single-shot multi-person mesh recovery due to the rapid progress of deep learning [4], [5]. Nevertheless, most of the existing methods are mainly suitable for the scenario of large spatial distances among people. It is noted that, due to the complex inter-person occlusions, pixel-level ambiguities make it hard to achieve good performance in multi-person scenes.

To solve the inter-person ambiguities in a single color image, a few methods [6], [7] use bounding boxes to crop the individual person and estimate SMPL parameters [8] for each one. However, these methods are not able to recover the meshes of interacting people since pixel-level inter-person occlusions are not explicitly taken into account. The occluded people in the image would further introduce severe ambiguities into network training and confuse the neural networks (*e.g*., 2D/3D pose estimation [9], [10]). Recently, centerHMR [11] provides a bottom-up solution to regress a parameter map from a single color image, which tries to avoid the above issue. However, the parameter-based representation still suffers from the non-unique problem (*i.e*., periodicity) of predicting occluded joints rotation.

Generally speaking, there are two main challenges in multi-person mesh recovery from a single color image. The first one is that it is difficult for the network to distinguish the target person and obtain his/her valid information from occluded human images, even if the images are cropped via bounding boxes. The second one is that only a little visible information could be used for the target human mesh recovery. To address the obstacles, our key idea is to use predicted 2D poses, accurate or not, to locate and separate the individual person, and reconstruct the occluded target human from the visible information. We utilize UV representation to describe human meshes, which provides dense correspondences for explicitly considering occlusions. With the UV representation, the performance of the mesh recovery is also boosted by a novel learning-based UV prior.

We first introduce a visible pose-mask module to avoid pixel-level ambiguities. This module estimates the visible heatmaps of the target person guided by a 2D pose detector [12], and then the estimated visible heatmaps are sequentially served as guidance for the visible human segmentation (see Fig.1). Our visible pose-mask module is suitable for any off-the-shelf 2D pose estimation methods (*e.g*., [12]–[14]) and can output both visible heatmaps and the mask of the target person. Compared to the human parser with bounding boxes, visible heatmaps and the mask are more suitable in multi-person scenarios. It is worth mentioning that the 2D pose detector in our module is also commonly required for conventional human mesh recovery methods (*e.g*., HMR [15]). Thus, the proposed strategy only consumes a little extra computation compared with previous works.

Even so, valid information of the occluded person extracted by the visible pose-mask module is still insufficient for reconstructing detailed human meshes. Thus, we propose a human mesh prior to handle this problem. Specifically, we choose a UV position map similar to [16] for human representation. This map-based representation could avoid highly non-linear mapping in regressing SMPL parameters and is suitable for convolution networks to reduce computational cost compared

This work was supported in part by the National Natural Science Foundation of China (No. 62076061), the "Young Elite Scientists Sponsorship Program by CAST" (No. YES20200025), and the "Zhishan Young Scholar" Program of Southeast University (No. 2242021R41083).

Buzhen Huang, Tianshu Zhang and Yangang Wang are with the School of Automation, Southeast University, Nanjing, China, 210096. All the authors from Southeast University are affiliated with the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China.

Buzhen Huang and Tianshu Zhang are co-first authors. The authorship was determined by a coin toss.

Corresponding author: Yangang Wang. E-mail: yangangwang@seu.edu.cn. Personal website: http://www.yangangwang.com.

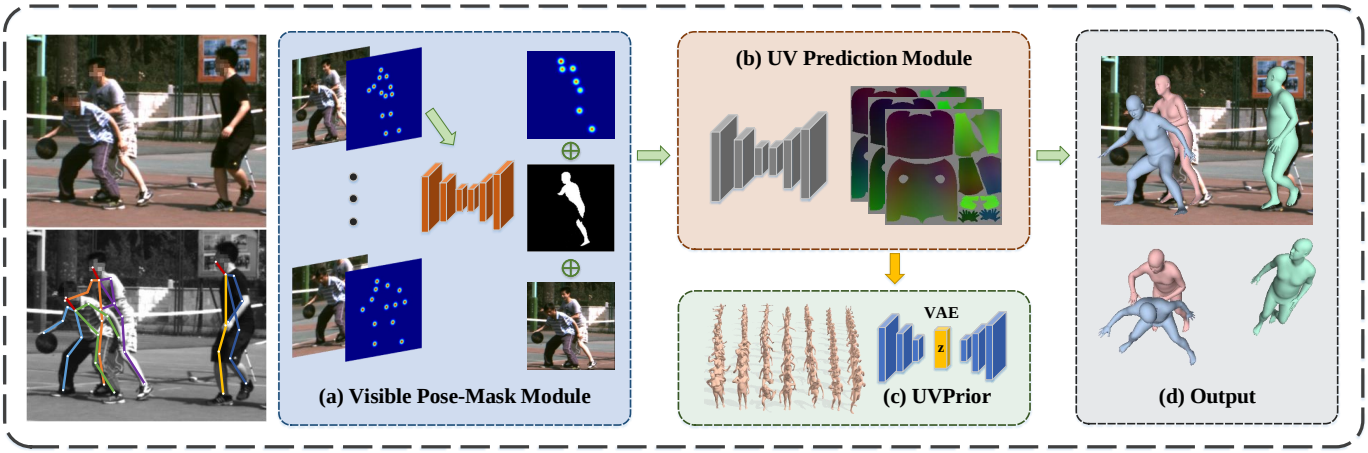



Fig. 1. **Overview of the proposed framework.** Given a challenging multi-person image, we first utilize the predicted 2D pose to locate and crop each individual. The heatmaps and image patch of each person are then fed to the visible pose-mask module (a) to estimate the visible heatmaps and masks. With the help of the proposed UVPrior (c), the UV Prediction Module (b) regresses a plausible UV position map from partial body cues. We can resample the predicted map to obtain a human mesh. Finally, we calculate the absolute position based on the regressed 3D joints and the visible 2D joints (d).

to other mesh-based methods [17], [18]. We adopt a variational auto-encoder (VAE) [19] to represent the human mesh prior, which is named UVPrior, to cover both pose and shape variations. To the best of our knowledge, UVPrior is the first one for learning-based priors in UV space. Given the visible information, UVPrior provides a reasonable hypothesis for the invisible part. Additionally, it prevents artifacts and penalizes implausible results, which are less-explored in [16], [20], [21].

Moreover, the estimated heatmaps and the 3D joints regressed from the UV position map are further combined to estimate the absolute position of human meshes. As shown in the experiments section, our method outperforms existing multi-person and single-person methods in both absolute and relative pose estimation. To better evaluate the performance of the proposed method, we further build a new benchmark named 3D Multi-person Basketball (3DMPB). 3DMPB contains more than 10K images and accurate 3D annotations, where all images are captured from real basketball scenes with numerous human-human interactions and inter-person occlusions. Our dataset may serve as a new challenging benchmark for the multi-person mesh recovery task.

The main contributions of this work are summarized as follows.

- We propose a visible pose-mask module to obtain valid features from inter-person occlusion cases, avoiding the pixel-level ambiguities in the top-down manner.
- We propose the first VAE-based UV prior, which covers both pose and shape prior knowledge. It also can be used to prevent the artifacts and penalize implausible results.
- We present a novel framework for absolute multi-person mesh recovery which achieves the state-of-the-art. We further build a multi-person benchmark named 3DMPB, which contains 2D segmentation and SMPL parameters. The dataset, codes are publicly available on our website.

## II. Related Work

**Single-person 3D mesh recovery**. At the early stage, lacking 3D annotations is the biggest challenge for 3D mesh recovery from single RGB images. [22]–[24] adopt an optimization-based framework to recover the SMPL [8] parameters from known or estimated 2D joints. To achieve real-time implementation, [15], [25], [26] use 2D cues as supervision and infer 3D mesh parameters directly from image features. As 3D data gradually enriches, human mesh recovery has made significant progress in accuracy and generalization. Sun *et al*. [27] proposed an end-to-end framework for recovering 3D human mesh from single images via a skeleton-disentangled representation. There are some approaches to estimate 3D locations of the mesh vertices using neural networks [28]–[30]. Instead of predicting 3D vertices, Zhu *et al*. [31] predicted vertex offsets and added them to a template mesh. More recent works [16], [20], [21], [32], [33] turned the single-person 3D mesh recovery into an image-to-image translation problem by encoding appearance and geometry into a UV map. It is worth to note that all of these methods can be applied to multi-person cases by detecting and cropping each person with the help of existing pose detectors. However, they could not achieve ideal performance without taking inter-person occlusions into account, which is the key difference between the single-person and multi-person problem.

