A Training Method For VideoPose3D With Ideology of Action Recognition

Jack / Hao Bai

Haob.19@intl.zju.edu.cn

ZJU-UIUC Institute

CONF
SPML

NOV 2021

What's based on?

- **VideoPose3D**: Pavllo, D., Feichtenhofer, C., Grangier, D., & Auli, M. (2019). 3d human **pose estimation** in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7753-7762).

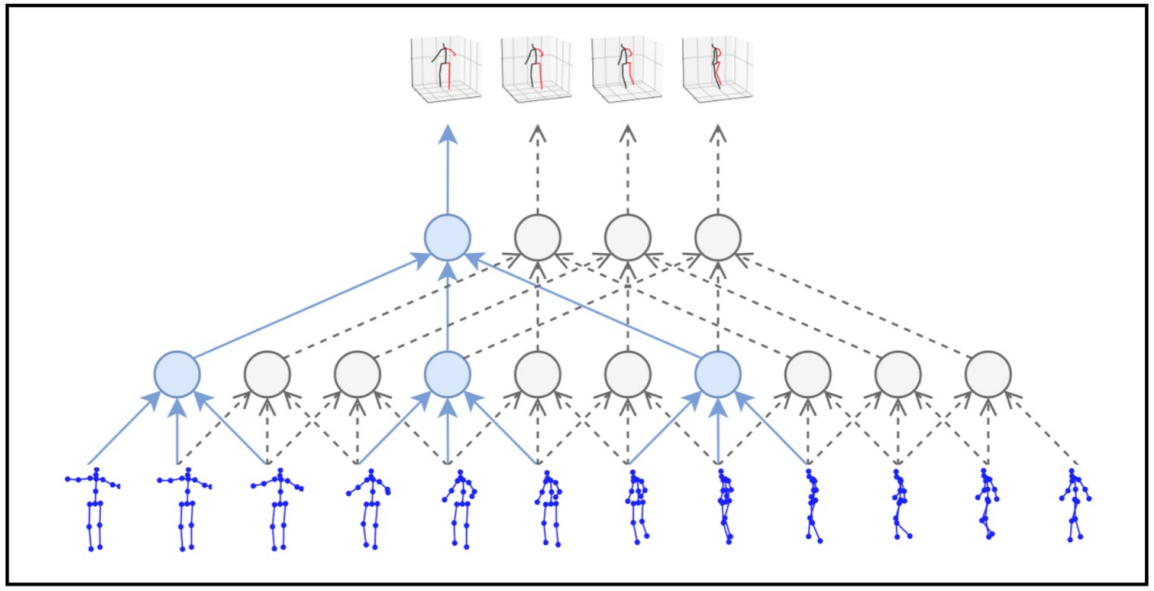

Fig. 1. The temporal convolutional model by [11].

Problem

- FCN (Fully Convolutional Network) + TCN (Temporal Convolutional Network) is TIME-CONSUMING.

