Delving Deep into Pixel Alignment Feature for Accurate Multi-View Human Mesh Recovery

Kai Jia, Hongwen Zhang*, Liang An, Yebin Liu*

Department of Automation, Tsinghua University, Beijing, China
kajia@umich.edu, zhanghongwen@mail.tsinghua.edu.cn, al17@mails.tsinghua.edu.cn, liuyebin@mail.tsinghua.edu.cn

Abstract

Regression-based methods have shown high efficiency and effectiveness for multi-view human mesh recovery. The key components of a typical regressor lie in the feature extraction of input views and the fusion of multi-view features. In this paper, we present Pixel-aligned Feedback Fusion (PaFF) for accurate yet efficient human mesh recovery from multi-view images. PaFF is an iterative regression framework that performs feature extraction and fusion alternately. At each iteration, PaFF extracts pixel-aligned feedback features from each input view according to the reprojection of the current estimation and fuses them together with respect to each vertex of the downsampled mesh. In this way, our regressor can not only perceive the misalignment status of each view from the feedback features but also correct the mesh parameters more effectively based on the feature fusion on mesh vertices. Additionally, our regressor disentangles the global orientation and translation of the body mesh from the estimation of mesh parameters such that the camera parameters of input views can be better utilized in the regression process. The efficacy of our method is validated in the Human3.6M dataset via comprehensive ablation experiments, where PaFF achieves 33.02 MPJPE and brings significant improvements over the previous best solutions by more than 29%. The project page with code and video results can be found at https://kairobo.github.io/PaFF/.

Introduction

Accurate and efficient recovery of the body mesh underlining a target human is undoubtedly helpful for substream tasks such as behavior understanding (Petrovich, Black, and Varol 2022) and human digitalization (Zheng et al. 2021), etc. With the employment of neural networks, regression-based methods (Kanazawa et al. 2018; Kolotouros, Pavlakos, and Daniilidis 2019; Guler and Kokkinos 2019; Lin, Wang, and Liu 2021) have shown promising results towards this goal. However, regression-based methods typically suffer from coarse alignment between the estimation and the person images. By analogizing the optimization methods (Bogo et al. 2016; Zhang et al. 2020b; Li, Oskarsson, and Heyden 2021), recent state-of-the-art approaches to monocular human mesh recovery (Song, Chen, and Hilliges 2020; Zanfir et al. 2021; Zhang et al. 2021a) make attempts to predict a neural descent to estimate parameters iteratively from feedback signals such as keypoint and part alignment errors (Song, Chen, and Hilliges 2020; Zanfir et al. 2021), or re-project the estimation to the original feature space to get pixel-level feedback features (Zhang et al. 2021a). These two trends of methods all utilize feedback features to update the estimation iteratively, which achieves better alignment quality than the previous regression-based methods. However, the accuracy of these monocular approaches remains far from satisfactory due to the underdetermined observations from a single image.

When deploying regression-based methods on multi-view setups, it is crucial to fuse multi-view information so that complementary observations can be considered in the regression process. Existing multi-view regression-based methods have proposed several fusion strategies, including view-by-view (Liang and Lin 2019; Yao et al. 2019), volumetric (Iskakov et al. 2019; Shin and Halilaj 2020), graph or transformer (Wu et al. 2021; Zhang et al. 2021b; Yagubbayli, Tonioni, and Tombari 2021; Shuai, Wu, and Liu 2022; He et al. 2020). View-by-view fusion methods (Liang and Lin 2019; Yao et al. 2019) perform estimation view by view and pass the estimation to the next view or stage, which gives improvement from their initial estimations. However, these methods do not consider all of the camera parameters in each stage, which can lead to a large multi-view misalignment. Volumetric fusion methods (Iskakov et al. 2019; Shin and Halilaj 2020) utilize back-projection operation to construct a feature volume and then use 3D convolution to fuse the spatial features. Nevertheless, these methods would introduce quantization errors in the discretization of the volume space. Moreover, noisy camera parameters and self-occlusion might lead to the situation that a target voxel in the 3D space corresponds to different body positions in 2D images, making the fusion ambiguous and less effective. Graph or transformer fusion methods (Wu et al. 2021; Zhang et al. 2021b) capture correlation between features from different views to search or infer the best fused feature for the final estimation. But these features are too sparse for the estimation of body mesh such as SMPL (Loper et al. 2015). In this work, we propose to fuse features on mesh vertices and show that such a vertex-wise fusion strategy is more suitable for multi-view human mesh recovery.

*Corresponding Authors

Based on the above motivations, we propose Pixel-aligned Feedback Fusion (PaFF) for multi-view human mesh recovery. PaFF uses the pixel-aligned feedback features in the feature extraction phase and regresses the mesh parameters using the features fused on mesh vertices. As illustrated in Fig. 1(a), the pixel-aligned features are extracted based on the feedback re-projection of the current estimation and reflect the alignment status on each input view. Fusing the alignment status on mesh vertices could provide more explicit guidance for our regressor to update the mesh parameters. When extracting the pixel-aligned features, an accurate estimation for orientation and translation is needed since it can align the estimation close to the body region in multi-view images and help us extract more informative feedback features for the pose and shape estimation. To this end, we disentangle the global orientation and translation from the pose and shape estimation and carefully design the orientation and translation estimators. Specifically, we incorporate camera parameters in the regression process to figure out the optimized orientation and translation that accord with each input view. Such a strategy enables our regressor to produce the global orientation and translation in an end-to-end manner and better handle the scale and rotation ambiguity issues. Different from the previous triangulation-based solution (Iskakov et al. 2019), our method does not depend on keypoint detection results and thus is free from detection mistakes under challenging cases such as occlusions.

