CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan *

Huawei Noah's Ark Lab
{zhihao.li,liu.jianzhuang,zhangzhensong,xusongcen,yanyouliang}@huawei.com

Abstract. Top-down methods dominate the field of 3D human pose and shape estimation, because they are decoupled from human detection and allow researchers to focus on the core problem. However, cropping, their first step, discards the location information from the very beginning, which makes themselves unable to accurately predict the global rotation in the original camera coordinate system. To address this problem, we propose to Carry Location Information in Full Frames (CLIFF) into this task. Specifically, we feed more holistic features to CLIFF by concatenating the cropped-image feature with its bounding box information. We calculate the 2D reprojection loss with a broader view of the full frame, taking a projection process similar to that of the person projected in the image. Fed and supervised by global-location-aware information, CLIFF directly predicts the global rotation along with more accurate articulated poses. Besides, we propose a pseudo-ground-truth annotator based on CLIFF, which provides high-quality 3D annotations for in-the-wild 2D datasets and offers crucial full supervision for regression-based methods. Extensive experiments on popular benchmarks show that CLIFF outperforms prior arts by a significant margin, and reaches the first place on the AGORA leaderboard (the SMPL-Algorithms track). The code and data are available at https://github.com/huawei-noah/noah-research/tree/master/CLIFF.

Keywords: 3D human pose, human body reconstruction, 2D to 3D, global rotation, projection, global location, pseudo annotation

1 Introduction

Given a single RGB image, 3D human pose and shape estimation aims to reconstruct human body meshes with the help of statistic models [3,29,37,52]. It is a fundamentally under-constrained problem due to the depth ambiguity. This problem attracts a lot of research because of its key role in many applications such as AR/VR, telepresence, and action analysis.

With the popular parametric human model SMPL [29], regression-based methods [15,20,35] learn to predict the SMPL parameters from image features in a data-driven way, and obtain the meshes from these predictions through a

* Corresponding author

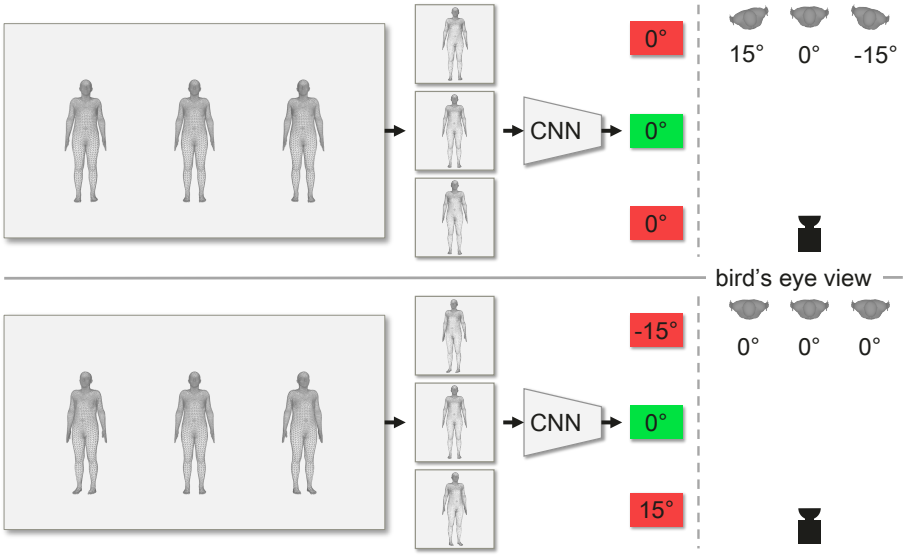


Fig. 1. In the upper part, the cropped images look similar and thus get close predictions from a CNN model. However, two of the three predictions are wrong (marked in red). Actually, they have clearly different global rotations relative to the original camera, which can be seen from the bird's eye view. In the lower part, the three global rotations are the same, but again two estimations are wrong. We only show the yaw angles of the global rotations here for simplicity.

linear function. Like most tasks in computer vision, there are two approaches to do this: top-down [15,22,19,20] and bottom-up [47,54]. The former first detects humans, then crops the regions of interest, and processes each cropped image independently. The latter takes a full image as input and gives the results for all individuals at once. The top-down approach dominates this field currently, because it is decoupled from human detection, and has high recall and precision performances thanks to the mature detection technique [41,42,5].

However, cropping, the first step in the top-down approach discards location information from the very beginning, which is essential in estimating the global rotation in the original camera coordinate system with respect to the full image. Take Fig. 1 as an example, where the original images come from a perspective camera with a common diagonal Field-of-View 55°. After cropping, the input images to the CNN model look similar, and thus get close predictions without surprise (see the upper part of Fig. 1). In fact, the three persons have clearly different global rotations, which can be inferred from the full image. The same problem exists for other 2D evidences such as 2D keypoints. As a result, the 2D reprojection loss calculated in the cropped images is not a proper supervision, which tends to twist articulated poses to compensate the global rotation error. In another word, missing the location information introduces extra ambiguity.

