Supplementary Materials

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

In these supplementary materials, we first derive Equation 7 given in the main paper, then take a further discussion about the CLIFF input and its performance when applied to videos, and finally provide more details about the CLIFF annotator.

1 Derivation of Equation 7

Equation 7. CLIFF computes the reprojection loss in the full frame instead of the cropped image, so we need to calculate the root translation $\mathbf{t}^{full} = [t_X^{full}, t_Y^{full}, t_Z^{full}]$ in the coordinate system of the original camera M_{full}. Inserting Equation 1 into Equation 7, we have:

$$\begin{aligned} t_X^{full} &= t_x + \frac{2 \cdot c_x}{b \cdot s}, \\ t_Y^{full} &= t_y + \frac{2 \cdot c_y}{b \cdot s}, \\ t_Z^{full} &= \frac{2 \cdot f_{CLIFF}}{b \cdot s}, \end{aligned} \tag{7.1}$$

where s, t_x, and t_y are the scale and translation parameters of the weak-perspective projection, (c_x, c_y) is the crop location relative to the full image center, b is the size of the original crop (detection result), and f_{CLIFF} is the focal length of the original camera. See Fig. 1 for the illustration.

A weak-perspective projection can be regarded as an orthogonal projection followed by a perspective projection [9]. As shown in Fig. 1, the human body is first projected (parallel to the $Z^{'}$ axis) onto the virtual plane $Z^{'} = t_Z^{full}$, and then onto the image plane $Z^{'} = f_{CLIFF}$ by a perspective projection. A T-pose human body of the mean shape is about $1.8m \times 1.8m$ (m denoting meters). We enclose it with a slightly enlarged box B of size $2m \times 2m$, and align the center of B at the root of the human body (the green point R in Fig. 1). B is projected to be a square region of size $b \cdot s$ in the image. Since the two triangles $\triangle$OGH and $\triangle$OPQ in Fig. 1 (in blue) are similar, we have:

$$\frac{2}{t_Z^{full}} = \frac{b \cdot s}{f_{CLIFF}}, \quad t_Z^{full} = \frac{2 \cdot f_{CLIFF}}{b \cdot s}. \tag{7.2}$$

Note that here b and f_{CLIFF} are in pixels, and t_Z^{full} is in meters.

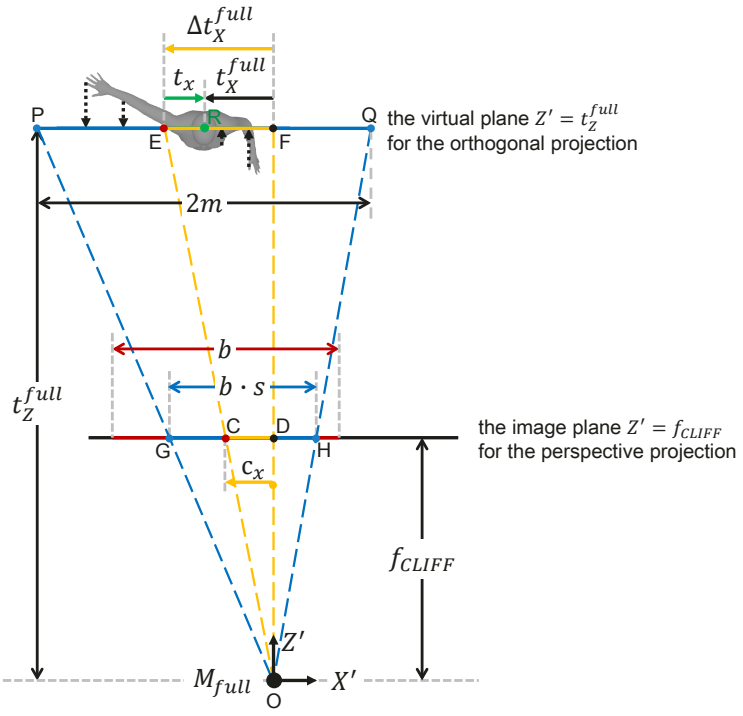

Fig. 1. The transformation from weak-perspective projection to perspective projection (bird's eye view). A weak-perspective projection can be regarded as an orthogonal projection followed by a perspective projection. Best viewed in color.