We conclude our contributions as follows:

- We propose Pixel-aligned Feedback Fusion (PaFF) for regression-based multi-view human mesh recovery. PaFF iteratively extracts pixel-aligned feedback features from each input view and fuses them on mesh vertices. The feedback features and vertex-wise fusion enables our regressor to update the parameters such that the body mesh is progressively aligned to each input view.
- We propose to disentangle the global orientation and translation from the estimation of mesh parameters since they are correlated to the camera parameters. In this way, the camera parameters can be better utilized in our regressor to overcome the scale and rotation ambiguity issues for a more accurate estimation of the global orientation and translation.
- Our method achieves state-of-the-art performances and brings significant improvements over previous methods in benchmark datasets with both calibration and calibration-free settings. PaFF provides an end-to-end solution for accurate, simple, yet efficient human mesh recovery from multi-view images.

Related Work

The recovery of human body mesh from RGB images has been actively studied in recent years (Bogo et al. 2016; Kanazawa et al. 2018; Kocabas, Athanasiou, and Black 2020; Caliskan et al. 2020; Zhang et al. 2021a; Sengupta, Budvytis, and Cipolla 2021; Kolotouros et al. 2021). Existing methods can be roughly divided into two paradigms, i.e., optimization-based methods (Liu et al. 2011; Xu et al. 2017; Bogo et al. 2016; Li, Oskarsson, and Heyden 2021; Zhang et al. 2021c; Huang et al. 2017; Zanfir et al. 2021; Ajanohoun, Paquette, and Vázquez 2021) and regression-based methods (Pavlakos et al. 2017; Omran et al. 2018; Kanazawa et al. 2018; Varol et al. 2018; Kolotouros et al. 2019; Rong et al. 2019; Rong, Shiratori, and Joo 2021; Guler and Kokkinos 2019; Kolotouros, Pavlakos, and Daniilidis 2019; Zhang et al. 2020a; Liang and Lin 2019; Sun et al. 2021; Zhang et al. 2021a; Lin, Wang, and Liu 2021; Xuan, Zhang, and Li 2022). We refer readers to (Tian et al. 2022) for a comprehensive survey in this field.

Regression-based methods for multi-view human mesh recovery (Liang and Lin 2019; Shin and Halilaj 2020; Sengupta, Budvytis, and Cipolla 2021; Zhang et al. 2021b) usually need to go through a feature extraction phase, a multi-view feature fusion phase, and an inference phase. To fuse the information from the multi-view, Liang and Lin (2019) proposes to estimate a human body stage by stage and view by view without using camera parameters. Shin and Halilaj (2020) fuses the multi-view features to a 3D feature volume using back-projection and then regresses the body parameters from the flattened volumetric feature, giving the state-of-art accuracy in the multi-view human body reconstruction, yet with quantization errors and possible occluded non-body regions incorporated.

While optimization-based methods (Bogo et al. 2016; Huang et al. 2017; Li, Oskarsson, and Heyden 2021; Ajanohoun, Paquette, and Vázquez 2021) can fit the estimation aligned to 2D evidence iteratively, regression-based methods (Kanazawa et al. 2018; Kolotouros, Pavlakos, and Daniilidis 2019) usually suffer from bad alignment to the human image region. Recently, Neural Descent methods mimicking optimization processes appeared in the human body estimation task (Carreira et al. 2016; Zanfir et al. 2021; Song, Chen, and Hilliges 2020; Zanfir et al. 2021; Corona et al. 2022). These methods use the iterative regression with feedback signals such as keypoint re-projection errors and body part alignment errors (Zanfir et al. 2021) to infer a neural descent of the parameters. Nevertheless, these methods can be sensitive to noisy 2D evidence. Recently, several methods extract the implicit feedback signals from the image features by projecting estimation to the image area (Zhang et al. 2021a,b), which contains richer alignment information. Zhang et al. (2021a) uses the mesh re-projection feedback feature while Zhang et al. (2021b) projects estimated 3D keypoints into each view and utilizes deformable convolution (Zhu et al. 2019) to extract efficient feedback features. Compared with the previous methods only relying on numerical errors (Song, Chen, and Hilliges 2020; Zanfir et al. 2021), these methods can utilize the rich image features to get a more aligned estimation and have more information in tasks such as shape estimation. It also relies less on the on-the-shelf models such as keypoints detection or part segmentation which might introduce additional noises. However, these methods are either not applied in the multi-view setting (Zhang et al. 2021a) or only used for the human 3D keypoints estimation task (Zhang et al. 2021b). Therefore, an efficient solution of the multi-view human reconstruction with the feedback loop is meaningful.

