

# A UNIFIED PRETRAINING FRAMEWORK FOR HUMAN MOTION ANALYSIS

**Anonymous authors**

Paper under double-blind review

We provide additional experiment results which are not elaborated in our manuscript due to space limits.

## 1 3D POSE ESTIMATION

Table 1: **Quantitative comparison of 3D human pose estimation using 2D GT keypoint sequences as input.** Numbers are MPJPE (mm) on Human3.6M.  $T$  denotes the clip length used by the method. We select the best results reported by each work.

MPJPE	$T$	Dire.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Martinez et al. (2017) ICCV'17	1	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Ci et al. (2019) ICCV'19	1	36.3	38.8	29.7	37.8	34.6	42.5	39.8	32.5	36.2	39.5	34.4	38.4	38.2	31.3	34.2	36.3
Xu & Takano (2021) CVPR'21	1	35.8	38.1	31.0	35.3	35.8	43.2	37.3	31.7	38.4	45.5	35.4	36.7	36.8	27.9	30.7	35.8
Wang et al. (2020) ECCV'20	96	23.0	25.7	22.8	22.6	24.1	30.6	24.9	24.5	31.1	35.0	25.6	24.3	25.1	19.8	18.4	25.6
Zheng et al. (2021) ICCV'21	81	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
Li et al. (2022) CVPR'22	351	27.7	32.1	29.1	28.9	30.0	33.9	33.0	31.2	37.0	39.3	30.0	31.0	29.4	22.2	23.0	30.5
Zhang et al. (2022) CVPR'22	243	21.6	22.0	20.4	21.0	20.8	24.3	24.7	21.9	26.9	24.9	21.2	21.5	20.8	14.7	15.6	21.6
Ours (Scratch)	243	18.8	<b>19.5</b>	19.2	18.0	19.2	22.4	21.2	21.1	24.4	26.5	20.0	19.3	18.7	11.6	13.0	19.5
Ours (Pretrained)	243	<b>18.2</b>	19.8	<b>18.4</b>	<b>16.6</b>	<b>19.1</b>	<b>20.7</b>	<b>20.9</b>	<b>19.5</b>	<b>23.6</b>	<b>24.0</b>	<b>19.4</b>	<b>17.9</b>	<b>17.1</b>	<b>10.9</b>	<b>11.8</b>	<b>18.5</b>

We compare the model performance when using 2D GT keypoint sequences as input. This experiment is free from the influence of different 2D detectors and directly evaluates the models' ability of 2D-to-3D lifting. As shown in Table 1, Our models outperform all the previous approaches.

## 2 ONE-SHOT ACTION RECOGNITION

Figure 1 shows the learned action representation for the novel classes. We can see that meaningful clusters emerge although the model has never seen these actions before. We perform one-shot action recognition using 1-nearest-neighbor with the exemplars and achieve 67.4% top-1 accuracy.

## 3 QUALITATIVE COMPARISON

We visualize the predictions of our models and compare them with baseline approaches. Figure 2 shows the comparison on 3D human pose estimation with VideoPose3D (Pavlo et al., 2019). Figure 3 shows the comparison on mesh recovery with VIBE (Kocabas et al., 2020). Our models provide better reconstruction results even on very challenging motions like *ballet* or *fencing*.