To fix this problem, we propose CLIFF: Carrying Location Information in Full Frames into 3D human pose and shape estimation, by making two major modifications to the previous top-down approach. First, CLIFF takes more holistic features as input. Besides the latent code of the resolution-fixed cropped image, the bounding box information is also fed to CLIFF, which encodes its discarded location and size in the original full image, providing the model with adequate information to estimate the global rotation. Second, CLIFF calculates the 2D reprojection loss with a broader view of the full frame. We obtain the predicted 3D joints in the original camera coordinate system, and project them onto the full image instead of the cropped one. Consequently, these predicted 2D keypoints have a projection process and perspective distortion similar to those of the person projected in the image, which is important for them to correctly supervise 3D predictions in an indirect way. Fed and supervised by global-location-aware information, CLIFF is able to predict the global rotation relative to the original camera along with more accurate articulated poses.

On the other hand, the regression-based methods need full supervision for SMPL parameters to boost their performances [22]. However, it costs a lot of time and effort to obtain these 3D annotations, using multi-view motion capture systems [13,32] or a set of IMU devices [48,31]. Moreover, the lack of diversity in actors and scenes limits the generalization abilities of these data-driven models. On the contrary, 2D keypoints are straightforward and inexpensive to annotate for a wide variety of in-the-wild images with diverse persons and backgrounds [27,2]. Hence, some CNN-based pseudo-ground-truth (pseudo-GT) annotators [22,14,34] are introduced to lift these 2D keypoints up to 3D poses for full supervision. Since these annotators are based on previous top-down models that are agnostic to the person location in the full frame, they produce inaccurate annotations, especially for the global rotation.

We propose a novel annotator based on CLIFF with a global perspective of the full frame, which produces high-quality annotations of human model parameters. Specifically, we first pretrain the CLIFF annotator on several datasets with available 3D ground truth, and then test it on the target dataset to predict SMPL parameters. Using these predictions as regularization and ground-truth 2D keypoints as weak supervision, we finetune the pretrained model on the target dataset, and finally test on it to infer SMPL parameters as pseudo-GT. With the implicit prior of the pretrained weights and the explicit prior of the SMPL parameter predictions, the CLIFF annotator alleviates the inherent depth ambiguity to recover feasible 3D annotations from monocular images.

Our contributions are summarized as follows:

- We reveal that the global rotations cannot be accurately inferred when only using cropped images, which is ignored by previous methods, and propose CLIFF to deal with this problem by feeding and supervising the model with global-location-aware information.
- Based on CLIFF, we propose a pseudo-GT annotator with strong priors to generate high-quality 3D annotations for in-the-wild images, which are demonstrated to be very helpful in performance boost.

– We conduct extensive experiments on popular benchmarks and show that CLIFF outperforms prior arts by significant margins on several evaluation metrics (e.g., 5.7mm MPJPE and 6.5mm PVE on 3DPW), and reaches the first place on the AGORA leaderboard (the SMPL-Algorithms track) way ahead of other methods.

2 Related Work

3D human pose estimation. 3D human pose is usually represented as a skeleton of 3D joints [45,51,53], or a mesh of triangle vertices [15,19,33,7,25,26]. These vertex locations are inferred directly by model-free methods [33,7,25,26], or obtained indirectly from parametric model (e.g., SMPL [29]) predictions of model-based methods [15,22,19]. Optimization-based methods [4] are first proposed to iteratively fit the SMPL model to 2D evidences, while regression-based ones [15] make the predictions in a straightforward way that may support real time applications. Both top-down [22,24,20,26] and bottom-up [47] approaches can do the job. Because our method is a top-down framework to regress the SMPL parameters, we only review the most relevant work, and refer the reader to [28] for a more comprehensive survey.

Input and supervision. Previous top-down methods take as input the cropped image [15] or/and 2D keypoints in the cropped region [7], perhaps with additional camera parameters [21]. They generally project the predicted 3D joints onto the cropped image to compute the reprojection loss for supervision. Since the location information is lost in the cropped image, it is difficult for them to estimate an accurate global rotation. To solve this problem, Kissos et al. [18] use the prediction with respect to the cropped image as initialization, and then use SMPLify [4] to refine the results for better pixel alignment. Since SMPLify computes the reprojection loss in the full image, they obtain a better global rotation in the end. However, as an optimization approach, SMPLify is very slow and may also harm the articulated pose estimation, as shown in EFT [14] and NeuralAnnot [34]. PCL [53] warps the cropped image to remove the perspective distortion, and corrects the global rotation via a post-processing, but the articulated poses cannot be corrected. Unlike these methods, CLIFF exploits the location information in both input and supervision, to predict more accurate global rotation and articulated poses simultaneously, without any post-processing.

Pseudo-GT annotators. It is well known that training with in-the-wild images can bring significant benefits to the generalization performances of regression-based models. However, it is hard to obtain the corresponding 3D ground truth, so pseudo-GT annotators are proposed. Optimization-based annotators [4] throw the images away, and fit the human model to 2D keypoints by minimizing the reprojection loss. CNN-based annotators [22,14,34,23,10] are recently introduced to get better results, taking the cropped images as input. They all need priors

to deal with the depth ambiguity. An extra model such as GMM [4], GAN [15,8] or VAE [40,19] is trained on a large motion capture dataset AMASS [30] to be an implicit prior. Other methods search for plausible SMPL parameters that may be close to the ground truth to be an explicit prior [36,10]. We propose a novel annotator based on CLIFF, which takes more than the cropped image as input and calculates the 2D supervision in the full image, and use the SMPL parameter predictions by the pretrained CLIFF annotator as an effective explicit prior, without an extra model and human actors mimicking the predefined poses [36,21].