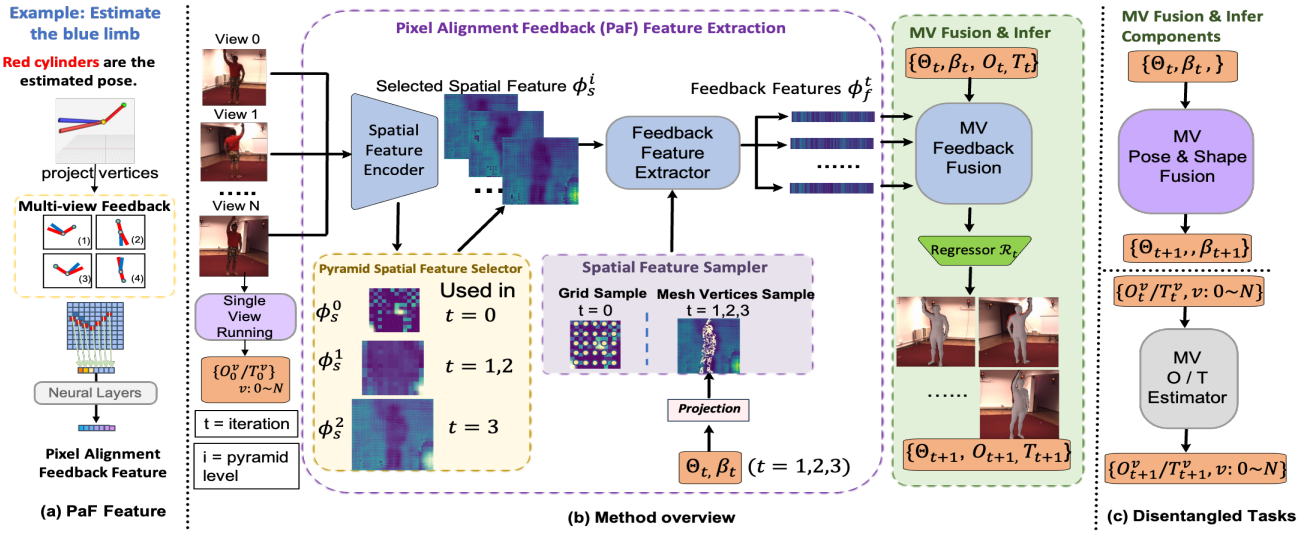

Figure 1: Overview of our proposed Pixel-aligned Feedback Fusion (PaFF) pipeline: (a) Pixel Alignment Feedback (PaF) Feature Extraction. (b) PaFF iteratively refines human body parameters' estimation with the guidance of the PaF feature. (c) The task of multi-view feedback fusion is disentangled into three tasks - Multi-view Pose & Shape Fusion, Multi-view Orientation Estimation, and Multi-view Translation Estimation to incorporate camera parameters in the end-to-end model.

Methodology

Overview

Given N-view images of single human body and camera parameters $\{K^v_{cam}, R^v_{cam}, T^v_{cam}, v = 0, 1, ..., N\}$, the multi-view human mesh recovery task requires estimating the body parameters (typically SMPL (Loper et al. 2015) pose parameter θ and shape parameter β), the global translation T_g, and the global orientation O_g simultaneously. To solve this problem effectively, our proposed method iteratively refines all the parameters with pixel alignment feedback (PaF) features, as illustrated in Fig. 1(b). Specifically, a pre-trained single-view pixel alignment feature extractor runs first to extract an image feature pyramid $\{\Phi^i_s, i = 0, 1, 2\}$ and gives the initialization estimation of body orientation O^v_0 and translation T^v_0 for each view. During multi-view fusion, PaFF adapts the Spatial Feature Sampler to sample multi-view feedback features Φ^t_f on the collected feature pyramid, followed by a multi-view feedback fusion module to aggregate all the feedback features in order to infer parameter updates for each iteration. Note that our model could work in arbitrary numbers of camera views given the camera parameters, yet $N = 4$ is used in the paper. There are 4 iterations of PaFF multi-view regression.

Multi-view Pixel Alignment Feedback Feature

The iterative regression process, which estimates the residual estimation updates iteratively to get the final estimation, has been proven to be beneficial for accurate human mesh recovery (Kanazawa et al. 2018; Kolotouros, Pavlakos, and Daniilidis 2019; Shin and Halilaj 2020). However, previous methods usually utilize feed-forward iteration, which means all iterations only engage the same global feature vector in each iteration. Here comes a critical drawback: decoding pose and shape from a global feature vector is hard to achieve pixel-aligned performance no matter how many iterations are performed. In order to obtain pixel-aligned performance, we seek to add image information to each iteration which serves as the *feedback*. To realize it, we make crucial improvements for multi-view feature extraction by projecting the estimated human mesh vertices to the image plane and constructing pixel-aligned feedback (PaF for short) features by concatenating sampled features from an intermediate feature map, as shown in Fig. 1(a). As the previous study (Dijk and Croon 2019) shows, the pixel position of a feature can be encoded by the neural network. We believe the relative position between the sampled pixels and the pixels inside the real human body region can be encoded into the feedback features and inform the misalignment between the estimated and the real human body in each camera view.

As illustrated in Fig. 1(b), we construct a PaF Feature Extraction Model, in which a Spatial Feature Encoder extracts a coarse-to-fine feature pyramid with three feature maps and a Spatial Feature Sampler samples estimated mesh vertices from the feature maps to get PaF features. During the initial iteration, where no previously estimated vertices are available, we apply grid sampling on the first feature map Φ^0_s to extract the initial point-wise features. There are three additional iterations to align the estimated body with multi-view 2D evidence. During these three iterations, we adopt a Mesh Vertex Sampling method to extract the PaF features from feature maps $\{\Phi^i_s, i = 1, 1, 2\}$ for iteration $t = 1, 2, 3$. Note that Φ^1_s is reused in iterations 1 and 2. Gradually, these feedback features would guide the sampled points closer to the true human area with an updated parameter set.

To make the extracted feedback feature more effective and

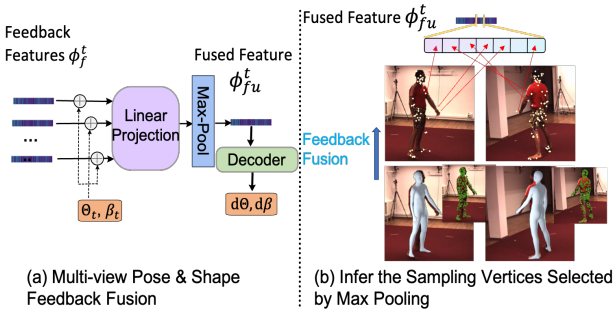

Figure 2: Multi-view Pose & Shape Feedback Fusion Module: a) Pose & shape multi-view feedback fusion Module. b) Visualize vertices selected by max-pooling: The estimated meshes from the previous iteration and the sampled vertices are shown at the bottom. In the top images, we visualize the sampling vertex in each view that contributes most to one feature dimension. By looking at the selected vertices on the left arm and legs, we can find that the vertices which reflect more estimation misalignment are easier to be chosen.