**Multi-person 3D mesh recovery**. It is a challenging problem to recover multi-person mesh from a single color image. Zanfir *et al*. [5] presented the first bottom-up trainable model for multi-person pose and shape estimation in monocular images. Body part scores parameterized by both 2d and 3d information are predicted and assembled into skeletons. Sun *et al*. [11] proposed another bottom-up representation, which encodes the SMPL parameters into a 2D map. The multiple human meshes can be resampled from the map predicted from the RGB image. In contrast to the bottom-up approaches, the top-down framework first detects all people and then handles each target person. MSC [6] optimizes the 3D shape of each person in the image using multiple scene constraints. Jiang *et al*. [7] also adopted this framework by using Faster-RCNN [34]. The RoI-aligned features are used to predict SMPL parameters. Different from previous methods using bounding boxes to differentiate persons, we are the first to use visible pose and

mask, which tackles the problem of pixel-level ambiguities.

**3D human prior**. The skeleton-based methods [35] integrate physical joint constraints into optimization objectives. All degrees of freedom of the skeleton are enforced to stay within their anatomical limits, but the method is still hard to prevent implausible poses with the soft constraints. Recently, HMR [15] limits the joint angles by using an adversarial network, but it increases representational redundancy. To obtain a more reasonable pose distribution, Bogo *et al*. [22] fitted a Gaussian mixture model to motion capture data from CMU [36] and used it during optimization. However, the optimization is unstable when choosing the closest model. There are also some methods using VAE to model pose prior in latent space [37], [38]. Zanfir *et al*. [39] promoted the VAE-based prior with the normalizing flow, which avoids balancing the reconstruction loss and the KL-divergence (Kullback-Leibler divergence) in traditional VAE training. However, the existing priors can only represent pose variation. Recently, a lot of mesh recovery methods can simultaneously represent the pose and shape [16], [20], [21]. Due to the lack of UV-based prior, the recovered meshes of these methods may have some artifacts. To this end, we propose a novel prior for UV representation, which provides pose and shape prior knowledge while also preventing artifacts. The proposed prior outperforms the discriminator-based prior in reconstructed mesh quality.

## III. Method

As illustrated in Fig.1, our method recovers multiple body meshes with absolute positions from a single color image. With a novel UV prior, the network takes the RGB image and predicted 2D pose as input and reconstructs the multiple people through the Visible Pose-Mask module and UV Prediction Module.

### A. 3D Human Representation

Choosing a proper representation is essential for the human mesh recovery problem. The UV map is widely used in rendering textures, which can also be extended to represent human body meshes (UV position map [20]). For the UV position map, it uses the correspondence between mesh vertices and the UV coordinates, and stores the XYZ location of mesh vertices in the RGB channels of a UV position map. The inner-triangle values are calculated by barycentric interpolation.

Several reasons account for the choice of this representation. Firstly, when parts of the human body are occluded, it is hard to represent occluded parts with joint rotation parameters. In contrast, the UV map could provide a dense correspondence between pixels and mesh vertices and represent the occluded vertex explicitly [16]. Secondly, regressing the SMPL parameters is a highly non-linear mapping and suffers from the non-unique problem (*i.e*., periodicity) [18], [40], while it can be well handled via UV representation. Moreover, non-parametric representation can also easily provide SMPL parameters by fitting meshes [16] and be applied to some scenarios where the body parts are missing (*e.g*., hair and clothes) [29].

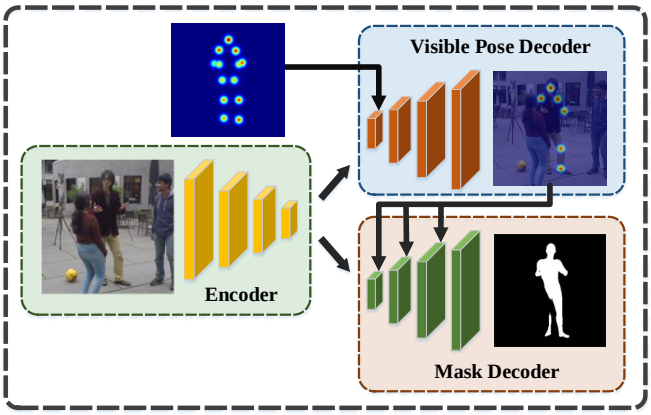



Fig. 2. **Visible Pose-Mask module.** We utilize the 2D pose as guidance for visible information extraction. The estimated full heatmaps are concatenated with latent features for visible heatmap refinement. Then, the mask of the target person is predicted from the estimated visible heatmaps and image features.

### B. Visible Pose-Mask Module

Due to various occlusions, reconstructing each individual from a single image is a challenging problem. Recent methods [6], [7] have difficulties to handle the inseparable cases in the image level (*e.g*., Fig.1 (a)). We first utilize the predicted 2D poses predicted by an off-the-shelf 2D pose detector [12] to locate the occluded person. Then a visible pose-mask module is designed to estimate visible heatmaps and mask to avoid the pixel-level ambiguities in a cascaded manner. Concatenating masks and heatmaps could provide visual attention for the network and ignore invalid features. Moreover, heatmaps restore articulation information, and masks provide important body shape cues [17], [25]. This information is crucial for human mesh recovery, especially for the UV position map, which combines both pose and shape. Besides, most of the existing 3D datasets are captured in constrained environments, causing a domain gap [17], [41], which is more evident in multi-person cases. Therefore, we introduce the visible heatmaps and mask to avoid pixel-level ambiguities, provide pose and shape guidance, and reduce the appearance domain gap.

It is difficult to obtain accurate segmentation and visible heatmaps from color images in inter-person occlusion cases. Inspired by [42], [43], we propose a visible pose-mask module to jointly predict visible heatmaps and mask. Specifically, we first detect all human poses in the given RGB image with an existing 2D pose detector [12]. Then, the proposed visible pose-mask module and the initial heatmaps of the detected pose are used to extract visible information for each target person. It is worth to note that other commonly used 2D pose detectors [13], [14] can also be applied to detect the initial poses. As shown in Fig.2, the visible pose-mask module contains two sub-nets: visible-pose net and mask-net. The encoder of the visible pose-mask module extracts features for the two decoders, and then the visible heatmaps are predicted from the image features and initial heatmaps. In the next step, predicted visible heatmaps are downsampled to the different scales and concatenated with the feature maps of the mask decoder to promote the visible human segmentation. The loss

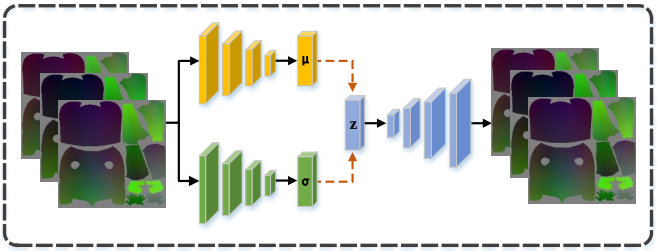



Fig. 3. **UVPrior** is trained in advance using traditional VAE structure. When training the UV Prediction Module, UVPrior encodes the estimated UV map and employs a regularization on the encoded latent code to penalize implausible results.