3 Approach

In this section, we first briefly review the commonly-used parametric model SMPL and the baseline method HMR, then propose our model CLIFF, and finally present a novel pseudo-GT annotator for in-the-wild images.

3.1 SMPL Model

As a parametric human body model, SMPL [29] provides a differentiable function, which takes dozens of parameters $\Theta = \{\boldsymbol{\theta}, \boldsymbol{\beta}\}$ as input, and returns a posed 3D mesh of 6890 vertices V, where $\boldsymbol{\theta} \in \mathbb{R}^{24\times3}$ and $\boldsymbol{\beta} \in \mathbb{R}^{10}$ are the pose and shape parameters, respectively. The pose parameters $\boldsymbol{\theta}$ consist of the global rotation of the root joint (pelvis) with respect to some coordinate system (e.g., the camera coordinate system in our work), and 23 local rotations of other articulated joints relative to their parents along the kinematic tree. The k joints of interest $J_{3D} \in \mathbb{R}^{k\times3}$ can be obtained by the linear combination of the mesh vertices, $J_{3D} = MV$, where M is a pretrained sparse matrix for these k joints.

3.2 HMR Model

HMR [15] is a simple and widely-used top-down method for 3D human pose and shape estimation. Its architecture is shown in Fig. 2(a). A square cropped image is resized to 224×224 and passed through a convolutional encoder. Then an iterative MLP regressor predicts the SMPL parameters $\Theta = \{\boldsymbol{\theta}, \boldsymbol{\beta}\}$ and weak-perspective projection parameters $P_{weak} = \{s, t_x, t_y\}$ for a virtual camera M_{crop} with respect to the cropped image (see Fig. 3), where s is the scale parameter, and t_x and t_y are the root translations relative to M_{crop} along the X and Y axes, respectively. With a predefined large focal length $f_{HMR} = 5000$, P_{weak} can be transformed to perspective projection parameters $P_{persp} = \{f_{HMR}, \mathbf{t}^{crop}\}$, where $\mathbf{t}^{crop} = [t_X^{crop}, t_Y^{crop}, t_Z^{crop}]$ denotes the root translations relative to M_{crop} along the X, Y, and Z axes respectively:

$$t_X^{crop} = t_x,\ t_Y^{crop} = t_y,\ t_Z^{crop} = \frac{2 \cdot f_{HMR}}{r \cdot s}, \tag{1}$$

where $r = 224$ denotes the side resolution of the resized square crop.

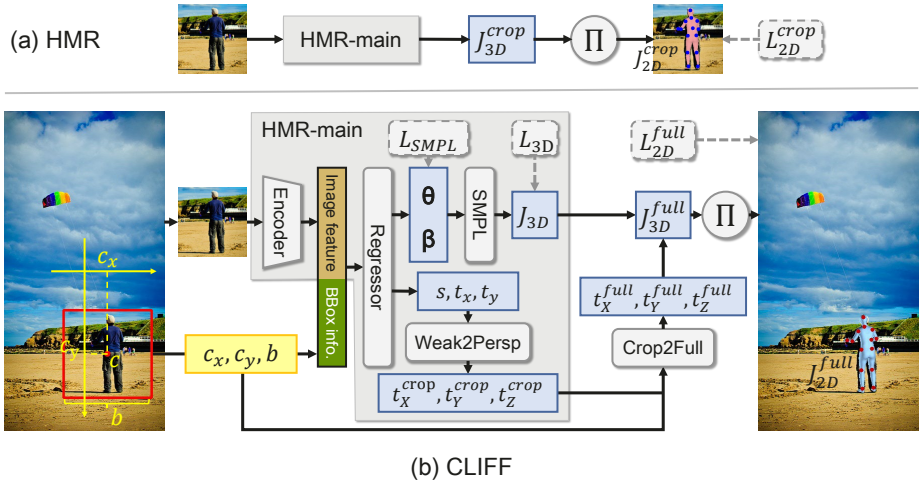


Fig. 2. (a) HMR architecture where the detail of its main part "HMR-main" is shown in the gray box in (b). (b) CLIFF architecture that takes HMR as its backbone and has two modifications: 1) CLIFF takes as input not only the cropped image but also the bounding box information. 2) CLIFF projects the predicted 3D joints onto the original full image plane to compute the 2D reprojection loss, while HMR does this in the cropped image.

The loss of HMR is defined as:

$$L^{HMR} = \lambda_{SMPL} L_{SMPL} + \lambda_{3D} L_{3D} + \lambda_{2D} L_{2D}^{crop}, \tag{2}$$

and its terms are calculated by:

$$L_{SMPL} = \left\| \Theta - \hat{\Theta} \right\|, L_{3D} = \left\| J_{3D} - \hat{J}_{3D} \right\|, L_{2D}^{crop} = \left\| J_{2D}^{crop} - \hat{J}_{2D}^{crop} \right\|, \tag{3}$$

where $\hat{\Theta}$, $\hat{J}_{3D}$, and $\hat{J}_{2D}^{crop}$ are the ground truth, and the predicted 2D keypoints in the cropped image are obtained by the perspective projection Π:

$$J_{2D}^{crop} = \Pi J_{3D}^{crop} = \Pi(J_{3D} + \mathbf{1}\mathbf{t}^{crop}), \tag{4}$$

where $\mathbf{1} \in \mathbb{R}^{k \times 1}$ is an all-ones matrix.