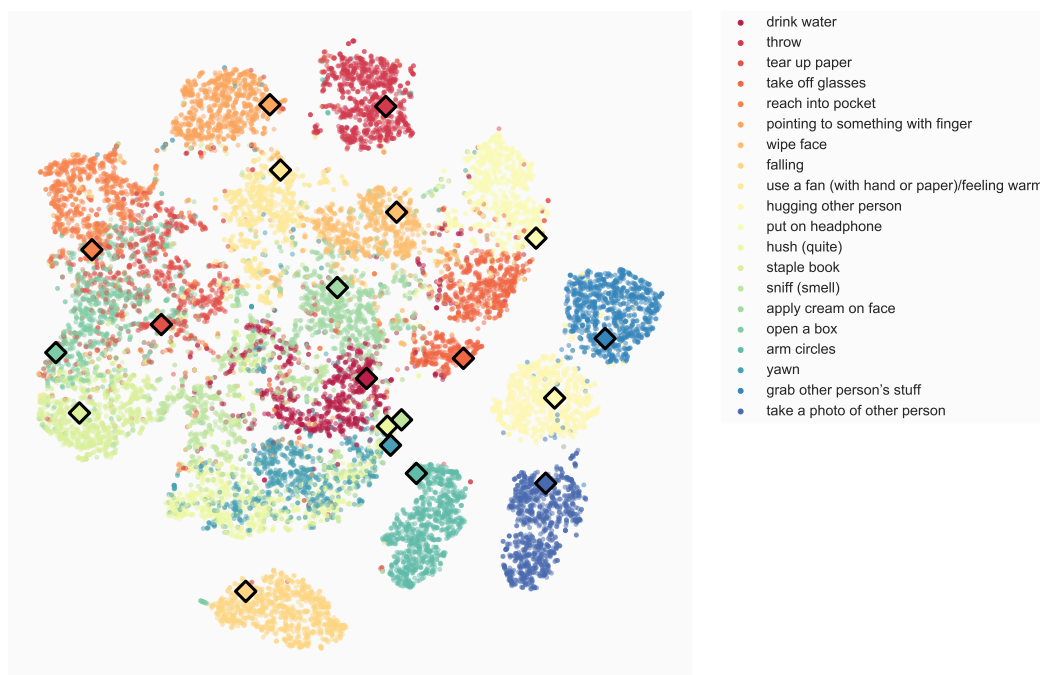


Figure 1: **Visualization of action representations for one-shot recognition.** We show the action representations of the 20 novel classes after training on the other 100 auxiliary classes with contrastive learning. We compute the pairwise cosine distance and apply t-SNE. Diamonds indicate the labeled exemplars for one-shot recognition.

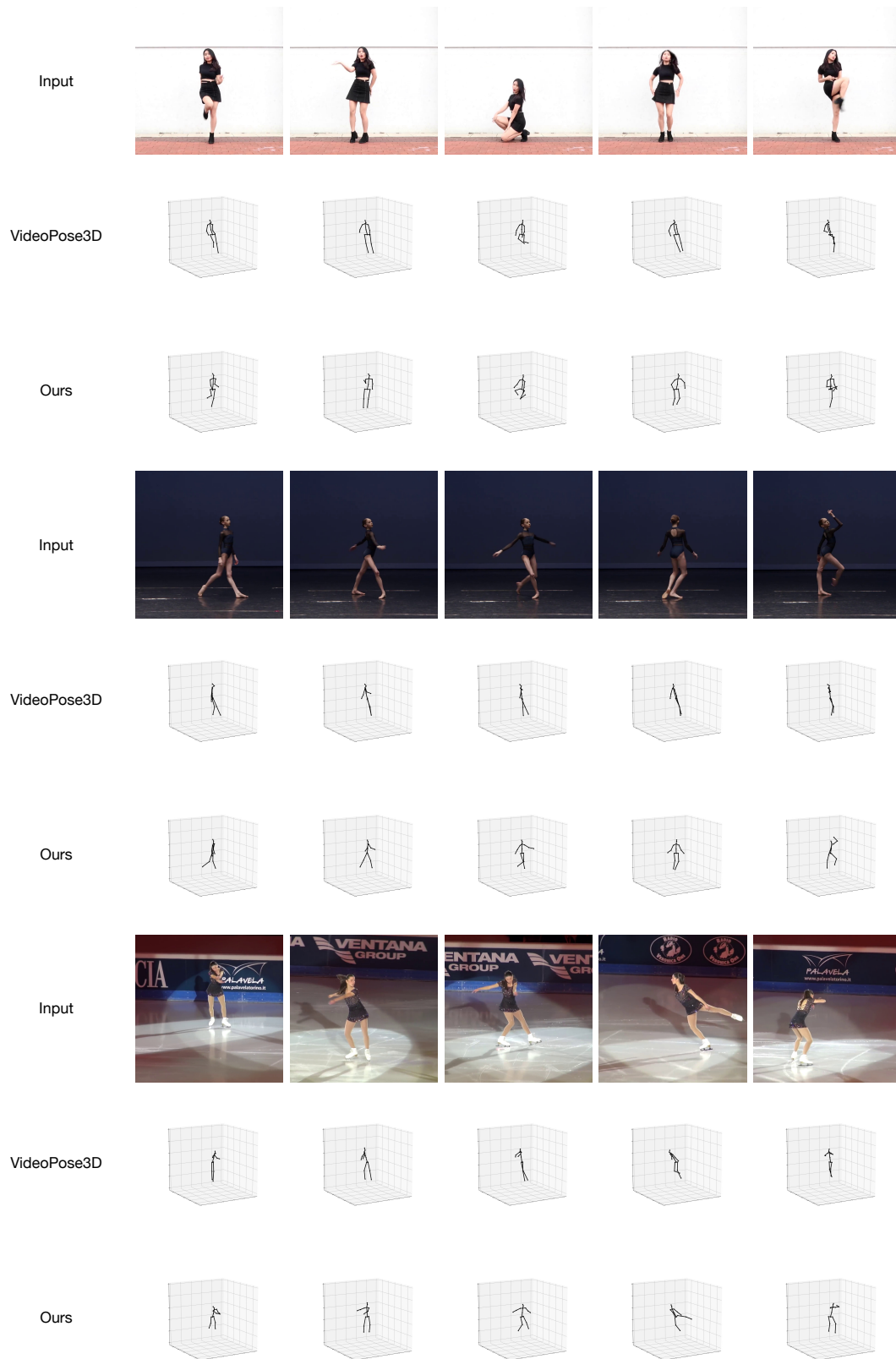


Figure 2: Qualitative comparison of 3D pose estimation on in-the-wild videos.