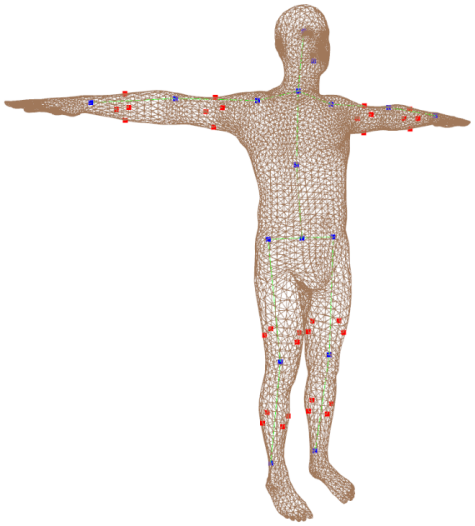


Fig. 4. **Visualization of anchor points (red points) and skeleton joints (blue points)**. The distance of anchor points illustrates the human shape. The distance of joints describes bone lengths.

function for the visible-pose net is denoted as:

$$L_{vp} = \sum_{i=1}^{N} \left\| H_{pre}^{i} - H_{gt}^{i} \right\|_2^2 + \lambda_o \sum_{i=1}^{N'} \left\| H_{o}^{i} \right\|_2^2, \tag{1}$$

where $N$ represents the number of joints, and $N'$ represents the number of occluded joints. The ground-truth visible heatmaps $H_{gt}$ are generated by checking whether the joint coordinate is within the mask region. This process could be simplified because the annotation provides the joint visibility, and we can directly check it. $H_o$ is the predicted heatmaps of occluded joints. We add the second term to erase these joints in visible heatmaps. The loss for the mask net is defined as:

$$L_{vm} = \left\| M_{pre} - M_{gt} \right\|_2^2. \tag{2}$$

The overall loss of the visible pose-mask module is

$$L_{pm} = \lambda_1 L_{vp} + \lambda_2 L_{vm}. \tag{3}$$

This cascaded module is trained in an end-to-end manner. The training data provides precise masks, 2D poses (not rendered masks), and the visibility of each joint. Due to the mutual promotion of mask and pose [43], our cascaded structure can obtain more accurate results than directly estimating the visible mask or visible pose from occluded images.

### C. UV Position Map Estimation

The UV prediction module encodes the image patch, estimated visible heatmaps and mask, and directly outputs the UV position map. Previous works [20], [21] only use the weighted L1 loss and total variation regularizer for supervision. Additionally, we propose a self-supervised symmetric loss that provides supervision on the bone length and body part shape using symmetric constraints. According to the human anatomy, the symmetric bone's length is consistent, e.g., right arm and left arm. As shown in Fig.4, we manually select four anchor points (red points) for each part of a limb to describe their shape. The distance between the two diagonal points depicts the width of each body part. We encourage the limb widths on the left and right to be consistent. Besides, the blue points are joints of the skeleton, which can be used to calculate the length of the bones. The benefit of this self-supervised loss is that it provides prior knowledge when some limbs are occluded. Symmetric shape loss is defined as:

$$L_{shape} = \sum_{i=1}^{N_l} \left| S - S_{sym} \right|, \tag{4}$$

where $S$ is the distance between two anchor points. Symmetric bone length loss is defined as:

$$L_{bone} = \sum_{i=1}^{N_l} \left| B - B_{sym} \right|, \tag{5}$$

where $N_l$ is the number of limbs. $B$ and $B_{sym}$ represent the bone lengths which are calculated by 3D joints.

The total loss is defined as:

$$L_{total} = \lambda_{L1} L_1 + \lambda_{tv} L_{tv} + \lambda_s L_{shape} + \lambda_b L_{bone} + \lambda_p L_{prior}, \tag{6}$$

where $L_{prior}$ is prior loss and will be discussed in next section. $L_1$ and $L_{tv}$ are the same as [16]:

$$L_1 = \sum_{j=1}^{H} \sum_{i=1}^{W} \beta_{i,j} \left( \left| P_{i,j} - P_{i,j}^{gt} \right| \right), \tag{7}$$

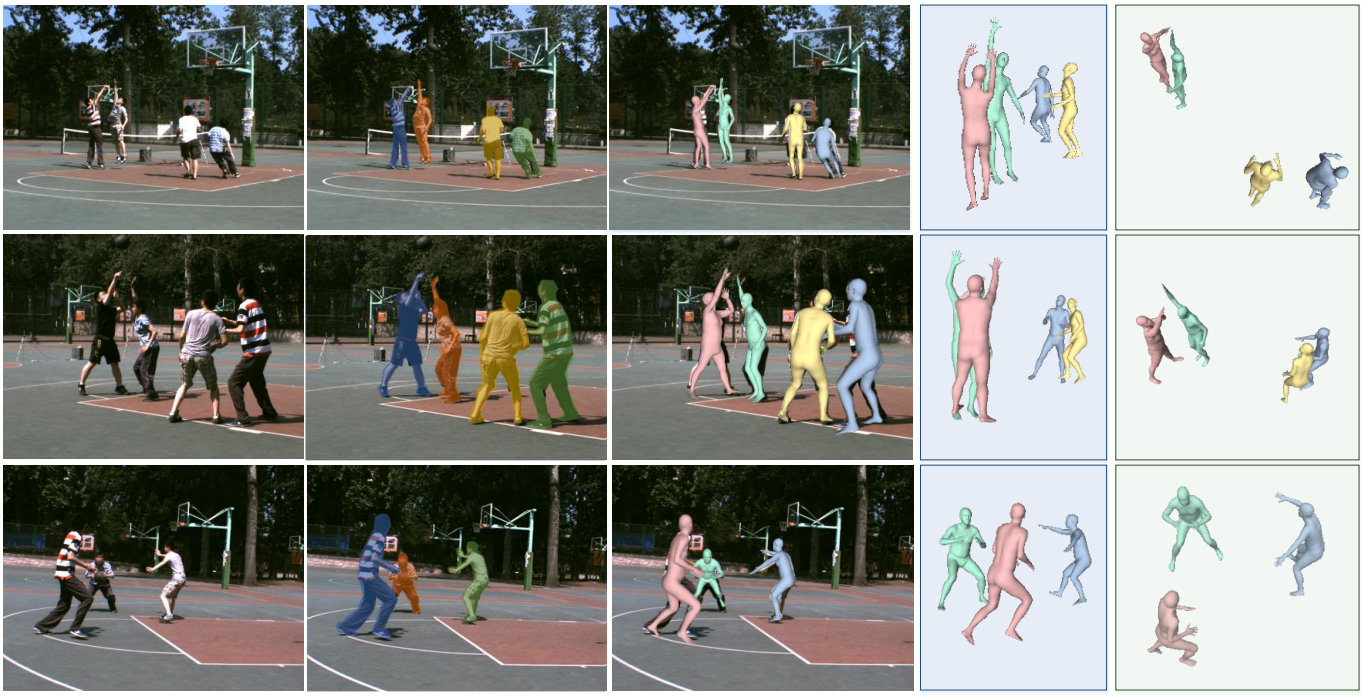
$\beta_{i,j}$ is a weight mask and the weight is inversely proportional to the part area. $P_{i,j}$ is the RGB value of pixel $(i, j)$ in the UV position map.

$$L_{tv} = \sum_{k} \sum_{(i,j) \in R_k} \left( \left| P_{i+1,j} - P_{i,j} \right| + \left| P_{i,j+1} - P_{i,j} \right| \right), \tag{8}$$

where $R_k$ is defined as the $k^{th}$ body part.

### D. UVPrior

Reconstructing full human meshes with limited visual cues will result in unreasonable pose and shape along with artifacts [16], [21]. As shown in Fig.3, we adopt a variational auto-encoder [19] structure with traditional reconstruction loss and KL-divergence loss to train a UV based prior, called UVPrior. Synthetic [4] and real data [44] provide massive feasible UV data for training. Unlike previous prior [37], which only considers pose information, UV prior not only contains the pose knowledge but also covers shape variations.



**(a) RGB image** **(b) Segmentation** **(c) Mesh** **(d) Different views**

Fig. 5. **Samples of the 3DMPB.** We provide multi-person images(a), segmentation(b), SMPL parameters(c), and camera parameters along with absolute position(d). 3DMPB may be a challenging benchmark for the problem of human mesh recovery.

In the training phase, UVPrior encodes each instance to a normal distribution and decodes the sampled code to reconstruct the input. We utilize two ResNet50 [45] as the backbone to regress mean and variance for a UV position map, respectively. The KL-divergence loss is applied to encourage a normal distribution. The normal distribution has several advantages in learning an expressive generative model (e.g., It is easily reparameterized and can be mapped to any other distribution). With the standard normal distribution constraint, the posterior distribution of the latent code is forced to be a normal distribution.

$$L_{KL} = KL(q(Z|R))||N(0, I)). \tag{9}$$

For the reconstruction loss, we find that the L1 loss between input and output can not reconstruct the input with high fidelity for map-based representation. We adopt the symmetric loss, weighted L1, and variation regularizer for training to obtain better results. The formulations are the same as Sec.III-C. The total loss for UVPrior is defined as:

$$L_{total} = \lambda_{L1}L_1 + \lambda_{tv}L_{tv} + \lambda_s L_{shape} + \lambda_b L_{bone} + \lambda_{kl}L_{KL} \tag{10}$$

where $\lambda_{L1}$ = 0.995, $\lambda_{tv}$ = 0.995, $\lambda_s$ = 1000, $\lambda_b$ = 1000, and $\lambda_{kl}$ = 0.005.

