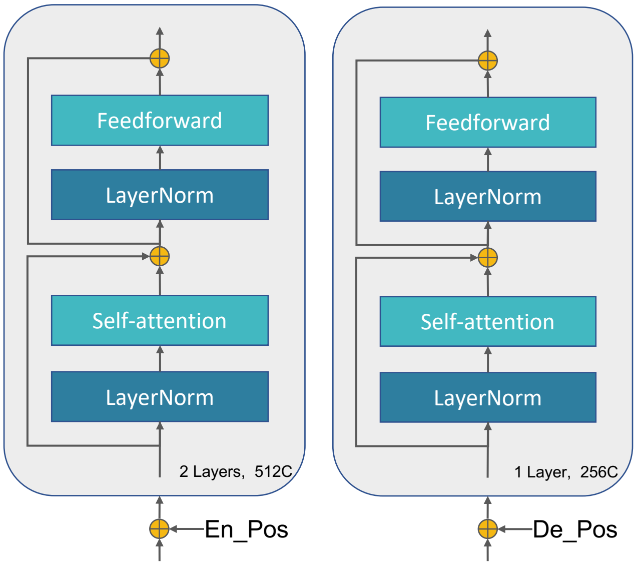
Supplementary Material *for* Global-to-Local Modeling for Video-based 3D Human Pose and Shape Estimation

Xiaolong Shen[1,2*], Zongxin Yang[1], Xiaohan Wang[1], Jianxin Ma[2], Chang Zhou[2], Yi Yang[1]
[1] ReLER, CCAI, Zhejiang University [2] DAMO Academy, Alibaba Group



(a) Global encoder.

(b) Global decoder.

Figure 1. Details of the global transformer.

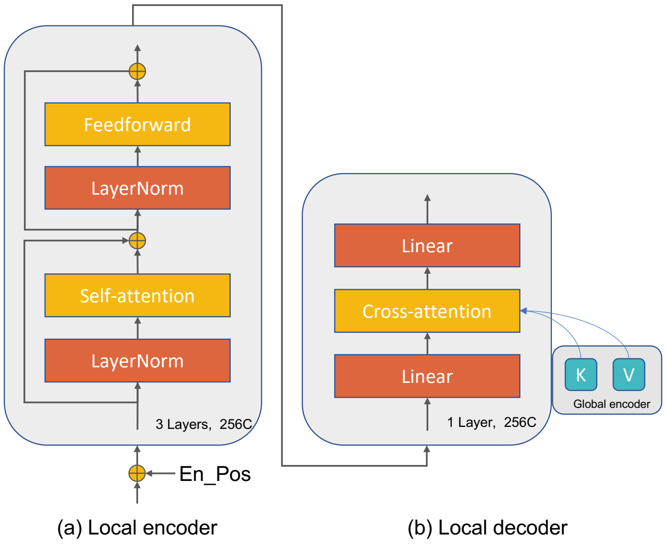


Figure 2. Details of the local transformer.

	PA-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓
w/o Detach	50.9	81.2	96.4	6.6
w/ Detach	**50.6**	**80.7**	**96.3**	**6.6**