3.3 CLIFF Model

As described above, previous top-down methods take only the cropped image as input, and calculate the reprojection loss in the cropped image, which may lead to inaccurate predictions, as emphasized in Section 1. To address this problem, we take HMR as the baseline and propose to make two modifications to build CLIFF, as shown in Fig. 2(b).

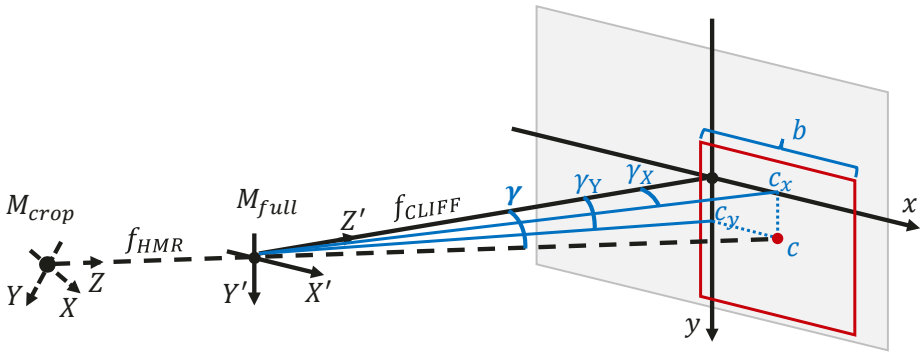


Fig. 3. The geometric relation between the virtual camera M_{crop} for the cropped image (the red rectangle) and the original camera M_{full} for the full image.

First, CLIFF takes more holistic features as input. Besides the encoded image feature, the additional bounding box information I_{bbox} of the cropped region is also fed to the regressor,

$$I_{bbox} = [\frac{c_x}{f_{CLIFF}}, \frac{c_y}{f_{CLIFF}}, \frac{b}{f_{CLIFF}}], \tag{5}$$

where (c_x, c_y) is its location relative to the full image center, b its original size, and f_{CLIFF} the focal length of the original camera M_{full} used in CLIFF (see Fig. 3). Besides the effect of normalization, taking f_{CLIFF} as the denominator gives geometric meaning to the first two terms in I_{bbox}:

$$\begin{aligned} \tan\gamma_X &= \frac{c_x}{f_{CLIFF}}, \\ \tan\gamma_Y &= \frac{c_y}{f_{CLIFF}}, \end{aligned} \tag{6}$$

where $\boldsymbol{\gamma} = [\gamma_X, \gamma_Y, 0]$ is the transformation angle from M_{crop} to the original camera M_{full} coordinate system with respect to the full image, as shown in Fig. 3. Therefore, fed with I_{bbox} as part of the input, the regressor can make the transformation implicitly to predict the global rotation relative to M_{full}, which also bring benefits to the articulated pose estimation. As for the focal length, we use the ground truth if it is known; otherwise we approximately estimate its value as $f_{CLIFF} = \sqrt{w^2 + h^2}$, where w and h are the width and height of the full image respectively, corresponding to a diagonal Field-of-View of 55° for M_{full}, following the previous work [18].

Second, CLIFF calculates the reprojection loss in the full image instead of the cropped one. The root translation is transformed from M_{crop} to M_{full}:

$$\begin{aligned} t_X^{full} &= t_X^{crop} + \frac{2 \cdot c_x}{b \cdot s}, \\ t_Y^{full} &= t_Y^{crop} + \frac{2 \cdot c_y}{b \cdot s}, \\ t_Z^{full} &= t_Z^{crop} \cdot \frac{f_{CLIFF}}{f_{HMR}} \cdot \frac{r}{b}, \end{aligned} \tag{7}$$

where $\mathbf{t}^{full} = [t_X^{full}, t_Y^{full}, t_Z^{full}]$ denotes the root translations relative to M_{full} along the $X^{'}$, $Y^{'}$, and $Z^{'}$ axes, respectively. The derivation can be found in our supplementary materials. Then we project the predicted 3D joints onto the full image plane:

$$J_{2D}^{full} = \Pi J_{3D}^{full} = \Pi(J_{3D} + \mathbf{1}\mathbf{t}^{full}), \tag{8}$$

to compute the 2D reprojection loss:

$$L_{2D}^{full} = \left\| J_{2D}^{full} - \hat{J}_{2D}^{full} \right\|, \tag{9}$$

where the ground truth $\hat{J}_{2D}^{full}$ is also relative to the full image center. Finally, the total loss of CLIFF is calculated by:

$$L^{CLIFF} = \lambda_{SMPL} L_{SMPL} + \lambda_{3D} L_{3D} + \lambda_{2D} L_{2D}^{full}. \tag{10}$$

The predicted 2D keypoints J_{2D}^{full} share a similar projection process and perspective distortion with the person in the image, especially when the focal length f_{CLIFF} is close to its ground truth. Thus, the corresponding loss L_{2D}^{full} can correctly supervise CLIFF to make more accurate prediction of 3D human poses, especially the global rotation, which is demonstrated in our experiments.