After training, the encoded latent code describes a manifold of human meshes, determining whether the input is physically plausible. When training the UV Prediction Module, we freeze the parameters of UVPrior. UVPrior encodes the estimated UV position map and employs a regularization on the encoded latent code to penalize implausible results. It is noted that we could use additional 2D data to train the UV prediction module with prior loss. The UV prior loss is defined as:

$$L_{prior} = \|z\|_2^2 . \tag{11}$$

The latent code z describes the plausibility of the UV position map.

Discriminator-based prior is widely used in parameter-based methods [15], [38]. However, the discriminator does not perform well compared to the VAE-based prior based in our experiments. Unlike the low-dimensional pose parameters, the UV position map contains complex body surface geometry. It is hard for the discriminator to judge the body mesh quality by using a one-dimension value. In contrast, our VAE model is trained with only real data and has a vertex-to-vertex correspondence resulting from the reconstruction loss.

### *E. Absolute Position Estimation*

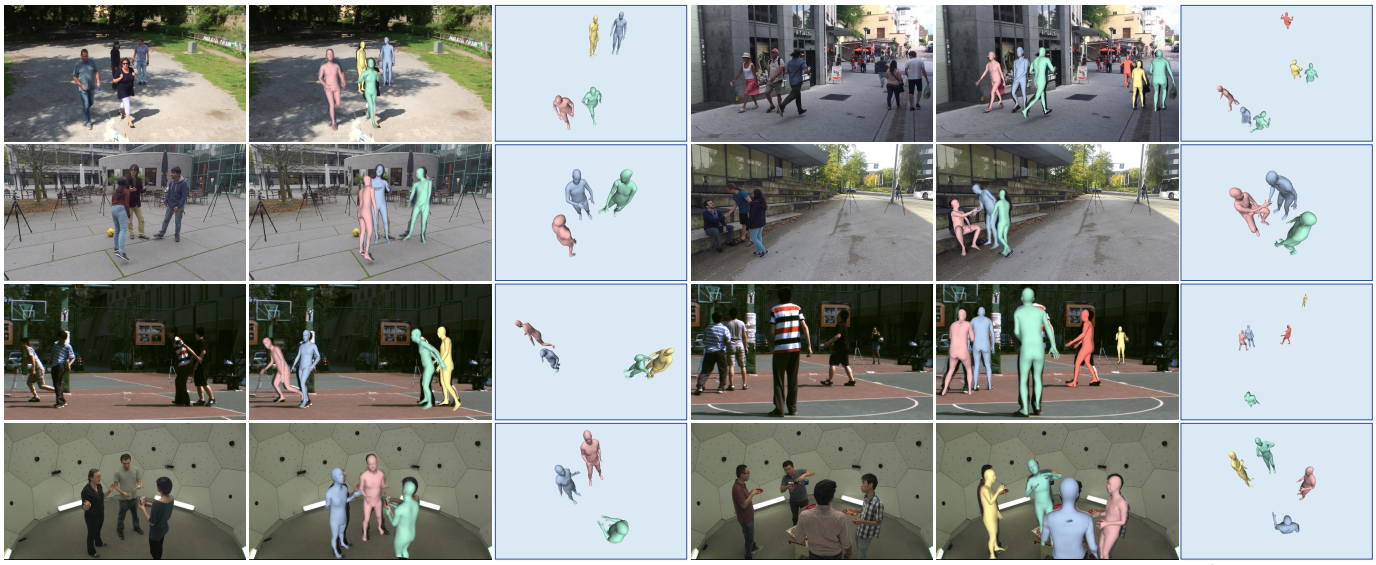
Due to the depth ambiguity, estimating people's absolute position from a monocular image is an ill-posed problem. To obtain an accurate absolute position, additional knowledge about the ground plane, reference objects, or absolute human height are necessary for traditional methods. Besides, visual perception of object scale and depth depends on the focal length and the image size. Therefore, learning-based methods are difficult to generalize on in-the-wild images. Fortunately, our UV representation is built in the metric system and could describe the real scale of humans. With camera parameters and estimated visible joints, the human mesh's translation can be estimated by matching 3D joints and 2D joints.

$$t^* = \arg\min_t \sum \| K(J_i + t) - P_i \|, \tag{12}$$

where $K$ is intrinsic matrix, $J$ denotes the 3D joints regressed from predicted UV-position map, $P$ is the estimated visible 2D joints.

TABLE I
COMPARISON AMONG DIFFERENT PUBLIC DATASETS RELATED TO MULTI-PERSON 3D POSE ESTIMATION. + DENOTES THE NUMBER OF OCCLUSION SAMPLES.

| Dataset | Occlusion | Outdoor | 2D Pose | 3D Pose | Segmentation | Mesh | Max Subj. |
|---|---|---|---|---|---|---|---|
| Campus and Shelf [46] | ++ | ✓ | ✓ | ✓ | – | – | 4 |
| MHHI [1] | ++ | – | ✓ | ✓ | ✓ | ✓ | 2 |
| CMU Panoptic [47] | ++ | – | ✓ | ✓ | – | – | 8 |
| 3DPW [48] | ++ | ✓ | ✓ | ✓ | – | ✓ | 2 |
| MuPoTS [10] | ++ | ✓ | ✓ | ✓ | – | – | 3 |
| **3DMPB** (ours) | +++ | ✓ | ✓ | ✓ | ✓ | ✓ | 4 |



**(a) Input image** **(b) Output mesh** **(c) Top down view** **(d) Input image** **(e) Output mesh** **(f) Top down view**

Fig. 6. **Qualitative results of our approach on 3DPW, MuPoTS, 3DMPB and Panoptic (a, d).** The output meshes and absolute positions are shown in (b, e) and (c, f).

## IV. 3DMPB DATASET

Recently, single-person 3D human datasets have seen great developments [49], [50]. However, it is still challenging to build a multi-person dataset under outdoor scenes due to inter-person occlusions and interactions, even for commercial systems. Pose diversity, background complexity, and subjects in the scene are the limitations of the existing datasets. A comparison with the existing multi-person datasets is shown in Tab.I. CMU Panoptic [47] is captured in a controlled environment and has a low background diversity. The only dataset with mesh, 3DPW [48], has a limited maximum number of people in each scene. To provide a more challenging benchmark for the problem of human mesh recovery, we propose **3DMPB**. 3DMPB is currently the only dataset with accurate human segmentations and 3D annotations. In addition, it provides a lot of human-human interactions and inter-person occlusions cases with challenging poses in real basketball scenes. Each frame has 1-4 subjects with challenging basketball poses and interactions. The set contains 6 subjects, 9 views, and 10K images. A few selected samples are shown in Fig.5. With accurate camera parameters, a multi-view fitting [51] is used to get the ground-truth SMPL parameters. Finally, the meshes with different identities and depths are rendered to obtain masks because most of the occlusions are inter-person occlusions.

## V. EXPERIMENTS

In this section, the quantitative and qualitative evaluations of our method are presented. We compare the proposed approach with state-of-the-art works on Human3.6M, CMU-Panoptic, 3DPW, MuPoTS, and our 3DMPB dataset. Furthermore, several ablation studies are conducted to illustrate the performance of our framework.

### A. Datasets

**Human3.6M** [49] is one of the most widely used single-person pose datasets. We use the subjects S1, S5, S6, S7, and S8 for training and keep the subjects S9 and S11 for testing. MoSH [55] is used to process the marker data in the original dataset to get SMPL parameters. We report the PA-MPJPE on this dataset.

**3DOH50K** [16] is a 3D human dataset with object occlusions. It provides accurate 3D annotations under various object-occlusions. We use the training set with 50310 images to train the Pose-Mask module and the UV prediction module.

**COCO** [56], **MPII** [57], **LSP** [58], **LSP Extended** [59], are in-the-wild 2D pose datasets. COCO dataset is used to train Pose-Mask module. EFT [60] provides SMPL parameters for these datasets.