Table 1. Gradient detachment

A. Datasets

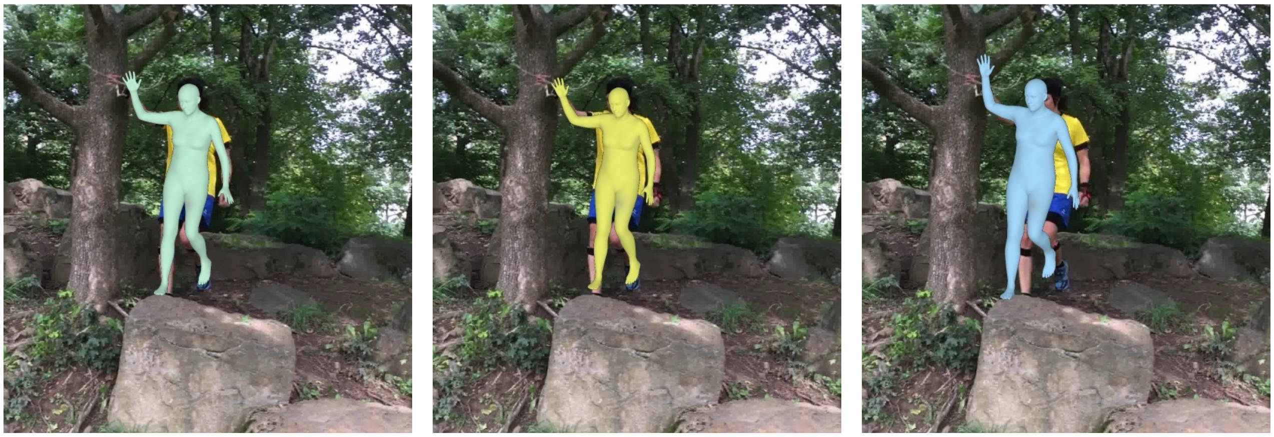
3DPW. 3DPW [8] is a challenging in-the-wild consisting of 60 videos, which are captured by a phone at 30 fps. Moreover, IMU sensors are utilized to obtain the near ground-truth SMPL parameters, *i.e.*, pose and shape. We utilize the official split to train and test our model, where the training, validation, and test sets are comprised of 24, 12, and 24 videos, respectively. For evaluation, we report MPVPE on 3DPW because it has ground-truth shape annotation.

Human3.6M. Human3.6M [4] is a large-scale dataset collected under a controlled indoor environment and includes 3.6M video frames. Folloing [3, 9], we train the model on 5 subjects (i.e., S1, S5, S6, S7, and S8) and test it on 2 subjects (i.e., S9 and S11). We set the frame rate of the dataset to 25 fps for training and testing.

MPI-INF-3DHP. MPI-INF-3DHP [5]] is a complex dataset captured at indoor and outdoor scenes with a markerless motion capture system. The 3D human pose annotations are computed by the multiview method. The training and testing sets are comprised of 8 and 6 subjects, respectively. Each subject has 16 videos captured in the indoor or outdoor environment. The total video frames are 1.3M. Following previous works [3, 9], we utilize the official training and testing split.

InstaVariety. InstaVariety is a 2D human pose dataset collected from Instagram. It consists of 28K videos, and the video length is an average of 6 seconds. The 2D annotation is generated from Openpose [2]. Following [3, 5], we use this dataset for training.

PoseTrack. PoseTrack [1] is also a 2D human dataset for multi-person pose estimation and tracking, which consists of 1.3K videos. Following [3], we use 792 videos for training.



(a) Ours (b) MPS-Net [9] (c) TCMR [3]

Figure 3. Comparison with other methods [3, 9]. **Please use Adobe Acrobat to view it.**