3.4 CLIFF Annotator

Full supervision from 3D ground truth (particularly the SMPL parameters) is crucial for regression-based methods to improve their performances on 3D human pose and shape estimation [22]. However, these annotations are scarce for in-the-wild datasets, since they require specialized devices and cost a lot of time and labor [13,31]. Recently, CNN-based pseudo-GT annotators [22,14,34] are proposed to address this problem. However, their base models are agnostic to the person locations in full frames, and thus they produce inaccurate annotations, especially for the global rotations, as mentioned in Section 1.

Hence, we propose an annotator based on CLIFF, which is fed and supervised with global-location-aware information, and thus produces better global rotation and articulated pose annotations simultaneously. As shown in Fig. 4, there are four steps in our pipeline to annotate an in-the-wild dataset with only 2D keypoint ground truth.

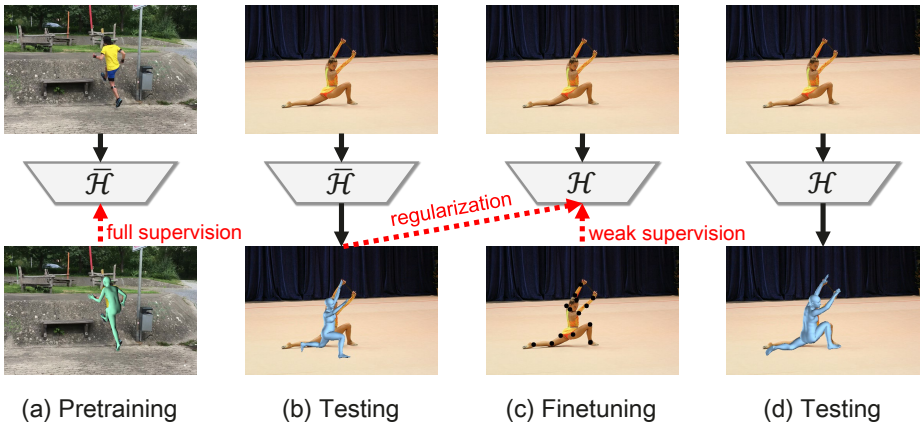


Fig. 4. Our pipeline to generate 3D annotations for a 2D dataset. (a) Pretrain CLIFF model $\bar{\mathcal{H}}$ with full supervision on available datasets. (b) Then test $\bar{\mathcal{H}}$ on the target dataset to predict SMPL parameters $\bar{\Theta}$. (c) Finetune $\bar{\mathcal{H}}$ with 2D keypoints as weak supervision and $\bar{\Theta}$ as regularization, obtaining the updated model $\mathcal{H}$. (d) Finally, test $\mathcal{H}$ on the target dataset to infer SMPL parameters as pseudo-GT.

1. We pretrain the CLIFF annotator $\bar{\mathcal{H}}$ on several datasets with ground-truth SMPL parameters, including 3D datasets and 2D datasets with pseudo-GT generated by EFT [14]. The pretrained weights serve as an implicit prior from these various datasets for the following optimization in Step 3.
2. Test the pretrained model $\bar{\mathcal{H}}$ on the target dataset to predict SMPL parameters $\bar{\Theta}$. Although these predictions may not be accurate, they can be an explicit prior to guide the optimization, and cost little without the need for crowd-sourced participants to mimick some predefined poses [36,21].
3. Finetune the pretrained model $\bar{\mathcal{H}}$ on the target dataset, using ground-truth 2D keypoints as weak supervision and $\bar{\Theta}$ as regularization, to get the updated annotator $\mathcal{H}$. Due to the depth ambiguity, these 2D keypoints are insufficient to supervise the optimization to recover their 3D ground truth. Thus, the priors are very important because they can prevent $\mathcal{H}$ from overfitting to these 2D keypoints and from offering implausible solutions.
4. Test $\mathcal{H}$ on the target dataset to get SMPL parameter predictions Θ as the final pseudo-GT. In our experiment, the reconstructed 3D meshes from these pseudo-GT are pixel-aligned to their 2D evidences, and also perceptually realistic which can be confirmed from novel views.

Compared with other annotators whose priors come from extra models [4,40] trained on another large motion capture dataset AMASS [30], the CLIFF annotator contains strong priors that are efficient to obtain with no need for an extra model and AMASS. More importantly, based on CLIFF, the annotator produces much better pseudo-GT which is very helpful to boost the training performance as shown in the experiments.

3.5 Implementation details

We train CLIFF for 244K steps with batch size 256, using the Adam optimizer [17]. The learning rate is set to $1 \times e^{-4}$ and reduced by a factor of 10 in the middle. The image encoder is pretrained on ImageNet [9]. The cropped images are resized to 224×224, preserving the aspect ratio. Data augmentation includes random rotations and scaling, horizontal flipping, synthetic occlusion [44], and random cropping [14,43]. To annotate an in-the-wild dataset, we train the CLIFF annotator for 30 epochs with learning rate $5 \times e^{-5}$ but no data augmentation. We use MindSpore [1] and PyTorch [38] for the implementation.