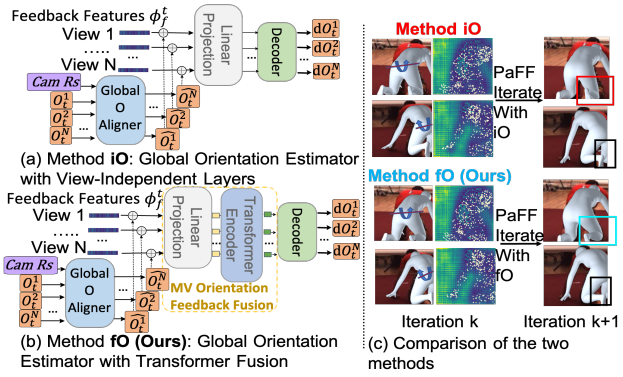

Figure 3: Global Orientation Estimator: (a) Method iO and (b) Method fO are two options for Global Orientation Estimator; (c) shows the motivation of method fO that uses transformer fusion to collect multi-view body parts misalignment signals in order to avoid depth ambiguity and occlusion problems. Estimated models and feature sampling are visualized in the left images, while updated estimations are shown in the right images. Blue arrows indicate the correction rotation needed for each view's estimation. Comparing the refinement effect in one iteration (k=2) of the two methods, method fO performs better in orientation estimation, which also leads to a better estimation of the right leg.

generalized, we pre-train the Pixel-alignment Feedback Extraction Model with the existing 2D datasets by concatenating a monocular inference module. The pre-trained model also takes four iterations to regress the final prediction. O^v and T^v estimations from the last iteration of each view's running would be reserved to initialize the multi-view estimation to make the body projection to align faster. Details of the pre-training are shown in the supplementary material.

Multi-view Feedback Fusion Module

After the extraction of PaF features from multi-view images, a multi-view fusion module will compile the misalignment information of different views before the final inference. As shown in Fig. 1(a), one view's misalignment information of the limb is not adequate to correct false body parts due to depth and shape ambiguity. Therefore, the goal of the feedback feature fusion is to leverage the multi-view misalignment information in one fused feature fully. Different from existing multi-view methods (Shin and Halilaj 2020; Zhang et al. 2021b) which fuse a single multi-view feature to infer all the parameters (θ, β, O_g and T_g), we disentangle the multi-view fusion and inference into three isolated tasks - i) pose & shape estimation, ii) global orientation estimation, and iii) global translation estimation. The benefit of the disentangling is to separate the inferring into two groups, body global pose (orientation and translation) and body local pose (change of joints and shape) in the iterative feedback regression process, which encourages the different modules to focus on their own task scale.

Multi-view Pose and Shape Fusion The PaF feature for pose and shape estimation contains information about the misalignment between the true body and the estimated body. Pose and shape estimations are highly entangled with each other since changes in joint angles or shape parameters can both result in a misalignment, in which bone's skew and shape vertices offset can be hard to distinguish. The angular misalignment from one view is not adequate to infer the true angular misalignment in 3D space, as shown in Fig. 1(a), while the true shape misalignment could be ambiguous in one view due to depth ambiguity. As shown in Fig. 2(a), we first use a linear projection layer to adapt the PaF feature for the pose and shape fusion task. With a belief that maximum feedback values reflecting more essential misalignment information and non-maximum values might contain noise induced by occlusion, to infer complete misalignment information from multi-view feedback features, we propose to apply max-pooling on the stacked multi-view PaF features. Fig. 2(b) shows the selected sampling vertices for the fused feedback features, which illustrates large misalignment information can be retained in the process. However, the max-pooling operation would inevitably filter some misalignment information contained in the non-maximum signals. The iterative update process can be seen as a solution for completing the misalignment information gradually.

Global Orientation, Translation Estimator Different from other multi-view regression-based methods (Liang and Lin 2019; Shin and Halilaj 2020), our method utilizes a feedback feature - PaF to refine the estimation of orientation and translation iteratively in an end-to-end manner. The orientation and translation are 'optimized' at the same time with pose and shape estimations yet with a limited number of iterations. An accurate estimation for orientation and translation is crucial to align the estimation to the multi-view body regions in order to extract informative feedback signals for local pose and shape updates. Following the monocular estimation methods for human reconstruction (Kanazawa et al.

2018; Kolotouros, Pavlakos, and Daniilidis 2019; Zhang et al. 2021a) to predict approximated camera parameters with the orthogonal projection assumption, we build a camera calibration-free version of our method, which has two independent neural networks for single-view body orientation estimation O^v and relative translation estimation TO^v using a default focal length while the shape and pose are still predicted as a joint estimation from multi-view feedback fusion. When calibrated cameras are given, global orientation O_g and global translation T_g can be estimated to rectify each view's camera prediction (O^v and TO^v). Specially, we choose not to update O and T after grid sampling and use the initial estimations since the grid points do not reflect any misalignment information.