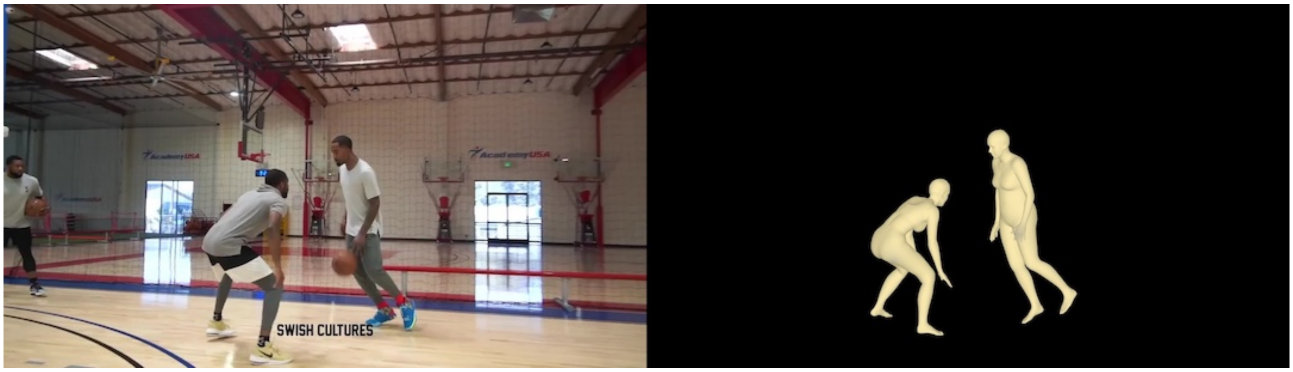


Figure 4. An Example of internet video. We sample every ten frames. **Please use Adobe Acrobat to view it.**

	PA-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓
Fix	52.1	83.4	99.5	6.4
Learnable	**50.6**	**80.7**	**96.3**	**6.6**

Table 2. Position embedding

B. Model details

Global transformer. The model details are shown in Figure 1. We utilize two layers of the encoder block with 512 dimensions. In the global decoder, we only apply one layer of the decoder block with 256 dimensions. The position embedding is learnable.

Local transformer. As shown in Figure 2, the encoder block is similar to the global encoder. We set three layers of the encoder block with 256 channel sizes. In addition, we employ cross-attention to the decoder and set the layer to one. The channel size is the same as the encoder.

C. Effect of gradient detachment

Table 1 shows the effect of gradient detachment. When we do not backward propagate the path of global estimation to HSCR, GLoT achieves the best performance. It is intuitively reasonable that fixing one is easier for optimization. In addition, w/o $Detach$ also obtains good results.

D. Effect of position embedding

In Table 2, we report the results of the different types of position embeddings. The learnable embedding obtains the best performance.

E. Inference time (GPU: V100) and MACs

Model	MACs (M)	Time (ms)	PA-MPJPE
TCMR	861.8	**11.7**	52.7
MPS-Net	318.4	17.6	52.1
Ours w/ Residual	**287.9**	13.0	51.5
Ours w/ HSCR	288.1	16.2	**50.6**

Table 3. Inference time and MACs.

We provide the results of inference time and MACs in Table 3. Our model achieves the lowest MACs. For inference time, our model (w/ HSCR) is slower than TCMR [3] but faster than the previous SOTA method MPS-Net [9]. We analyze that the reason for slower than TCMR is the self-attention mechanism used in our model and MPS-Net. Al-