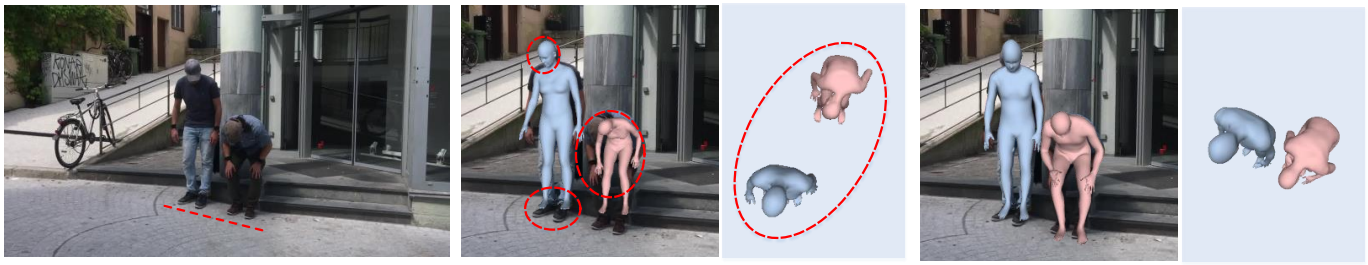
**CMU-Panoptic** [47] is a dataset with multiple people captured in a controlled scene. For evaluation, we follow the protocol of [6] to choose the testing set from four activities (Haggling, Mafia, Ultimatum, Pizza).

**(a) Input image** **(b) Sun et al.** **(c) Jiang et al.** **(d) Ours**

Fig. 7. Qualitative comparisons with Sun et al. [11] and Jiang et al. [7]. Even though our UV position map representation has few artifacts, our method still achieves better overall performance and is more close to the input image. Visible heatmaps and masks provide important guidance and reduce the distractions of pixel-level ambiguities. Moreover, benefiting from the UVPrior, our results are quite reasonable with heavy occlusions.

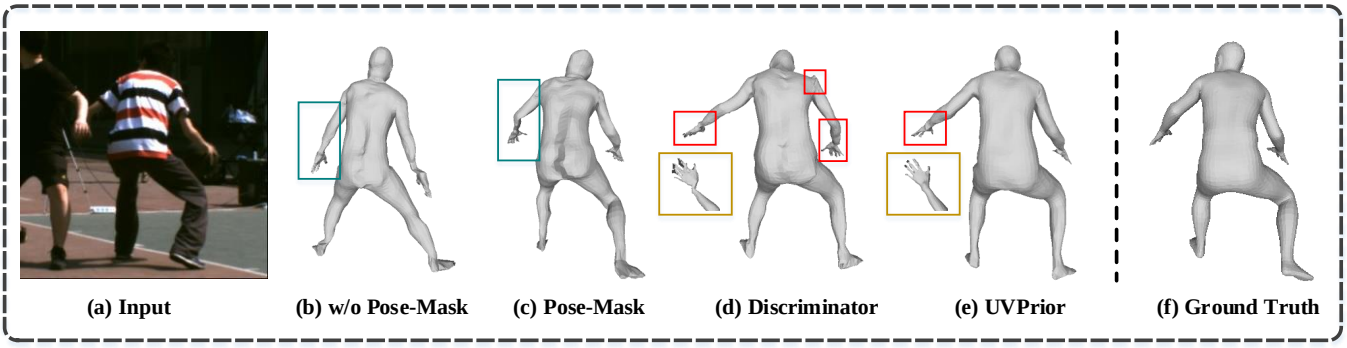
TABLE II
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON 3DPW, MUPOTS AND OUR 3DMPB.

| | Method | *3DPW* PA-MPJPE ↓ | *3DMPB* PA-MPJPE ↓ | *MuPoTS*$_{matched}$ PCK$_{rel}$ ↑ | PCK$_{abs}$ ↑ | *MuPoTS*$_{all}$ PCK$_{rel}$ ↑ | PCK$_{abs}$ ↑ |
|---|---|---|---|---|---|---|---|
| Pose only | Zhen et al. [52] | 92.3 | 113.8 | 80.5 | 38.7 | 73.5 | 35.4 |
| | Moon et al. [53] | 79.2 | 75.4 | 82.5 | 31.8 | 81.8 | 31.5 |
| | Li et al. [54] | 57.4 | – | – | – | **82.0** | **43.8** |
| Pose and shape | Jiang et al. [7] | 67.7 | 102.6 | 72.2 | – | 69.1 | – |
| | Sun et al. [11] | 57.4 | 72.0 | 73.3 | – | 70.7 | – |
| | Choi et al. [17] (predicted 2D pose) | 58.9 | 89.2 | – | – | – | – |
| | **Ours** (predicted 2D pose) | **57.1** | **69.5** | **82.8** | **39.5** | 78.0 | 37.6 |



**(a) Input** **(b) Jiang et al.** **(c) Ours**

Fig. 8. Our method obtains more accurate absolute position and body pose compared to Jiang et al. [7]. In our method, absolute depth is independent of the size of meshes which avoids the coupling of camera parameters and body size. More discussion is provided in Sec.V



**(a) Input** **(b) w/o Pose-Mask** **(c) Pose-Mask** **(d) Discriminator** **(e) UVPrior** **(f) Ground Truth**

Fig. 9. **Ablation on different modules.** Our Pose-Mask module removes ambiguous image features and promotes accurate mesh recovery (c). Compared with the discriminator (d), the UVPrior-supervised network generates preferable local details.

**MuPoTS-3D** [10] is obtained with a multi-view markerless motion capture system. For evaluation, a 3D percentage of correct keypoints (3DPCK$_{rel}$) with a threshold of $15cm$ is used after root alignment with ground-truth. Besides, we use 3DPCK$_{abs}$, which is the 3DPCK without root alignment, to evaluate the absolute camera-centered coordinates.

**3DPW** [48] is a 3D multi-person dataset and contains both indoor and outdoor scenes. For a fair comparison, none of the approaches get trained on 3DPW. We achieve the state-of-the-art on this impactive dataset.

**3DMPB** are captured from real basketball scenes. Due to numerous interactions and occlusions, it is a challenging benchmark for multi-person mesh recovery and pose estimation tasks. See Sec.IV for details.

### *B. Implementation Details*

We have three modules in total: visible pose-mask module, UV prediction module, and UVPrior module. UVPrior is trained at first, and the other two modules are trained separately because different parts use different training data. Existing datasets are not able to provide full supervision for the whole framework in an end-to-end manner. Input images are resized to 256*256. The detailed structure of the pose-mask module is provided in the supplementary material, and the training data is from COCO [56] and 3DOH50K [16] with precisely labeled masks and 2D poses. The backbone of the UV prediction module is ResNet50. The human mesh is normalized using a $2m$ bounding-box to obtain a unified UV position map. For UVPrior training, in addition to the AMASS dataset [44], we synthesize 0.5M data using the same method as [4]. We adopt Adam optimizer [61] and Leaky-ReLu [62] for training. All the networks are trained on a single NVIDIA TITAN RTX GPU with a learning rate of 1e-4 and a batch size of 32. Following the previous work [7], we use weak-perspective projection for a fair comparison. We use the LSP joint regressor for evaluation. In order to speed up our method, we apply multiprocessing in the pose detecting stage, preprocessing stage, and absolute position computing stage. The total running time reduces from 2000ms to 800ms.

### *C. Comparison with the state-of-the-art*

To demonstrate the effectiveness of our method, we performed quantitative evaluations on multiple commonly used multiple and single person datasets. Since only a small amount of work focuses on multi-person mesh recovery, we also compared the proposed approach with some single-person

TABLE III
**RESULTS ON HUMAN3.6M (PROTOCOL 2).** OUR METHOD IS COMPARABLE WITH SINGLE-PERSON METHODS.

| | Method | PA-MPJPE ↓ |
|---|---|---|
| Single-person | Kanazawa et al. [15] | 56.8 |
| | Kolotouros et al. [29] | 50.1 |
| | Zhang et al. [16] | 41.7 |
| | Kocabas et al. [38] | 41.4 |
| | Kolotouros et al. [26] | **41.1** |
| | Choi et al.(predicted 2D pose) [17] | 46.3 |
| | Choi et al.(accurate 2D pose) [17] | **38.4** |
| Multi-person | Jiang et al. [7] | 52.7 |
| | Sun et al. [11] | 50.1 |
| | **Ours** (predicted 2D pose) | 41.6 |
| | **Ours** (accurate 2D pose) | **36.0** |

TABLE IV
COMPARISON WITH MULTI-PERSON MESH RECOVERY METHODS. WE REPORT **MPJPE** ↓ ON THE PANOPTIC DATASET.

| Method | Haggling | Mafia | Ultim. | Pizza | Average |
|---|---|---|---|---|---|
| Sun et al. [11] | 151.5 | 168.7 | 182.8 | 202.1 | 176.3 |
| Zanfir et al. [6] | 140.0 | 165.9 | 150.7 | 156.0 | 153.4 |
| Zanfir et al. [5] | 141.4 | 152.3 | 145.0 | 162.5 | 150.3 |
| Jiang et al. [7] | 129.6 | **133.5** | 153.0 | 156.7 | 143.2 |
| **Ours** (predicted 2D pose) | **104.2** | 136.0 | **123.2** | **151.0** | **128.6** |

pose estimation baselines. Detailed comparisons are described in the following.