4 Experiments and Results

Datasets. Following previous work, we train CLIFF on a mixture of 3D datasets (Human3.6M [13] and MPI-INF-3DHP [32]), and 2D datasets (COCO [27] and MPII [2]) with pseudo-GT provided by the CLIFF annotator. The evaluation is performed on three datasets: 1) 3DPW [31], an in-the-wild dataset with 3D annotations from IMU devices; 2) Human3.6M, an indoor dataset providing 3D ground truth from optical markers with a multi-view setup; 3) AGORA [39], a synthetic dataset with highly accurate 3D ground truth. We use the 3DPW and AGORA training data when conducting experiments on them respectively.

Evaluation metrics. The three standard metrics in our experiments are briefly described below. They all measure the Euclidean distances (in millimeter (mm)) of 3D points between the predictions and ground truth.

MPJPE (Mean Per Joint Position Error) first aligns the predicted and ground-truth 3D joints at the pelvis, and then calculates their distances, which comprehensively evaluates the predicted poses and shapes, including the global rotations.
PA-MPJPE (Procrustes-Aligned Mean Per Joint Position Error, or reconstruction error) performs Procrustes alignment before computing MPJPE, which mainly measures the articulated poses, eliminating the discrepancies in scale and global rotation.
PVE (Per Vertex Error, or MVE used in the AGORA evaluation) does the same alignment as MPJPE at first, but calculates the distances of vertices on the human mesh surfaces.

4.1 Comparisons with State-of-the-art Methods

Quantitative results. We compare CLIFF with prior arts, including video-based methods [16,6,19,50] that exploit temporal information, and frame-based ones [7,24,26,22] that process each frame independently. They could be model-based [15,20] or model-free [33,25], and most of them are top-down methods, except for one bottom-up [47]. As shown in Table 1, CLIFF outperforms them

Table 1. Performance comparison between CLIFF and state-of-the-art methods on 3DPW, Human3.6M and AGORA

		3DPW			Human3.6M		AGORA	
	Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	MVE ↓
video	HMMR [16]	116.5	72.6	-	-	56.9	-	-
	TCMR [6]	86.5	52.7	102.9	-	-	-	-
	VIBE [19]	82.7	51.9	99.1	65.6	41.1	-	-
	MAED [50]	79.1	45.7	92.6	56.4	38.7	-	-
model-free	I2L-MeshNet [33]	93.2	58.6	110.1	-	-	-	-
	Pose2Mesh [7]	89.5	56.3	105.3	64.9	46.3	-	-
	HybrIK [24]	80.0	48.8	94.5	54.4	34.5	-	-
	METRO [25]	77.1	47.9	88.2	54.0	36.7	-	-
	Graphormer [26]	74.7	45.6	87.7	51.2	34.5	-	-
model-based	HMR [15]	130.0	81.3	-	-	56.8	180.5	173.6
	SPIN [22]	96.9	59.2	116.4	-	41.1	153.4	148.9
	SPEC [21]	96.5	53.2	118.5	-	-	112.3	106.5
	HMR-EFT [14]	85.1	52.2	98.7	63.2	43.8	165.4	159.0
	PARE [20]	79.1	46.4	94.2	-	-	146.2	140.9
	ROMP [47]	76.7	47.3	93.4	-	-	116.6	113.8
	CLIFF (Res-50)	72.0	45.7	85.3	50.5	35.1	91.7	86.3
	CLIFF (HR-W48)	**69.0**	**43.0**	**81.2**	**47.1**	**32.7**	**81.0**	**76.0**

Table 2. Ablation study of CLIFF on Human3.6M

Method	MPJPE ↓	PA-MPJPE ↓
CLIFF w/o CI & CS	85.2	54.5
CLIFF w/o CI	84.0	52.4
CLIFF	**81.4**	**52.1**

Table 3. Direct comparison among pseudo-GT annotators on 3DPW

Annotator	MPJPE ↓	PA-MPJPE ↓	PVE ↓
ProHMR [23]	-	52.4	-
BOA [11]	77.2	49.5	-
EFT [14]	-	49.3	-
DynaBOA [10]	65.5	40.4	82.0
Pose2Mesh [7]	65.1	34.6	-
CLIFF (Ours)	**52.8**	**32.8**	**61.5**

by significant margins in all metrics on these three evaluation datasets. With the same image encoder backbone (ResNet-50 [12]) and similar computation cost, CLIFF beats its baseline HMR-EFT, reducing the errors by more than 13mm on MPJPE and PVE. In the case of similar PA-MPJPE to other methods, CLIFF still has lower MPJPE and PVE, since it has a better global rotation estimation. With HRNet-W48 [46], CLIFF has better performance and distinct advantages over previous state-of-the-art, including METRO [25] and Mesh Graphormer [26] which have similar image encoder backbones (HRNet-W64) and transformer-based architectures [49]. CLIFF reaches *the first place* on the AGORA leaderboard (the SMPL-Algorithms track) way ahead of other methods (whose results are from the leaderboard).