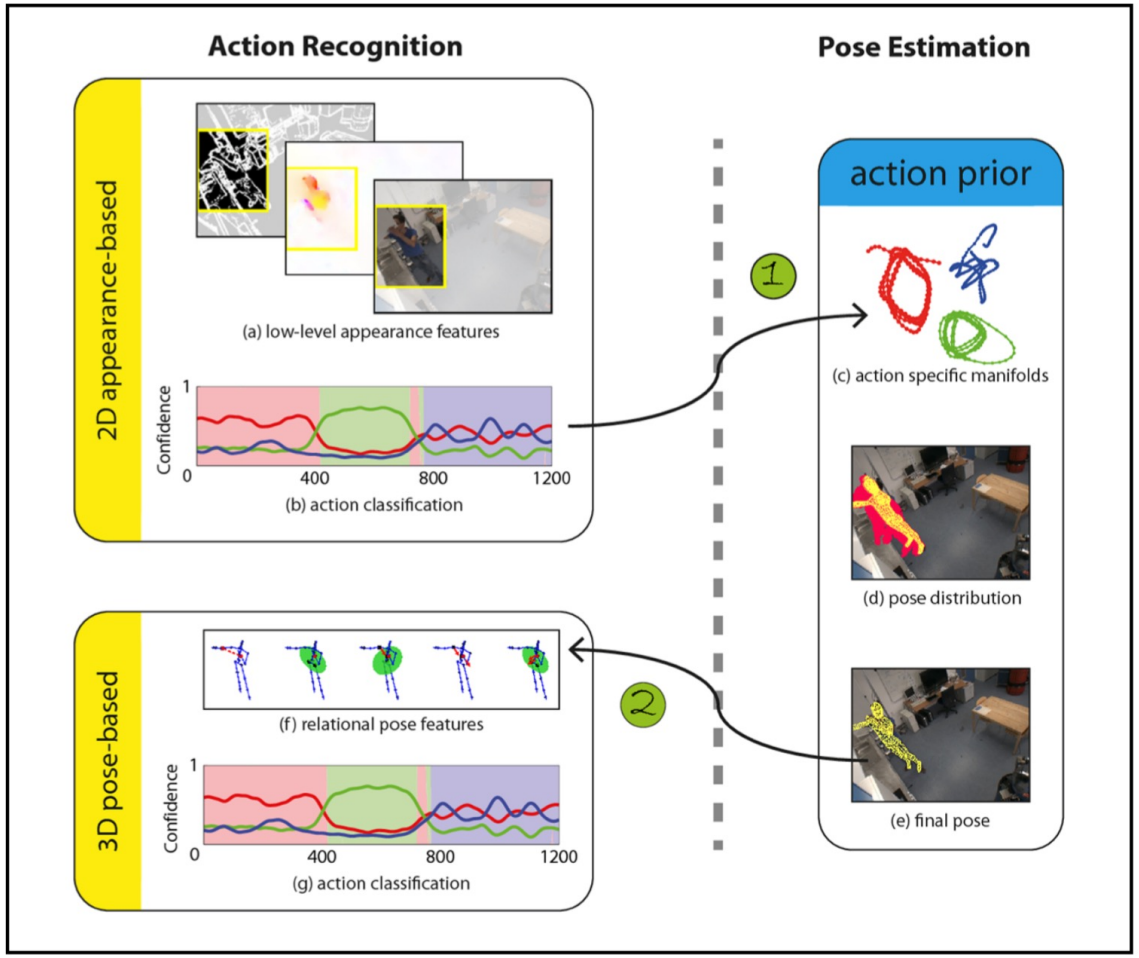

Fig. 2. Coupled Action Recognition and Pose Estimation model by [3].

Inspiration

- Yao, Angela, et al. "Does human action recognition benefit from pose estimation?"." Proceedings of the 22nd British machine vision conference-BMVC 2011. BMV press, 2011.
- Pose estimation can help with action recognition.
- Vice Versa?

The Approach

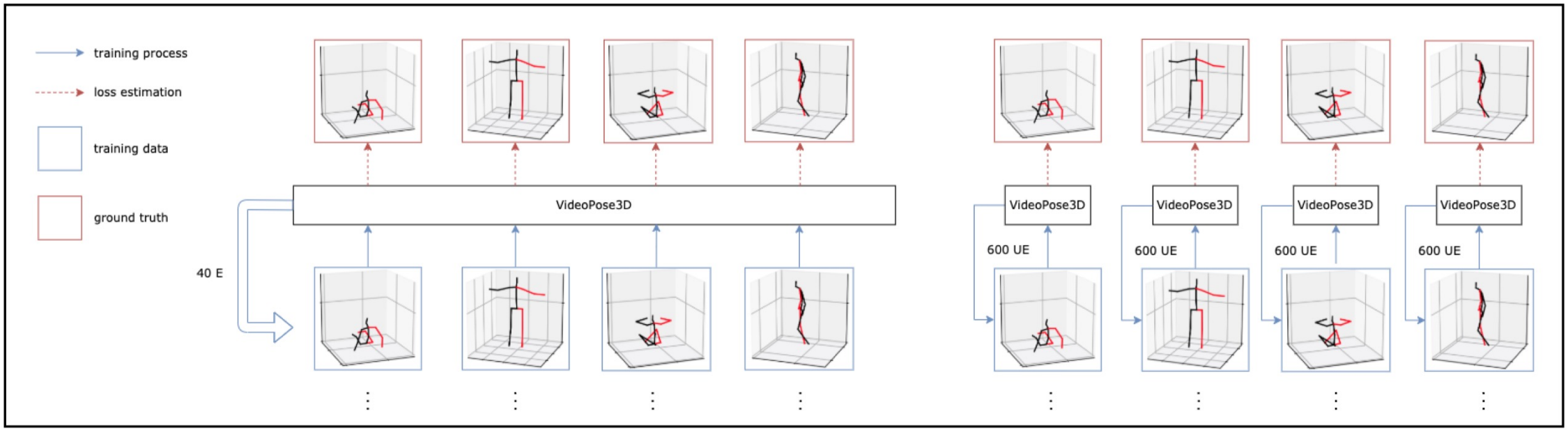

Fig. 3. Comparison of models between our model and *VideoPose3D* for common pose-estimation tasks

Hao Bai(∗). "A Training Method For VideoPose3D with Ideology of Action Recognition." 2021 The International Conference on Signal Processing and Machine Learning (CONF-SPML 2021).