Let D (the projection of F) be the image center, and C (the projection of E) be the crop (i.e., detection result) center. Then the root translation of the human body along the $X^{'}$ axis is calculated by:

$$t_X^{full} = t_x + \Delta t_X^{full}, \tag{7.3}$$

where Δt_X^{full} is the $X^{'}$ coordinate of point E. Since the two triangles $\triangle$OCD and $\triangle$OEF in Fig. 1 (in yellow) are similar, we have:

$$\frac{\Delta t_X^{full}}{t_Z^{full}} = \frac{c_x}{f_{CLIFF}}. \tag{7.4}$$

Combining Equations 7.2 and 7.4, we obtain:

$$\Delta t_X^{full} = \frac{2 \cdot c_x}{b \cdot s}. \tag{7.5}$$

Similarly, it also holds for the root translation along the $Y^{'}$ axis:

$$\begin{aligned} t_Y^{full} &= t_y + \Delta t_Y^{full} \\ &= t_y + \frac{2 \cdot c_y}{b \cdot s}. \end{aligned} \tag{7.6}$$

Fig. 2. Qualitative results on the AGORA test samples. CLIFF works well in crowded and occluded scenes, even when large body parts are missing in the BBoxes.

The orthogonal projection in the weak-perspective projection omits the $Z^{'}$ coordinate discrepancy inside the human body, which assumes the human body is far from the camera whose focal length is unrealistically large (corresponding to a very small field-of-view) [5,8]. This is not true for many cases. Thus we use the perspective projection with an appropriate focal length to calculate the 2D reprojection loss, because this is how the original image is captured. However, we still let the model predict the weak-perspective projection parameters, since for most cases, $t_x \in [-1, 1]$, $t_y \in [-1, 1]$, $s \in [0, 1]$, meaning that they have the normalization property, which makes them suitable to be the CNN predictions.

2 Impact of the BBox Quality

The BBox quality is important to our method, just like other top-down methods. However, taking the BBox information as the additional input does not make our method rely more on the BBox quality. As demonstrated in the AGORA evaluation, we use the BBox predicted by Mask R-CNN which is trained on COCO without finetuning on AGORA; yet CLIFF still reaches the first place on the leaderboard, outperforming other top-down and bottom-up methods by large margins. Note that AGORA contains a lot of crowded and severely occluded scenes, as shown in Fig. 2. CLIFF is robust to inaccurate BBox detection, mainly thanks to the data augmentation such as random scaling and cropping.

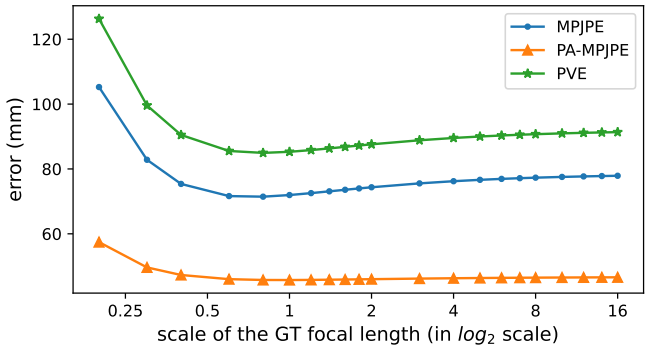

Fig. 3. Impact of the focal length on estimation errors.

Table 1. Comparison between CLIFF and video-based methods on 3DPW

Annotator	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Accel Error ↓
HMMR [6]	116.5	72.6	-	14.3
TCMR [2]	86.5	52.7	102.9	**7.1**
VIBE [7]	82.7	51.9	99.1	23.4
MAED [10]	79.1	45.7	92.6	17.6
CLIFF (Res-50)	72.0	45.7	85.3	24.7
CLIFF (Res-50) w/ OneEuro	74.0	46.2	87.6	10.6
CLIFF (HR-W48)	**69.0**	**43.0**	**81.2**	20.5
CLIFF (HR-W48) w/ OneEuro	70.1	43.1	82.3	11.3

3 Impact of the Focal Length as Part of the Input

We conduct this experiment on the 3DPW test set by perturbing the focal length from its GT value f_{GT}. As shown in Fig. 3, CLIFF is robust (with less than 5% error increase) when the estimated focal length is in $[0.4f_{GT}, 3f_{GT}]$. The estimation $f_{CLIFF} = \sqrt{w^2 + h^2}$ is within this range for most cases (except for super telephotos). Moreover, in practical applications, f_{GT} is often known, making the performance guaranteed.