The body orientation can be seen as a 'root joint' of the human body. One change of orientation can skew all body parts with image evidence, which gives us the intuition of using pixel alignment features to infer the skew of orientation estimation. As illustrated in Fig. 3(b), we first align the initial single-view orientation estimations with a Global Orientation Aligner Algorithm using camera rotations. The Global Orientation Aligner Algorithm first filter the most orientation-skewed view and update the orientation estimation for each view as $\widehat{O_t^v}$ with camera rotation parameters (check supplementary for the details). After the Global Orientation Aligner, PaF features are extracted with the updated orientation estimations. Inside the PaF feature, the information reflecting the orientation misalignment mainly comes from the overall estimated body parts' skewing. Due to depth ambiguity and occlusion (which can be seen in Fig. 3(c)), black and red boxes), one view's body parts' skewing might not be complete for the orientation correction. So, we design a transformer-based orientation feedback fusion module to capture the co-relationship between multi-view orientation misalignment and complete each view's orientation misalignment information by the attention mechanism. We construct the fusion module with a multi-head transformer encoder with concatenated multi-view feedback features, and single-view orientation estimation as input queries (Vaswani et al. 2017). Since the angle of the correction rotation for each view is the same (multi-view features are fusible), while the rotation axis is different (as shown in Fig. 3(c)), we use each view's updated PaF feature to infer a correction rotation for each view independently. As shown in Fig. 3(b), multi-view PaF features would go through the transformer encoder to predict orientation update dO_t^v in the 6D representation rotation (Zhou et al. 2019). We first update O^v for each view in an addictive fashion, i.e., $O_{t+1}^v = O_t^v + dO_t^v$ and then apply the Global O Aligner to re-align the estimations with camera rotations. A non-transformer version of global O estimator in Fig. 3(a) is compared with our method to illustrate the effect of transformer-based multi-view fusion in Fig. 3(c).

The global translation of the human body T_g can be estimated by triangulation based on keypoint detections (Iskakov et al. 2019; Tu, Wang, and Zeng 2020) or be ignored by predicting offsets from a template model (Lin, Wang, and Liu 2021). However, keypoints detection can be noisy with occlusion and require an additional off-the-shelf model, and translation estimation is needed for our reprojection process. As the 2D image feature is hard to reflect perspective camera pose, we are following the monocular methods, which estimate an orthogonal camera of the human body (Kanazawa et al. 2018; Kolotouros, Pavlakos, and Daniilidis 2019; Zhang et al. 2021a) to predict a relative translation and a body scale for each view. As the relative translation from an orthogonal camera can lead to a scale ambiguity, we solve a global scale by using the camera information. Specifically, we assume the estimated body is aligned after independent orthogonal camera prediction. Then the global location of 3D pelvis and the pelvis 2D keypoint of the estimated body lies in the same camera ray for each view. Since the scale of the human body is the same given by a single shape parameter, we solve an adaptive global scale for the estimated body by estimating a global translation with camera parameters from a linear equation. The adaptive scale is helpful to rescale the body mesh, which will improve the accuracy of the absolute keypoints location estimation. (check derivation details in supplementary)

Training

There are two stages of training. The first stage is to pre-train the Pixel Alignment Feedback Feature Extraction Model with some monocular human capture datasets. In the second stage, the feature extraction module is fixed since the feature is informative enough after the pre-training process. And we train the multi-view fusion modules without the global translation estimation. After the second stage, the body alignment for multi-view images is relatively accurate so that the Global Translation Estimator can be applied to solve scale ambiguity. We use 2D/3D keypoints, joint angles, and shape parameters as the supervision signals and add regularization if shape data is not available. (details in supplementary)

Experiments

The experiments are designed to answer the following questions. 1). How is the performance of PaFF? Does it have a generalization ability? 2). How do the Global Orientation Estimator and the Global Translation Estimator boost the performance? 3). How does the individual component choice affect the performance, and why?

Datasets and Evaluation Metrics We trained our PaF feature extractor on Human3.6M (Ionescu et al. 2013), MPI-INF-3DHP (Mehta et al. 2017), COCO (Lin et al. 2014), MPII (Andriluka et al. 2014), LSP (Johnson and Everingham 2010), LSP Extended (Andriluka et al. 2014) using monocular images for the first training stage. Then we freeze the PaF feature extractor's weight and train the rest of the model on Human3.6M and MPI-INF-3DHP with extrinsic and intrinsic camera parameters (For calibration-free PaFF, we do not use them). The train/test split for Human3.6M and MPI-INF-3DHP follows the previous multi-view estimation works (Liang and Lin 2019; Shin and Halilaj 2020). To evaluate Human3.6M, we remove MPI-INF-3DHP during training since the 3D keypoints ground truth in the dataset is relatively inaccurate. To evaluate MPI-INF-3DHP, we keep MPI-INF-3DHP during training but would constrain some

Methods	MPJPE	PA-MPJPE	Multi-view	Cameras	Parametric
Deep Triangulation (Iskakov et al. 2019)	**20.8***	-	Yes	Known	Keypoints
SPIN (Kolotouros et al. 2019)	62.5	41.1	No	Not Known	Yes
I2L-MeshNet (Moon and Lee 2020)	55.7*	41.1*	No	Not Known	Vertices
Mesh Graphomer (Lin, Wang, and Liu 2021)	51.2*	34.5*	No	Not Known	Vertices
PyMAF (Zhang et al. 2021a)	57.7	40.5	No	Not Known	Yes
MuVS (Huang et al. 2017)	58.2	47.1	Yes	Known	Yes
Shape Aware (Liang and Lin 2019)	79.9	45.1	Yes	Not Known	Yes
Shin and Halilaj (2020)	46.9	32.5	Yes	Known	Yes
ProHMR(Kolotouros et al. 2021)	62.2	34.5	Yes	Not Known	Yes
Yu et al. (2022)	-	33.0	Yes	Not Known	Yes
Calib-free PaFF(Ours)	**44.8**	**28.2**	Yes	Not Known	Yes
PaFF(Ours)	**33.0**	**26.9**	Yes	Known	Yes

Table 1: Result on the Human3.6M. Calib-free means calibration-free. * denotes the results of non-parametric representations.

abnormal losses given by noisy 3D data. For the evaluation Human3.6M, we use MPJPE, PA-MPJPE, and PVE as the evaluation metrics following (Liang and Lin 2019). For the evaluation of MPI-INF-3DHP, we use the same metrics - MPJPE, PCK and AUC with Liang and Lin (2019) and Shin and Halilaj (2020). To evaluate the body orientation estimation, we use the angle error between the estimated orientation rotation matrix and the ground truth matrix, which denotes as 'O Err'. To show the generalization ability of PaFF, we additionally train PaFF on MTC Dataset (Xiang, Joo, and Sheikh 2019) by mixing training with Human 3.6M, of which results will be shown in the supplementary.