Control the Variables

Common Problem

$$t_0 = t_{unit} \cdot n_{ac} \quad (1)$$

$$\frac{t_0}{t_{unit}} = \left\lfloor \frac{\sum_{i \in actions} f(i)}{f(action)} \right\rfloor \quad (2)$$

Action-based Problem

$$\sum_{i \in actions} n_{VP}(i) = n_{AB}(action) \quad (3)$$

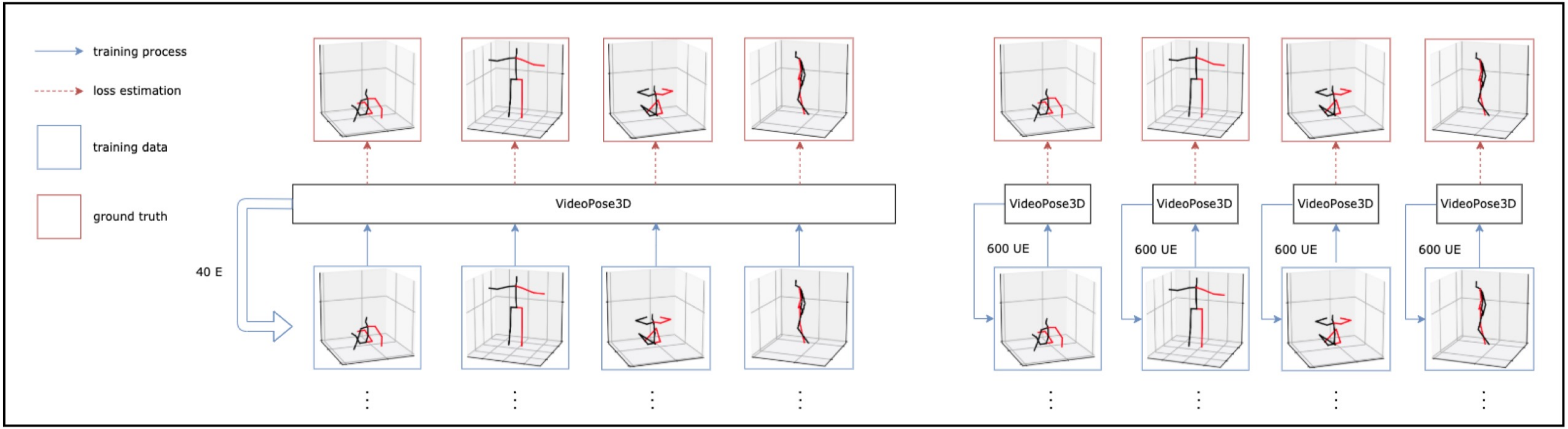

Fig. 3. Comparison of models between our model and *VideoPose3D* for common pose-estimation tasks

How about the results?

Hao Bai(*). "A Training Method For VideoPose3D with Ideology of Action Recognition." 2021 The International Conference on Signal Processing and Machine Learning (CONF-SPML 2021).

TABLE I. PROTOCOL I. MPJPE ERROR.

Arguments	Model	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WkD.	Walk	WkT.	Avg
F=1, UE=15	VideoPose3D	67.0	**65.8**	68.7	72.5	**74.2**	**87.2**	**65.5**	**73.1**	85.6	117	73.6	**70.8**	81.1	63.5	67.3	75.5
	Ours	**61.9**	69.2	**62.0**	**68.4**	75.7	88.2	74.5	76.9	**81.5**	**97.9**	**71.1**	80.9	**80.7**	**49.9**	**59.2**	**73.2**
F=27, UE=15	VideoPose3D	**54.5**	**61.4**	56.3	**58.6**	**61.3**	**68.6**	**57.6**	**60.6**	70.5	84.7	60.5	**59.1**	68.2	51.8	**53.1**	**61.8**
	Ours	70.5	71.9	**52.1**	61.4	69.0	82.6	70.6	82.1	**70.4**	**79.4**	**60.3**	79.5	**56.1**	**44.2**	56.2	67.1
F=243, UE=1200	Pavlakos [18]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
	Luvizon [8]	49.2	51.6	47.6	50.5	51.8	60.3	48.5	51.7	61.5	70.9	53.7	48.9	57.9	44.4	48.9	53.2
	VideoPose3D	**46.6**	**47.4**	45.2	**46.2**	**49.0**	**56.7**	**46.4**	**47.2**	**59.9**	68.2	**48.1**	**46.2**	49.4	32.9	**34.3**	**48.2**
	Ours	47.9	48.8	**45.1**	48.4	51.7	62.8	47.1	59.4	61.2	**64.7**	48.2	59.3	**46.6**	**31.4**	35.5	50.5