4 Smoothness Comparison with Video-Based Methods

We can apply CLIFF to a video frame by frame, and perform temporal smoothing to reduce jitter, such as OneEuro filtering [1]. Video-based methods [6,2,7,10] usually make temporally smooth 3D predictions, which is their advantage over frame-based methods. However, they cost much computation by processing additional adjacent frames. Here we compare CLIFF with these video-based methods, especially on the smoothness evaluation, as shown in Table 1. The metric for evaluating temporal smoothness is acceleration error, which measures the average difference between ground truth 3D acceleration and the predicted 3D acceleration of each joint in mm/s^2. CLIFF, as a frame-based method, achieves

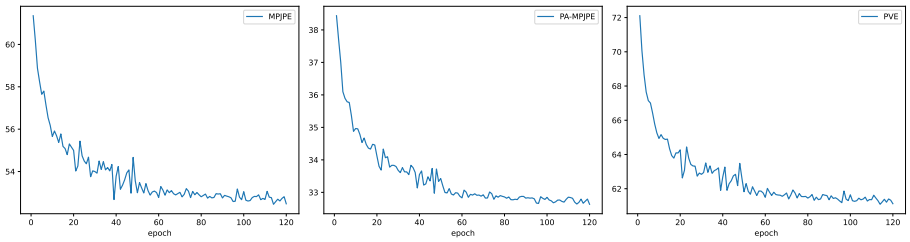

Fig. 4. The curves of training the CLIFF annotator on the 3DPW test data. The evaluation errors do not diverge even for long training, which means there is no need for the CLIFF annotator to choose a generic stopping criterion carefully.

Table 2. Ablation study of the CLIFF annotator on 3DPW

Annotator	MPJPE ↓	PA-MPJPE ↓	PVE ↓
ProHMR	-	52.4	-
BOA	77.2	49.5	-
EFT	-	49.3	-
DynaBOA	65.5	40.4	82.0
Pose2Mesh	65.1	34.6	-
HMR-based (Ours)	63.6	38.7	72.6
CLIFF-based (Ours)	**52.8**	**32.8**	**61.5**

comparable smoothness performance with video-based methods. With the additional OneEuro filtering as post-processing which costs negligible extra computation, the smoothness performance is improved significantly with slightly larger pose errors, which are still much smaller than those of the competitors.

5 CLIFF Annotator Training

In Fig. 4, we show the evaluation error curves of training the CLIFF annotator on the 3DPW test data. The learning rate starts from $5 \times e^{-5}$, and is reduced by a factor of 10 at the 45th epoch. We can obtain a fine model before the 60th epoch, and the evaluation errors do not diverge even for longer training (120 epochs in total), and may decrease for a better performance. It means that the CLIFF annotator is robust in the optimization, because the proposed priors prevent the annotator from overfitting to the 2D keypoints and from producing implausible poses. Consequently, there is no need for our annotator to choose a generic stopping criterion carefully, which is a serious problem for EFT [4].

6 Ablation Study of the CLIFF Annotator

We implement the proposed pseudo-GT annotator based on HMR, and compare it to the CLIFF-based one on 3DPW. As shown in Table 2, the errors increase

when switching the base model from CLIFF to HMR, but the HMR-based annotator is still better than other SOTA methods. Note that Pose2Mesh [3] as a model-free method produces only 3D vertices but no SMPL parameters.

7 More Qualitative Pseudo-GT Results

In Fig. 5, we show additional qualitative results in the CLIFF annotator experiments. We test the pretrained annotator on the target images to get predictions as the explicit prior, which may not be accurate but usually plausible. The final pseudo-GT achieves better pixel alignment, and maintains the plausibility with the help of the proposed priors.