Qualitative results. As shown in Fig. 5, we compare CLIFF with PARE [20] on the 3DPW testing data, which is one of the best cropped-image-based methods. We render the reconstructed meshes using the original camera with ground-truth intrinsic parameters. Even though accurate articulated poses can also be



Fig. 5. Qualitative comparison on 3DPW. From left to right: input images, CLIFF predictions, PARE predictions, and their visualizations from novel views (green for the ground truth, blue for CLIFF, and red for PARE).

obtained by PARE, we can see clear pixel-misalignment of its results overlaid on the images, suffering from its inferior global rotation estimation. From the novel viewpoints, we can see that the predicted meshes by CLIFF overlay with the ground truth better than those by PARE, thanks to its more accurate global rotation estimation.

4.2 Ablation Study

We take HMR as the baseline and make two modifications to build CLIFF: additional input of the bounding box information (CI, denoting the CLIFF Input), and the 2D reprojection loss calculated in the full frame for supervision (CS, denoting the CLIFF Supervision). As shown in Table 2, we conduct an ablation study on Human3.6M, since it has accurate 3D ground truth. Without CI providing enough information, MPJPE increases significantly, indicating a worse global rotation estimation. It causes larger errors when we also drop CS that can guide CLIFF to better predictions. This study validates our intuition that global-location-aware information helps the model predict the global rotation and obtain more accurate articulated poses.

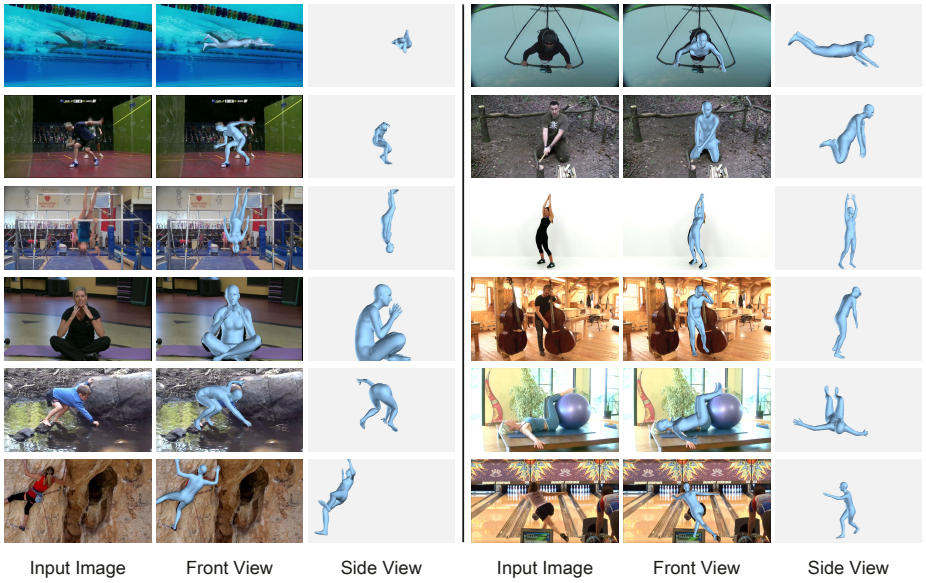


Fig. 6. Qualitative results of the CLIFF annotator on the MPII dataset. From left to right: input images, the front view and side view of the pseudo-GT.

Table 4. Indirect comparison among pseudo-GT annotators by training on COCO and then evaluating on 3DPW and Human3.6M

	3DPW			Human3.6M	
Annotator	MPJPE ↓	PVE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
SPIN [22]	101.2	119.2	65.2	115.0	66.9
EFT [14]	98.8	115.8	62.0	110.9	60.7
CLIFF (Ours)	**85.4**	**100.5**	**53.6**	**96.1**	**54.8**

4.3 Annotator Comparisons

Direct comparison. In Table 3, we directly compare different pseudo-GT annotators on 3DPW, for it is an in-the-wild dataset with ground-truth SMPL parameters. The CLIFF annotator outperforms other methods in all metrics. Even with similar PA-MPJPE to Pose2Mesh [7], CLIFF reduces MPJPE by 12.3mm, and PVE by 20.5mm. Compared to EFT [14] that finetunes a pretrained model on each example, the CLIFF annotator is trained in a mini-batch manner, which helps it maintain the implicit prior all the way. With the additional explicit prior, there is no need for our annotator to choose a generic stopping criterion carefully. It only takes about 30 minutes for the CLIFF annotator to annotate the whole 35,515 images with 4 Tesla V100 GPUs (the finetuning and final testing steps described in Section 3.4).

Indirect comparison. We train CLIFF on the COCO dataset with pseudo-GT from different annotators, and show their results on 3DPW and Human3.6M. The training lasts for 110K steps without learning rate decay. As shown in Table 4, the CLIFF annotator has much better performance than SPIN [22] and EFT [14] (more than 13mm margins on MPJPE and PVE). It demonstrates that the CLIFF annotator can generate high-quality pseudo-GT for in-the-wild images with only 2D annotations, which helps to improve performances significantly.

Qualitative results. In Fig. 6, we show pseudo-GT samples generated by the CLIFF annotator. With good 2D keypoint annotations, the reconstructed meshes are pixel-aligned to the image evidence. From the side view, we can see that they are also perceptually realistic without obvious artifacts, thanks to the strong priors in the CLIFF annotator.