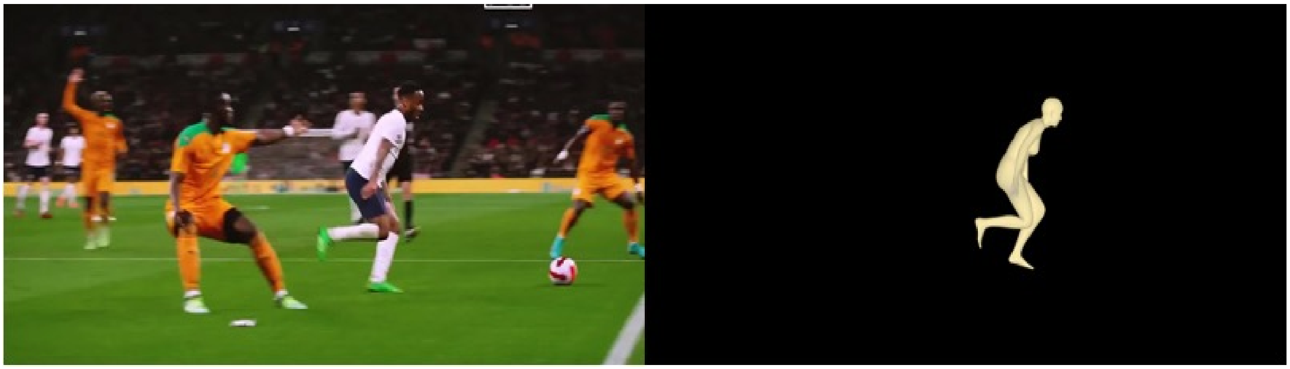


Figure 5. An Example of internet video. **Please use Adobe Acrobat to view it.**

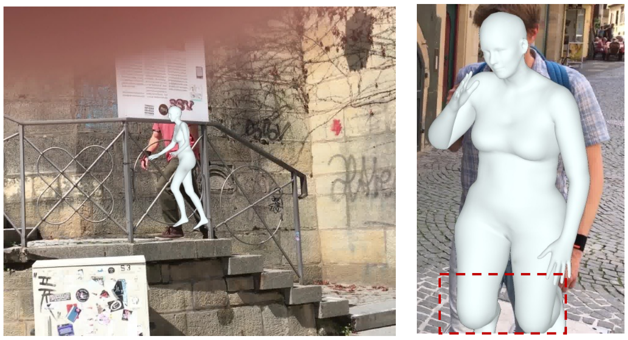


Figure 6. Some failure cases.

though our model (w/ HSCR) is slower than TCMR by 4.5 ms, it shows a significant improvement in PA-MPJPE. Besides, we provide the inference time of our model (w/ Residual) for comparing the time consumption of the HSCR. It is worth noting that our model (w/ Residual) reduces 1.2 PA-MPJPE with a time consumption of only 1.3 ms compared with TCMR.

F. Input length of the global encoder

length	PA-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓
32	51.2	82.0	98.3	6.7
24	51.2	82.7	98.5	6.7
16	**50.6**	**80.7**	**96.3**	**6.6**

Table 4. Input length of the global encoder.

Although the 16-frame input length is commonly used in this task, we consider that exploring more input lengths is valuable. In Table 4, we supply the study of longer input lengths, 24 and 32. The 16-frame setting achieves the best results. A possible reason is that our lightweight global encoder can not sufficiently model longer temporal relations.

G. More qualitative results

We show the comparison results with other methods in Figure 3. We observe (1) The results of MPS-Net [9] suffer from insufficient local details. (2) The results of TCMR [3] do not capture the actual human global location of the frames. Figure 4 and 5 are multi-person internet videos, we first use a multi-object tracker to process videos and then utilize our method for each tracked person, following the previous methods [3, 9].

H. Failure cases

As shown in Figure 6, we provide some failure cases, mainly including occlusion. We divide the occlusion into two types, *i.e.*, object occlusion (Left Figure) and truncation of the frame (Right Figure, some joints are outside of the frame). We consider that these cases are caused by long-term occlusion, which means the input frames are all occluded by the object or truncated by the camera, leading to failures in temporal modeling.

I. Future works

We plan to use this framework in similar tasks, *i.e.*, hand pose and shape estimation [6, 10]. This task will provide a more robust hand representation for downstream tasks, *e.g.*, sign language recognition [7]. Moreover, we believe that exploring multi-person interaction in a video would be a good idea. While there are some methods in image-based tasks to deal with occlusion problems caused by multiple people, video-based methods in this area are still unexplored.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. *CVPR*, 2017. 1

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *PAMI*, 2018. 1

[3] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consis-

tent 3d human pose and shape from a video. *CVPR*, 2020. 1, 2, 3

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014. 1

[5] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. *international conference on 3d vision*, 2016. 1

[6] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *TOG*, 2017. 3

[7] Xiaolong Shen, Zhedong Zheng, and Yi Yang. Stepnet: Spatial-temporal part-aware network for sign language recognition. In *arXiv*, 2022. 3

[8] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate {3D} human pose in the wild using {IMUs} and a moving camera. *ECCV*, 2018. 1

[9] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *CVPR*, pages 13211–13220, 2022. 1, 2, 3

[10] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *CVPR*, pages 813–822, 2019. 3