References

1. Casiez, G., Roussel, N., Vogel, D.: 1€ filter: A simple speed-based low-pass filter for noisy input in interactive systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2012)
2. 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)
3. 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)
4. 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)
5. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
6. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR (2019)
7. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: CVPR (2020)
8. 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)
9. Shimshoni, I., Basri, R., Rivlin, E.: A geometric interpretation of weak-perspective motion. PAMI (1999)
10. 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)

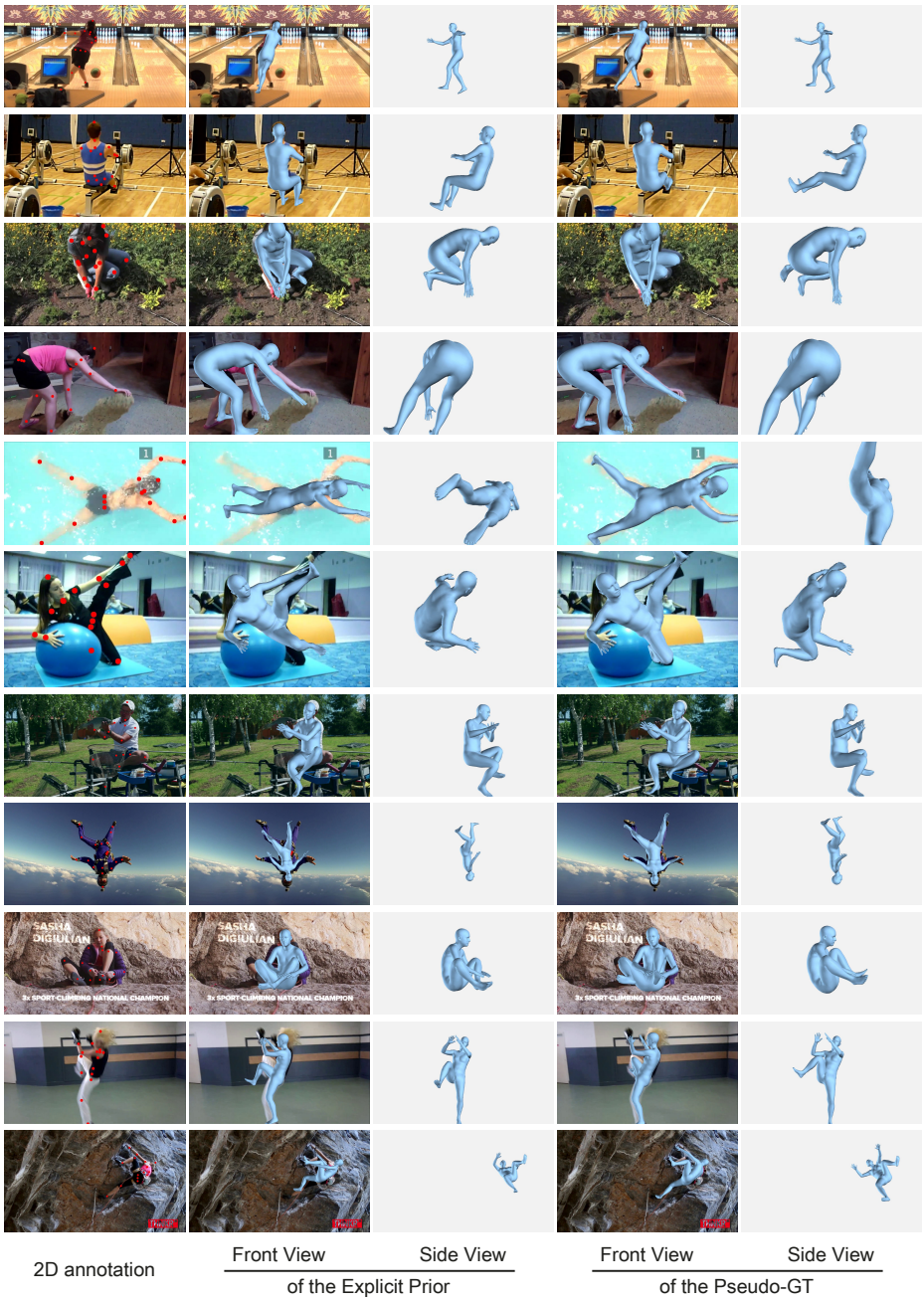

Fig. 5. More pseudo-GT samples from the CLIFF annotator. From left to right: 2D annotation, front view of the explicit prior, side view of the explicit prior, front view of the pseudo-GT, and side view of the pseudo-GT.