5 Discussion

In Section 3, we show how to build CLIFF based on HMR by making two modifications. We believe that the idea can also be applied to many other methods. First, it can benefit regression-based top-down methods that work on the cropped-region features (e.g., image, keypoint, edge, and silhouette). As for bottom-up methods that treat all the subjects without distinction of their different locations, we can take another form to encode the location information, for example, a location map which consists of a normalized coordinate for each pixel. Going beyond 3D human pose estimation, we can apply the idea to other 3D tasks that involve object global rotations (e.g., 3D object detection and 6-DoF object pose estimation). Even when there are perfect 3D annotations and thus no need for the 2D reprojection loss calculated in the full image, it is still important to take the global-location-aware information as input.

6 Conclusion

Although translation invariance is a key factor for CNN models to succeed in computer vision, we argue that the global location information in full frames matters in 3D human pose and shape estimation, and the global rotations cannot be accurately inferred when only using cropped images. To address this problem, we propose CLIFF by feeding and supervising the model with global-location-aware information. CLIFF takes not only the cropped image but also its bounding box information as input. It calculates the 2D reprojection loss in the full image instead of the cropped one, projecting the predicted 3D joints in a way similar to that of the person projected in the image. Moreover, based on CLIFF, we propose a novel pseudo-GT annotator for in-the-wild 2D datasets, which generates high-quality 3D annotations to help regression-based models boost their performances. Extensive experiments on popular benchmarks show that CLIFF outperforms state-of-the-art methods by a significant margin and reaches the first place on the AGORA leaderboard (the SMPL-Algorithms track).

References

1. Mindspore. https://www.mindspore.cn
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
3. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: SIGGRAPH (2005)
4. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: ECCV (2016)
5. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: CVPR (2019)
6. Choi, H., Moon, G., Chang, J.Y., Lee, K.M.: Beyond static features for temporally consistent 3d human pose and shape from a video. In: CVPR (2021)
7. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: ECCV (2020)
8. Davydov, A., Remizova, A., Constantin, V., Honari, S., Salzmann, M., Fua, P.: Adversarial parametric pose prior. arXiv:2112.04203 (2021)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
10. Guan, S., Xu, J., He, M.Z., Wang, Y., Ni, B., Yang, X.: Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation. arXiv:2111.04017 (2021)
11. Guan, S., Xu, J., Wang, Y., Ni, B., Yang, X.: Bilevel online adaptation for out-of-domain human mesh reconstruction. In: CVPR (2021)
12. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV (2016)
13. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. PAMI (2013)
14. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In: 3DV (2021)
15. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
16. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR (2019)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
18. Kissos, I., Fritz, L., Goldman, M., Meir, O., Oks, E., Kliger, M.: Beyond weak perspective for monocular 3d human pose estimation. In: ECCV (2020)
19. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: CVPR (2020)
20. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: ICCV (2021)
21. Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: Spec: Seeing people in the wild with an estimated camera. In: ICCV (2021)
22. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
23. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: ICCV (2021)

24. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: CVPR (2021)
25. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: CVPR (2021)
26. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: ICCV (2021)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
28. Liu, W., Bao, Q., Sun, Y., Mei, T.: Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective. arXiv:2104.11536 (2021)
29. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. TOG (2015)
30. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: ICCV (2019)
31. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018)
32. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3DV (2017)
33. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: ECCV (2020)
34. Moon, G., Lee, K.M.: Neuralannot: Neural annotator for in-the-wild expressive 3d human pose and mesh training sets. arXiv:2011.11232 (2020)
35. Moon, G., Lee, K.M.: Pose2pose: 3d positional pose-guided 3d rotational pose prediction for expressive 3d human pose and mesh estimation. arXiv:2011.11534 (2020)
36. Muller, L., Osman, A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: CVPR (2021)
37. Osman, A.A., Bolkart, T., Black, M.J.: Star: Sparse trained articulated human body regressor. In: ECCV (2020)
38. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
39. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: Agora: Avatars in geography optimized for regression analysis. In: CVPR (2021)
40. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)
41. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv:1804.02767 (2018)
42. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems (2015)
43. Rockwell, C., Fouhey, D.F.: Full-body awareness from partial observations. In: ECCV (2020)
44. Sárándi, I., Linder, T., Arras, K.O., Leibe, B.: How robust is 3d human pose estimation to occlusion? arXiv:1808.09316 (2018)
45. Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A.: Monocular 3d human pose estimation by generation and ordinal ranking. In: ICCV (2019)

46. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR (2019)
47. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: ICCV (2021)
48. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total capture: 3d human pose estimation fusing video and inertial sensors. In: BMVC (2017)
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems (2017)
50. Wan, Z., Li, Z., Tian, M., Liu, J., Yi, S., Li, H.: Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In: ICCV (2021)
51. Wehrbein, T., Rudolph, M., Rosenhahn, B., Wandt, B.: Probabilistic monocular 3d human pose estimation with normalizing flows. In: ICCV (2021)
52. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: CVPR (2020)
53. Yu, F., Salzmann, M., Fua, P., Rhodin, H.: Pcls: Geometry-aware neural reconstruction of 3d pose with perspective crop layers. In: CVPR (2021)
54. Zhang, J., Yu, D., Liew, J.H., Nie, X., Feng, J.: Body meshes as points. In: CVPR (2021)