Implementation Details We built the Spatial Feature Encoder in the PaF feature extractor upon ResNet-50 (He et al. 2016) and the Feedback Feature Extractor in the PaF feature extractor with 1D convolution layers to downsize the sampled feature. The Multi-view Orientation Fusion Module is constructed by a multi-head transformer encoder with 5 heads and 2 layers. The decoders are all constructed by fully connected layers. We use Adam Kingma and Ba (2014) optimizer with a fixed learning rate 1e-5. The first stage of training takes 30 epochs with batch_size of 64 to learn an effective feedback feature extraction. The second stage of training takes 10 epochs. The batch_sizes for the second stage is 16, and the number of views is 4. For more details about implementation details, please refer to supplementary.

Main Results

Human3.6M We evaluate our PaFF model on Human3.6M dataset and compare it with the previous best method (Shin and Halilaj 2020) and other existing 3D human pose estimation methods, as shown in Table. 1. Our PaFF model achieves 33.02 MPJPE improving the previous best method (Shin and Halilaj 2020) by 29.6% and 26.9 PA-MPJPE improving by 17.2%. By comparing the calibration-free PaFF, Liang and Lin (2019) and Shin and Halilaj (2020), we show PaFF shows the state-of-the-art performance without knowing multi-view cameras. By comparing with MuVS Huang et al. (2017) - a multi-view fitting pipeline using 2D estimated keypoints and 2D silhouette, we show our PaFF can perform better, which implies the advantage of not relying on noisy 2D detection when cameras are few. By comparing with Deep Triangulation (Iskakov et al. 2019), it is seen that PaFF performs close to the direct 3D keypoints estimation.

Methods	MPJPE	PCK	AUC
Liang and Lin (2019)	59.0	95.0	65.0
Shin and Halilaj (2020)	50.2	97.4	65.5
PaFF(Ours)	**48.4**	**98.6**	**67.3**

Table 2: Results on the MPI-INF-3DHP. The higher results of PCK and UAC mean better performances.

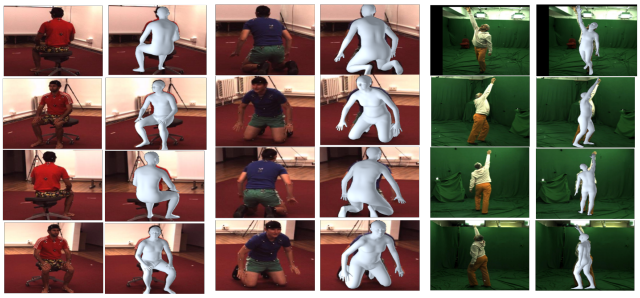

Figure 4: Qualitative results on Human3.6M and MPI-INF-3DHP. Each row means a different view.

MPI-INF-3DHP We evaluate our model on MPI-INF-3DHP to show the generalization of PaFF. As shown in Table 2, our model outperforms the other methods in all of the metrics. It does not show a significant improvement as in Human3.6M due to the noisy 3D labels in MPI-INF-3DHP (Shin and Halilaj 2020).

Qualitative Experiments (More in Supplementary)

Human Mesh Recovery We demonstrate three qualitative examples in Fig. 4 on Human3.6M and MPI-INF-3DHP datasets. Benefiting from the iterative refine process with PaF feedback features, our method could accurately aggregate multi-view information to handle handling object occlusion (chair) and self-contact (the first example). The second and third examples demonstrate the robustness of our PaFF for estimating unusual poses.

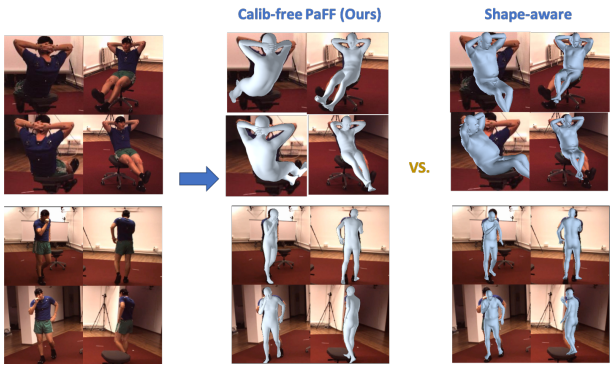

Figure 5: Visualization of the effect of Multi-view PaF Feature by comparing the calibration-free PaFF and Shape-aware (Liang and Lin 2019) in two cases.

Methods	MPJPE	PA-MPJPE	PVE
AP	36.1	29.3	53.9
SOFTMAX_SUM	34.9	28.5	51.3
Transformer + MP	34.4	28.4	51.7
MP(ours)	**33.0**	**26.9**	**48.9**

Table 3: Multi-view Aggregation Functions for Pose & Shape. AP is average pooling. SOFTMAX_SUM is to do softmax first among multi-views then do summation. MP is max-pooling. Transformer + MP goes through a transformer fusion module first then do max-pooling.

Multi-view PaF Feature Fusion To fairly compare with the calibration-free method (Liang and Lin 2019) ('Shape-aware') in Fig. 5, we adapt our method to a calibration-free version ('Calib-free PaFF'). Note that our 'Calib-free PaFF' could align body better because our multi-view PaF feature fusion can alleviate estimation misalignment.