We first evaluated the performance of our approach on Human3.6M. As shown in Tab.III, our method is comparable with the state-of-the-art single-person baseline in terms of PA-MPJPE. Specifically, [16] also uses a UV position map to represent a human body. Our method generates more accurate results with the guidance provided by the heatmaps and mask. To verify the performance for the inter-human occluded cases, we evaluated our method on Panoptic by MPJPE metric. The images of the Panoptic dataset are severely affected by inter-person occlusion. In Fig.6 (row 4) and Tab.IV, the results demonstrate that our method effectively reduces the interference of occlusions and outperforms previous methods by 14.6 in terms of MPJPE. More qualitative results on single-person and occluded cases are shown in Fig.10.

3DPW and 3DMPB are outdoor multi-person datasets, which contain more complex occlusions and backgrounds. Since some previous works did not report the results on these benchmarks, we obtained the results by retesting the pre-trained models. For a fair comparison, we did not use any training data from these two datasets. In Fig.7, the qualitative comparison with [7] and [11] are shown. [11] regress a scale parameter to determine the depth relationship, and we found it is difficult to estimate coherent results when the depth of two people is similar. Since [7] first regresses bounding-boxes in the image, the results may be redundant in crowd cases. We also found that a wrong depth will cause a severe deviation of human shape, see Fig.8. However, absolute human depth is independent of mesh recovery in our method, so the shape prediction will not be affected by the depth estimation. The results in Tab.II, Fig.11 and Fig.6 demonstrate that our method performs well in various multi-person scenarios.

TABLE V
**COMPARISON OF COMPUTATIONAL COSTS.**

| | Method | Param ↓ |
|---|---|---|
| SMPL-based | Kanazawa et al. [15] | **26.8M** |
| | Kolotouros et al. [26] | 27.0M |
| | Sun et al. [11] | 29.0M |
| Mesh-based | Zhang et al. [40] | 102.3M |
| | Choi et al. [17] | 77.1M |
| | Kolotouros et al. [29] | 42.7M |
| | Zhang et al. [16] | **30.0M** |
| | **Ours** | 58.2M |

TABLE VI
**ABLATION ON 3DMPB.** TO FULLY TEST THE PERFORMANCE OF EACH PART WITH SUFFICIENT OCCLUSIONS, WE CONDUCTED THE ABLATION STUDY ON 3DMPB. THE DATASET IS SPLIT INTO TRAINING SET AND TESTING SET FOR TRAINING AND TESTING.

| Method | MPJPE ↓ | PA-MPJPE ↓ | MPVPE ↓ |
|---|---|---|---|
| baseline(RGB to UV) | 93.2 | 66.4 | 123.4 |
| +heatmap | 76.2 | 52.8 | 111.9 |
| +heatmap+Pose-Mask(SMPL) | 193.2 | 59.3 | 208.2 |
| +heatmap+Pose-Mask(UV) | 65.5 | 47.6 | 102.5 |
| +heatmap+Pose-Mask+$L_{sym}$ | 65.4 | 47.4 | 100.3 |
| +heatmap+Pose-Mask+disc. | 65.1 | 47.1 | 95.2 |
| **+heatmap+Pose-Mask+UVPrior** | **62.9** | **45.6** | **91.9** |
| **+heatmap$_{gt}$+mask$_{gt}$+UVPrior** | **55.0** | **38.1** | **75.8** |

We found that Human3.6M only has few occluded cases, and our results are close to other methods. Some deviations may come from the calculation of joint position because we directly predict full mesh without the strong body prior of SMPL parameters. However, the gap between other baselines and our method increases coinciding with the increased occlusion threshold, *aka* Panoptic $>$ 3DMPB $>$ 3DPW $>$ Human3.6M. It turns out that our method is more robust to occlusions and outperforms existing mesh-based methods.

We also conducted experiments on MuPoTS to evaluate the performance of absolute depth estimation. [7] is a state-of-the-art method for absolute mesh estimation. Since depth ordering-aware loss is used as supervision, the estimated results can only reflect the ordinal depth relation. A qualitative comparison in Fig.8 shows that an incorrect shape may seriously affect the depth estimation in [7]. To quantify the effectiveness of our method, we also compared our approach with other absolute pose estimation methods. As shown in Tab.II, our method surpasses the state-of-the-art methods.

Moreover, we compare the computational costs with SMPL-based and Mesh-based methods in Tab.V. Compared to SMPL-based methods, our UV-based method has more computational complexity because of the high-dimension output. To handle the challenging multi-person problem, there is a trade-off of performance and complexity. On the other hand, one of the advantages of UV representation is that we can use pure convolution operation, which highly reduces the complexity compared to other mesh-based outputs with graph convolution and many linear layers.

### *D. Ablation Study*

**UVPrior**. We conducted several experiments to prove the importance of our proposed UVPrior. The UVPrior is a 3D human prior, which provides additional information to recover complete mesh from limited cues. In Tab.VI (row 4 and row 7), we found all metrics are decreased, and the vertex error

Fig. 10. More qualitative results on single-person and object-occluded cases. Our proposed framework is also compatible with other cases like self-occlusions and object-occlusions and achieves competitive performance with single-person methods.

TABLE VII
VISIBLE MASK ABLATION ON COCO2017VAL.

| Method | AP ↑ | $AP_M$ ↑ | $AP_L$ ↑ |
|---|---|---|---|
| Mask R-CNN [63] | 0.532 | 0.433 | 0.648 |
| **Ours** | **0.554** | **0.544** | 0.641 |

TABLE VIII
2D VISIBLE POSE ABLATION ON COCO2017VAL.

| Method | MPJPE-2D$_{0.2}$ ↓ | MPJPE-2D$_{0.3}$ ↓ | MPJPE-2D$_{0.4}$ ↓ | MPJPE-2D$_{0.5}$ ↓ |
|---|---|---|---|---|
| AlphaPose [12] | 17.17 | 17.39 | 17.55 | 17.62 |
| **Ours** | **6.00** | **6.00** | **5.60** | **5.29** |

is greatly reduced by 10.6. Accordingly, Fig.9 (e) shows the mesh quality has been significantly improved. In addition, we compared our UVPrior with discriminator-based prior.

For the discriminator, we directly adopt the structure of 16*16 PatchGAN [64], and the adversarial loss is:

$$L_{adv} = \mathbb{E}_{P\sim G}\left[\left(\mathcal{D}(\hat{\boldsymbol{P}}) - 1\right)^2\right] \tag{13}$$

The objective for discriminator is:

$$L_{\mathcal{D}} = \mathbb{E}_{P\sim R}\left[\left(\mathcal{D}(\boldsymbol{P}) - 1\right)^2\right] + \mathbb{E}_{P\sim G}\left[\mathcal{D}(\hat{\boldsymbol{P}})^2\right] \tag{14}$$

Where $\hat{P}$ is the output UV position map of the generator, and $P$ is the real data from AMASS dataset. Tab.VI (row 6 and row 7) and Fig.9 (d, e) show that UVPrior is a better choice to provide body prior knowledge in terms of UV representation. We also tried a simple structure that directly discriminates the whole UV position map. The results are worse than the PatchGan discriminator which still proves that the VAE-based UVPrior is more suitable.

**Visible Pose-Mask Module**. We then conducted experiments to demonstrate the effectiveness of the proposed Pose-Mask module on tackling pixel-level ambiguities. We also evaluate this module to verify that it outperforms state-of-the-art human segmentation and pose estimation works on multi-person cases.