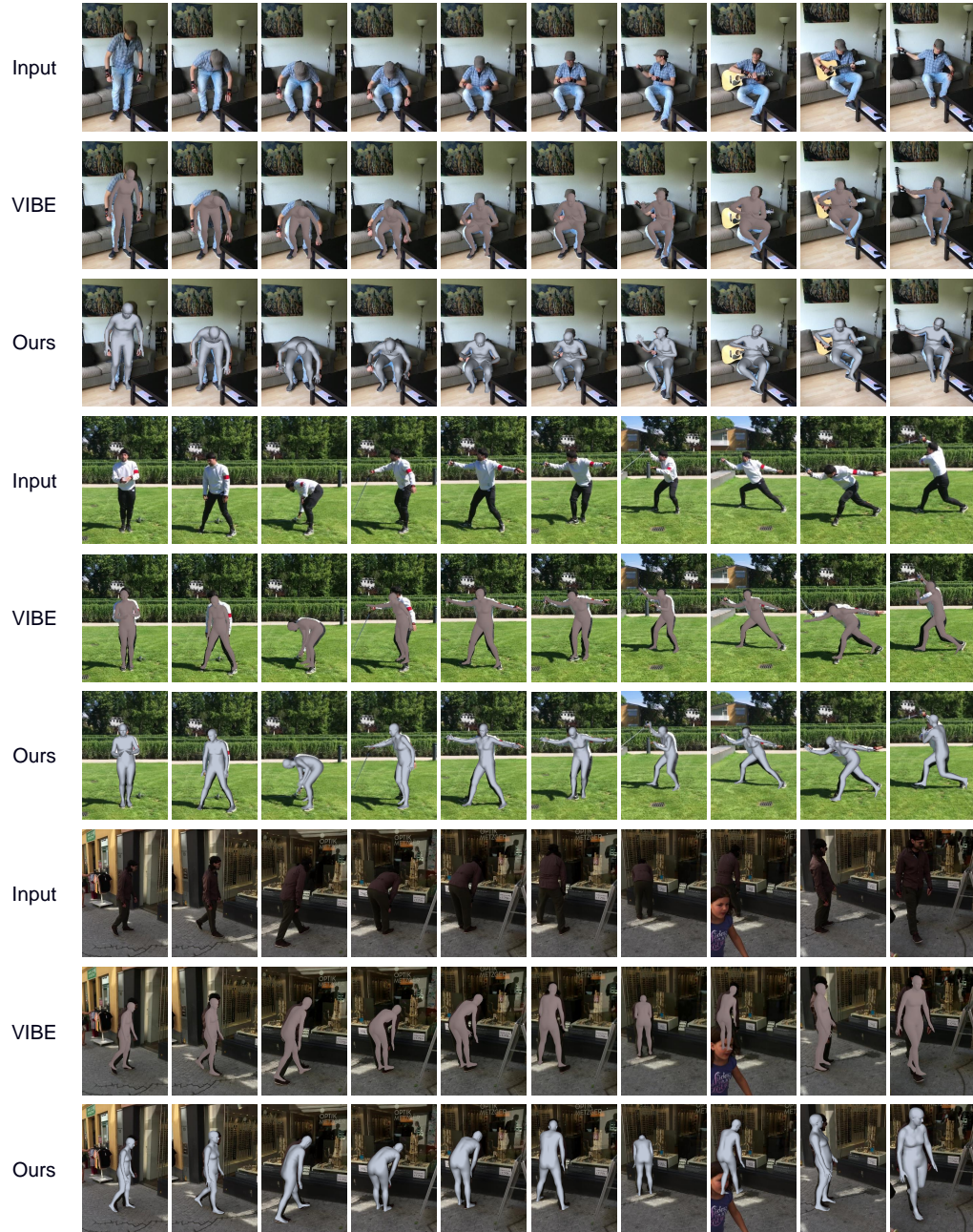


Figure 3: Qualitative comparison of mesh recovery on 3DPW test set.

## REFERENCES

- Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *ICCV*, pp. 2262–2271, 2019.
- Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pp. 5253–5263, 2020.
- Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13147–13156, 2022.
- Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, pp. 2640–2649, 2017.
- Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, pp. 7753–7762, 2019.
- Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision*, pp. 764–780. Springer, 2020.
- Tianhan Xu and Wataru Takano. Graph stacked hourglass networks for 3d human pose estimation. In *CVPR*, 2021.
- Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *CVPR*, 2022.
- Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. *ICCV*, 2021.