Ablation Study

Multi-view Fusion Architecture Choice for Multi-view Pose & Shape Estimation Fusion using a Transformer encoder for pose & shape estimation has a lower performance than the fully connected layer + max-pooling in Table. 3. The reason for the phenomenon might be that the self-attention mechanism would mess up some views' feedback signals for the multiple pose & shape parameters. Moreover, there are additional options for the aggregation function, such as average pooling (AP) and Softmax_Sum (do softmax first, then sum up). By comparing the performance of these aggregation functions in Table 3, we find that the max-pooling operation performs the best, which shows the effectiveness of keeping the maximum essential feedback signals from the feedback features.

Different Orientation Estimation Methods We compare three different orientation estimation methods as shown in Table. 4. By comparing 'Ind O' and 'Ind O + Align', we can see a clear improvement in MPJPE after applying Global O Aligner, which shows the benefit of aligning with camera rotations. We find that the performance of the transformer encoder + Global O Aligner structure is superior to the view independent O estimation ('Ind O + Align'), which is due to the ability to renew each view's alignment information with the transformer's cross view attention. The angular O Err is improved from the top to the bottom but with a small step, in which some large orientation angle improvements for some hard cases are averaged in the large data.

Methods	MPJPE	PA-MPJPE	PVE	O Err
Ind O	44.9	28.2	65.2	6.5°
Ind O + Align	34.3	27.7	49.8	5.8°
Tran + Align	**33.0**	**26.9**	**48.9**	**5.1°**

Table 4: Global Orientation Fusion Methods: Ind O is the same with Calib-free PaFF except for using real focal length; Ind O + Align additionally uses Global O Aligner. Tran + Align (ours) uses transformer multi-view fusion and Global O Aligner. O Err is in degrees.

Methods	MPJPE	PA-MPJPE	PVE
w./o. Scale	36.5	26.9	58.6
w. Scale (ours)	**33.0**	**26.9**	**48.9**

Table 5: Global Translation Estimation. 'w./o. Scale' does not estimate a global translation. 'w. Scale' estimates a global translation and an adaptive scale.

Different Translation Estimation Settings In Table. 5, by comparing 'w./o. Scale' and 'w. scale', we have observed clear improvements in PMJPE and PVE after adapting inferred body scale, which proves the advantage of solving scale ambiguity induced by the orthogonal cameras.

Conclusions

We present an end-to-end Pixel-aligned Feedback Fusion (PaFF) model to recover a single human mesh from multi-view images. Different from the existing multi-view methods, we extract Pixel Alignment Feedback (PaF) features from images and fuse them with a novel Feedback Fusion Module to infer the misalignment of the current estimation. The feedback fusion module is divided into three fusion modules to disentangle human pose & shape, global orientation, and global translation in an end-to-end manner. Furthermore, we conduct quantitative experiments on Human3.6M and MPI-INF-3DHP to verify the efficacy of our method, which shows a significant improvement over the previous state-of-the-art. Qualitative results further demonstrate PaFF's potential to deal with the uncommon pose, self-occlusion, and close-contact challenges. However, the limited multi-view training data can limit the model for instant use in the wild. In the future, to further improve the performance against the subtle misaligned, a more accurate parametric annotation and a dataset with more diverse human shapes is necessary. Furthermore, our PaFF is an effective and general pipeline, which can be extended to other parametric model regression tasks such as whole-body motion capture (Zhang et al. 2022), 3D hand reconstruction (Zhou et al. 2020; Li et al. 2022), and multiple human motion capture (Huang et al. 2021; Dong et al. 2021).

Acknowledgements

The work is supported by the National Key R&D Program of China (Grant No. 2021ZD0113501) and the China Postdoctoral Science Foundation (Grant No. 2022M721844). We would like to thank Yuxiang Zhang, and Mengcheng Li for their help, feedback, and discussions for this paper.

References

Ajanohoun, J.; Paquette, E.; and Vázquez, C. 2021. Multi-View Human Model Fitting Using Bone Orientation Constraint and Joints Triangulation. In *2021 IEEE International Conference on Image Processing (ICIP)*, 1094–1098. IEEE.

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 3686–3693.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European conference on computer vision*, 561–578. Springer.

Caliskan, A.; Mustafa, A.; Imre, E.; and Hilton, A. 2020. Multi-view consistency loss for improved single-image 3d reconstruction of clothed people. In *Proceedings of the Asian Conference on Computer Vision*.

Carreira, J.; Agrawal, P.; Fragkiadaki, K.; and Malik, J. 2016. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4733–4742.

Corona, E.; Pons-Moll, G.; Alenyà, G.; and Moreno-Noguer, F. 2022. Learned Vertex Descent: A New Direction for 3D Human Model Fitting. arXiv:2205.06254.

Dijk, T. v.; and Croon, G. d. 2019. How do neural networks see depth in single images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2183–2191.

Dong, J.; Fang, Q.; Jiang, W.; Yang, Y.; Huang, Q.; Bao, H.; and Zhou, X. 2021. Fast and robust multi-person 3D pose estimation and tracking from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Guler, R. A.; and Kokkinos, I. 2019. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10884–10894.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, Y.; Yan, R.; Fragkiadaki, K.; and Yu, S.-I. 2020. Epipolar transformer for multi-view human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 1036–1037.

Huang, B.; Shu, Y.; Zhang, T.; and Wang, Y. 2021. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *2021 International Conference on 3D Vision (3DV)*, 710–720. IEEE.

Huang, Y.; Bogo, F.; Lassner, C.; Kanazawa, A.; Gehler, P. V.; Romero, J.; Akhter, I.; and Black, M. J. 2017. Towards accurate marker-less human shape and pose estimation over time. In *2017 international conference on 3D vision (3DV)*, 421–430. IEEE.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339.

Iskakov, K.; Burkov, E.; Lempitsky, V.; and Malkov, Y. 2019. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7718–7727.