Tab.VI (row 2) shows that introducing heatmaps to human mesh recovery can bring a significant improvement. It verifies

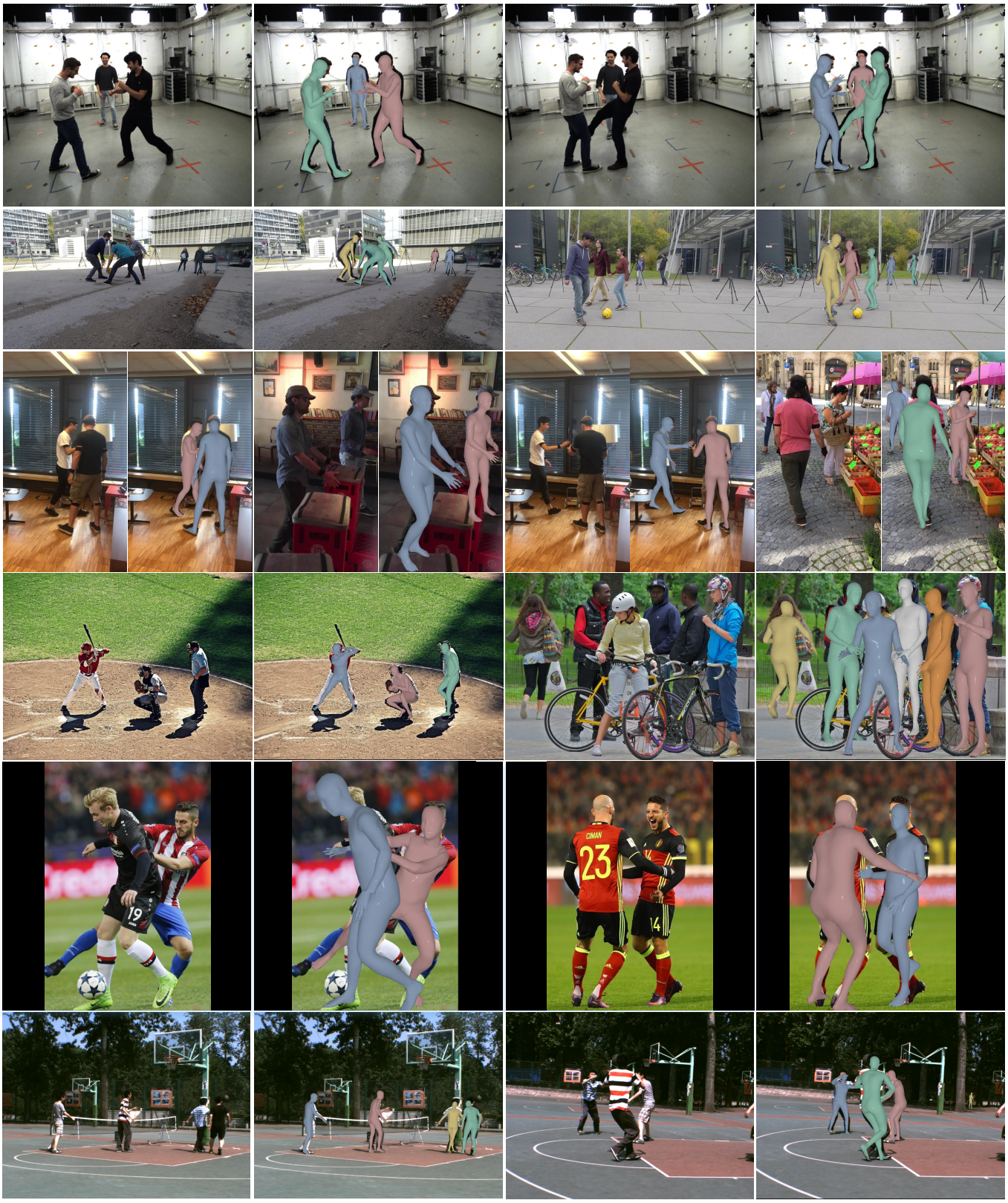


Fig. 11. Qualitative results on multi-person testing set and internet images. Video results are provided in the supplementary material. Our method can obtain promising results with the in-the-wild color images as input.

**(a) Input** **(b) Mask RCNN** **(c) Visible Pose-Mask**

Fig. 12. Qualitative comparison with Mask RCNN [63]. Our visible pose-mask module separate each person correctly from a human-human occluded image.

the necessity of pose information. Qualitative and quantitative results in Fig.9 (b, c) and Tab.VI (row 1 and row 4) show misleading predictions, and pixel-level ambiguities are removed by incorporating this module.

Besides the above analysis of overall performance, we also conduct experiments on this module isolatedly. We first compared with Mask-RCNN [63] with the latest trained model provided by Detectron2 [65] to demonstrate the effectiveness of our module on visible human segmentation. With the help of visible poses, our method outperforms the Mask-RCNN according to Tab.VII, especially for the $AP_m$. Fig.12 shows a qualitative comparison with Mask R-CNN [63]. The results show that it is hard for the state-of-the-art segmentation framework to separate each instance from inter-person occlusion images. The inter-person occlusion delivers severe pixel-level ambiguity to network regression. However, due to the pose guidance, the visible pose-mask module performs well in these cases.

For visible pose estimation, we use different confidence thresholds and compare our method with the state-of-the-art [12]. Both the accuracy of visibility and joint position precision for multi-person have a significant improvement. The results are shown in Tab.VIII.

**UV Representation**. To verify the superiority of UV representation, we compared it with the SMPL parameters. The setting for the SMPL regression task is the same as the UV except for the last layer (using linear layer to regress parameters). The gap between UV and SMPL representation comes from occlusions. The parameters are unconstrained for the occluded joints because they can not represent explicit occlusions and do not have a dense correspondence which results in the large MPJPE arising from wrong root orientation (Tab.VI (row 3 and row 4)).

## VI. Discussion

Our framework allows us to recover multiple human meshes with absolute position from a single color image. However, it is difficult for neural networks to prevent interpenetration even it has been considered in loss function [7]. As shown in Fig.13 (row 4), the interpenetration also occurs in close physical contact cases. We can solve this by adopting an additional optimization with a collision loss, which is not our main concern. Another type of failure case results from the off-the-shelf 2D detector (Fig.13). When the target is not detected, our method could not provide the corresponding output mesh. The improvement of multi-person 2D pose estimation will promote the performance of our method. In the future, we hope to reconstruct closely interacting people without interpenetration and provide more smooth and physically plausible results.

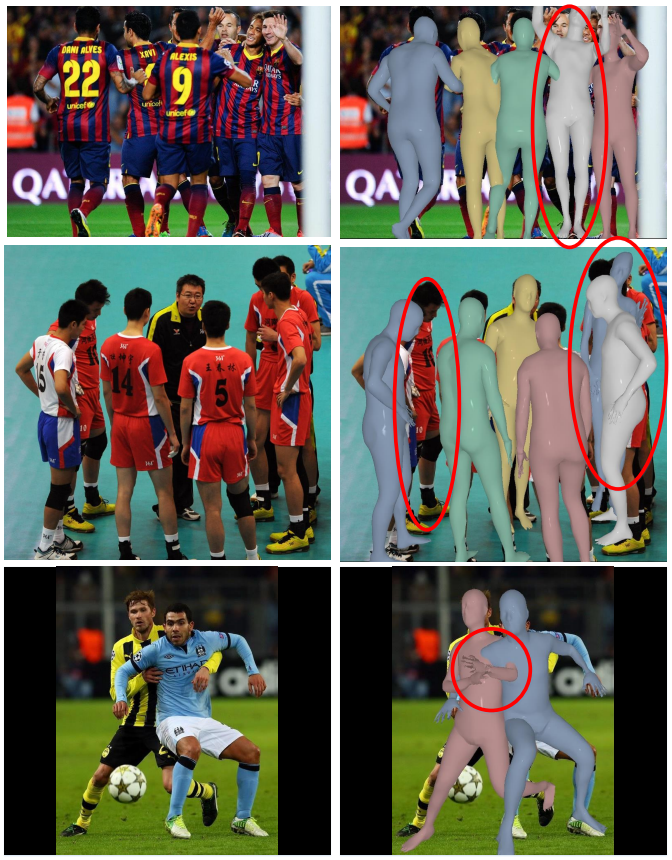


Fig. 13. Failure cases. When no visible 2D keypoints are detected, the human cannot be reconstructed (row 1, 2, 3). Interpenetration cannot be avoided for the results directly output from the neural network (row 4).

## VII. Conclusion

In this paper, we propose a novel framework for single-shot multi-person mesh recovery. To deal with the challenging

pixel-level ambiguities, we utilize the visible pose and mask to distinguish the target person and propose a novel UV-based prior to reconstruct plausible meshes with limited information. More concretely, we introduce a pose-mask mutual promotion mechanism for target separation and design a UVPrior to provide prior knowledge for the body mesh. We also demonstrate that the VAE-based prior is more competitive than the discriminator for UV representation since it can describe more local details. Our method is evaluated in various benchmarks and outperforms strong baseline methods. To thoroughly evaluate inter-person occlusion cases, we further build a 3D in-the-wild multi-person benchmark, 3DMPB. We believe this benchmark will promote future research on multi-person mesh recovery and other topics.

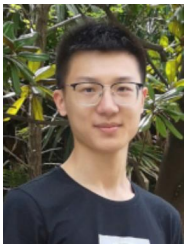
## ACKNOWLEDGMENT

We thank the anonymous reviewers to improve this paper. The authors also would like to thank all participants who have contributed to the 3DMPB dataset for their precious time and efforts.

## REFERENCES

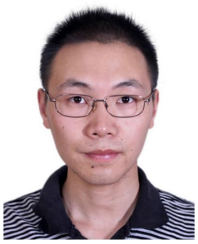
[1] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt, "Markerless motion capture of interacting characters using multi-view image segmentation," in *CVPR*, 2011, pp. 1249–1256.
[2] K. Li, N. Jiao, Y. Liu, Y. Wang, and J. Yang, "Shape and pose estimation for closely interacting persons using multi-view images," in *Computer Graphics Forum*, vol. 37, no. 7. Wiley Online Library, 2018, pp. 361–371.
[3] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *CVPR*, 2018, pp. 8320–8329.
[4] J. Liu, N. Akhtar, and A. Mian, "Temporally coherent full 3d mesh human pose recovery from monocular video," *arXiv preprint arXiv:1906.00161*, 2019.
[5] A. Zanfir, E. Marinoiu, M. Zanfir, A.-I. Popa, and C. Sminchisescu, "Deep network for the integrated 3d sensing of multiple people in natural images," in *NIPS*, 2018.
[6] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Monocular 3d pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints," in *CVPR*, 2018.
[7] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidis, "Coherent reconstruction of multiple humans from a single image," in *CVPR*, 2020.
[8] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.
[9] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *CVPR*, 2019, pp. 10 863–10 872.
[10] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, and C. Theobalt, "Single-shot multi-person 3d pose estimation from monocular rgb," in *3DV*, 2018.
[11] Y. Sun, Q. Bao, W. Liu, Y. Fu, and T. Mei, "Centerhmr: a bottom-up single-shot method for multi-person 3d mesh recovery from a single image," 2020.
[12] H. Fang, S. Xie, Y. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *ICCV*, 2017.
[13] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1.
[14] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.
[15] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *CVPR*, 2018.
[16] T. Zhang, B. Huang, and Y. Wang, "Object-occluded human shape and pose estimation from a single color image," in *CVPR*, June 2020.
[17] H. Choi, G. Moon, and K. M. Lee, "Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose," in *ECCV*, 2020.
[18] G. Moon and K. M. Lee, "I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image," *arXiv preprint arXiv:2008.03713*, 2020.
[19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.
[20] P. Yao, Z. Fang, F. Wu, Y. Feng, and J. Li, "Densebody: Directly regressing dense 3d human pose and shape from a single color image," 2019.
[21] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang, "3d human mesh regression with dense correspondence," in *CVPR*, June 2020.
[22] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *ECCV*, 2016.
[23] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *CVPR*, 2019.
[24] J. Song, X. Chen, and O. Hilliges, "Human body model fitting by learned gradient descent," 2020.
[25] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3d human pose and shape from a single color image," in *CVPR*, 2018, pp. 459–468.
[26] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," 2019.
[27] S. Yu, Y. Yun, L. Wu, G. Wenpeng, F. YiLi, and M. Tao, "Human mesh recovery from monocular images via a skeleton-disentangled representation," in *ICCV*, 2019.
[28] A. Venkat, C. Patel, Y. Agrawal, and A. Sharma, "Humanmeshnet: Polygonal mesh recovery of humans," 2019.
[29] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *CVPR*, June 2019.
[30] K. Lin, L. Wang, Y. Jin, Z. Liu, and M.-T. Sun, "Learning nonparametric human mesh reconstruction from a single image without ground truth meshes," 2020.
[31] L. Zhu, K. Rematas, B. Curless, S. Seitz, and I. Kemelmacher-Shlizerman, "Reconstructing nba players," in *ECCV*, 2020.
[32] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2shape: Detailed full human body geometry from a single image," in *ICCV*, 2019.
[33] V. Lazova, E. Insafutdinov, and G. Pons-Moll, "360-degree textures of people in clothing from a single image," in *3DV*, 2019.
[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.
[35] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, "Livecap: Real-time human performance capture from monocular video," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 2, pp. 1–17, 2019.
[36] http://mocap.cs.cmu.edu/.
[37] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *CVPR*, 2019.
[38] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *CVPR*, 2020.
[39] A. Zanfir, E. G. Bazavan, H. Xu, B. Freeman, R. Sukthankar, and C. Sminchisescu, "Weakly supervised 3d human pose and shape reconstruction with normalizing flows," *arXiv preprint arXiv:2003.10350*, 2020.
[40] H. Zhang, J. Cao, G. Lu, W. Ouyang, and Z. Sun, "Learning 3d human shape and pose from dense body parts," in *ACM MM*, 2020.
[41] H. Tu, C. Wang, and W. Zeng, "Voxelpose: Towards multi-camera 3d human pose estimation in wild environment," in *ECCV*, 2020.
[42] S. H. Zhang, R. Li, X. Dong, P. Rosin, and S. M. Hu, "Pose2seg: Detection free human instance segmentation," in *CVPR*, 2019.
[43] Y. Wang, C. Peng, and Y. Liu, "Mask-pose cascaded cnn for 2d hand pose estimation from single color image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2018.
[44] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *ICCV*, 2019.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[46] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3d pictorial structures for multiple human pose estimation," in *CVPR*, 2014.

[47] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social interaction capture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 190–204, 2019.
[48] T. V. Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *ECCV*, 2018.
[49] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, jul 2014.
[50] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," 2017.
[51] B. Huang, Y. Shu, T. Zhang, and Y. Wang, "Dynamic multi-person mesh recovery from uncalibrated multi-view cameras," in *3DV*, 2021.
[52] J. Zhen, Q. Fang, J. Sun, W. Liu, W. Jiang, H. Bao, and X. Zhou, "Smap: Single-shot multi-person absolute 3d pose estimation," in *ECCV*, 2020.
[53] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image," in *ICCV*, 2019.
[54] J. Li, C. Wang, W. Liu, C. Qian, and C. Lu, "Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation," in *ECCV*, 2020.
[55] M. M. Loper, N. Mahmood, and M. J. Black, "MoSh: Motion and shape capture from sparse markers," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 33, no. 6, pp. 220:1–220:13, Nov. 2014. [Online]. Available: http://doi.acm.org/10.1145/2661229.2661273
[56] T. Y. Lin, M. Maire, S. Belongie, J. Hays, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[57] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, June 2014.
[58] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *BMVC*, 2010.
[59] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *CVPR*, 2011, pp. 1465–1472.
[60] H. Joo, N. Neverova, and A. Vedaldi, "Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation," *arXiv preprint arXiv:2004.03686*, 2020.
[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[62] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML*, 2013.
[63] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2961–2969.
[64] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networkss," in *ICCV*, 2017.
[65] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.



**Buzhen Huang** received the bachelor's degree in Automation from Hangzhou Dianzi University, Hangzhou, China, in 2019. He is currently pursuing a Ph.D. degree at Southeast University, Nanjing, China. His research interests include computer vision and computer graphics.

**Tianshu Zhang** is currently a third year master student at Southeast University, Nanjing, China. He received his honors degree from Chien-Shiung Wu College of Southeast University in 2019. His research interests include computer vision, computer graphics and human mesh recovery.



**Yangang Wang** received his B.E. degree from Southeast University, Nanjing, China, in 2009 and his Ph.D. degree in control theory and technology from Tsinghua University, Beijing, China, in 2014. He was an associate researcher at Microsoft Research Asia from 2014 to 2017. He is currently an associate professor at Southeast University. His research interests include image processing, computer vision, computer graphics, motion capture and animation.