Johnson, S.; and Everingham, M. 2010. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *bmvc*, volume 2, 5.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5253–5263.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2252–2261.

Kolotouros, N.; Pavlakos, G.; and Daniilidis, K. 2019. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4501–4510.

Kolotouros, N.; Pavlakos, G.; Jayaraman, D.; and Daniilidis, K. 2021. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11605–11614.

Li, M.; An, L.; Zhang, H.; Wu, L.; Chen, F.; Yu, T.; and Liu, Y. 2022. Interacting attention graph for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2761–2770.

Li, Z.; Oskarsson, M.; and Heyden, A. 2021. 3D Human Pose and Shape Estimation Through Collaborative Learning and Multi-View Model-Fitting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1888–1897.

Liang, J.; and Lin, M. C. 2019. Shape-aware human pose and shape reconstruction using multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4352–4362.

Lin, K.; Wang, L.; and Liu, Z. 2021. Mesh graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12939–12948.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, Y.; Stoll, C.; Gall, J.; Seidel, H.-P.; and Theobalt, C. 2011. Markerless motion capture of interacting characters using multi-view image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1249–1256. IEEE.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6): 1–16.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, 506–516. IEEE.

Moon, G.; and Lee, K. M. 2020. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, 752–768. Springer.

Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; and Schiele, B. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, 484–494. IEEE.

Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7025–7034.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. arXiv:2204.14109.

Rong, Y.; Liu, Z.; Li, C.; Cao, K.; and Loy, C. C. 2019. Delving deep into hybrid annotations for 3d human recovery in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5340–5348.

Rong, Y.; Shiratori, T.; and Joo, H. 2021. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Sengupta, A.; Budvytis, I.; and Cipolla, R. 2021. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16094–16104.

Shin, S.; and Halilaj, E. 2020. Multi-view human pose and shape estimation using learnable volumetric aggregation. arXiv:2011.13427.

Shuai, H.; Wu, L.; and Liu, Q. 2022. Adaptive Multi-view and Temporal Fusing Transformer for 3D Human Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Song, J.; Chen, X.; and Hilliges, O. 2020. Human body model fitting by learned gradient descent. In *European Conference on Computer Vision*, 744–760. Springer.

Sun, Y.; Bao, Q.; Liu, W.; Fu, Y.; Black, M. J.; and Mei, T. 2021. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11179–11188.

Tian, Y.; Zhang, H.; Liu, Y.; and Wang, L. 2022. Recovering 3d human mesh from monocular images: A survey. arXiv:2203.01923.

Tu, H.; Wang, C.; and Zeng, W. 2020. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision*, 197–212. Springer.

Varol, G.; Ceylan, D.; Russell, B.; Yang, J.; Yumer, E.; Laptev, I.; and Schmid, C. 2018. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 20–36.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, S.; Jin, S.; Liu, W.; Bai, L.; Qian, C.; Liu, D.; and Ouyang, W. 2021. Graph-based 3d multi-person pose estimation using multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11148–11157.

Xiang, D.; Joo, H.; and Sheikh, Y. 2019. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10965–10974.

Xu, L.; Liu, Y.; Cheng, W.; Guo, K.; Zhou, G.; Dai, Q.; and Fang, L. 2017. Flycap: Markerless motion capture using multiple autonomous flying cameras. *IEEE transactions on visualization and computer graphics*, 24(8): 2284–2297.

Xuan, H.; Zhang, J.; and Li, K. 2022. MHPro: Multi-Hypothesis Probabilistic Modeling for Human Mesh Recovery. In *Proceedings of the the CAAI International Conference on Artificial Intelligence*.

Yagubbayli, F.; Tonioni, A.; and Tombari, F. 2021. LegoFormer: Transformers for Block-by-Block Multi-view 3D Reconstruction. arXiv:2106.12102.

Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; and Quan, L. 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5525–5534.

Yu, Z.; Zhang, L.; Xu, Y.; Tang, C.; TRAN, L.; Keskin, C.; and Park, H. S. 2022. Multiview Human Body Reconstruction from Uncalibrated Cameras. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Zanfir, A.; Bazavan, E. G.; Zanfir, M.; Freeman, W. T.; Sukthankar, R.; and Sminchisescu, C. 2021. Neural descent for visual 3d human pose and shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14484–14493.

Zhang, H.; Cao, J.; Lu, G.; Ouyang, W.; and Sun, Z. 2020a. Learning 3D human shape and pose from dense body parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, H.; Tian, Y.; Zhang, Y.; Li, M.; An, L.; Sun, Z.; and Liu, Y. 2022. PyMAF-X: Towards Well-aligned Full-body Model Regression from Monocular Images. arXiv:2207.06400.

Zhang, H.; Tian, Y.; Zhou, X.; Ouyang, W.; Liu, Y.; Wang, L.; and Sun, Z. 2021a. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11446–11456.

Zhang, J.; Cai, Y.; Yan, S.; Feng, J.; et al. 2021b. Direct Multi-view Multi-person 3D Pose Estimation. *Advances in Neural Information Processing Systems*, 34.

Zhang, Y.; An, L.; Yu, T.; Li, X.; Li, K.; and Liu, Y. 2020b. 4D association graph for realtime multi-person motion capture using multiple video cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1324–1333.

Zhang, Y.; Li, Z.; An, L.; Li, M.; Yu, T.; and Liu, Y. 2021c. Lightweight multi-person total motion capture using sparse multi-view cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5560–5569.

Zheng, Z.; Yu, T.; Liu, Y.; and Dai, Q. 2021. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, Y.; Barnes, C.; Lu, J.; Yang, J.; and Li, H. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5745–5753.

Zhou, Y.; Habermann, M.; Xu, W.; Habibie, I.; Theobalt, C.; and Xu, F. 2020. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5346–5355.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9308–9316.