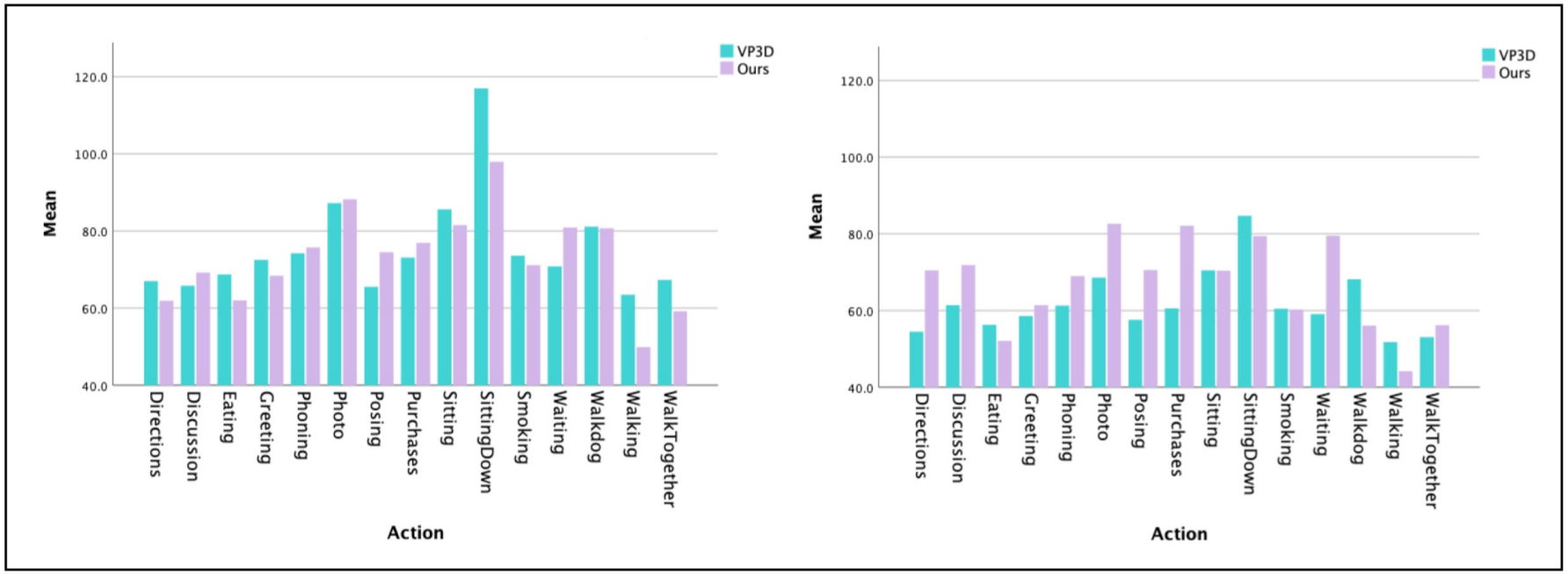

Fig. 4. Visualization for MPJPE error comparison. **Left**: error under 1F, 15UE. **Right**: error under 27F, 15UE.

TABLE II. PROTOCOL II. V-MPJPE ERROR.

Arguments	Model	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WkD.	Walk	WkT.	Avg
F=1, UE=15	VideoPose3D	**10.7**	**12.2**	10.2	12.4	11.1	**11.0**	10.5	12.3	**11.4**	**13.8**	11.0	**10.6**	12.7	12.9	12.4	11.7
	Ours	10.9	13.0	**9.59**	**11.8**	**10.3**	11.4	**10.1**	**10.5**	12.2	14.6	**10.9**	11.5	**12.5**	**12.6**	**12.2**	**11.6**
F=27, UE=15	VideoPose3D	3.84	4.02	3.14	4.42	3.31	3.58	**3.44**	3.93	3.14	4.21	3.36	**3.34**	4.68	4.51	4.08	3.80
	Ours	**3.78**	**4.28**	**3.01**	**4.27**	**3.14**	**3.50**	3.79	**3.65**	**2.84**	**3.90**	**3.05**	3.55	**4.37**	**3.77**	**3.49**	**3.63**
F=243, UE=1200	VideoPose3D	**3.01**	3.21	**2.33**	**3.55**	2.33	**2.83**	**2.77**	3.23	2.11	3.01	2.44	**2.43**	3.82	3.33	**2.86**	**2.49**
	Ours	3.14	**3.20**	2.48	3.62	**2.31**	2.91	3.04	**3.11**	**2.07**	**2.88**	**2.23**	2.73	**3.66**	**3.04**	2.97	3.08

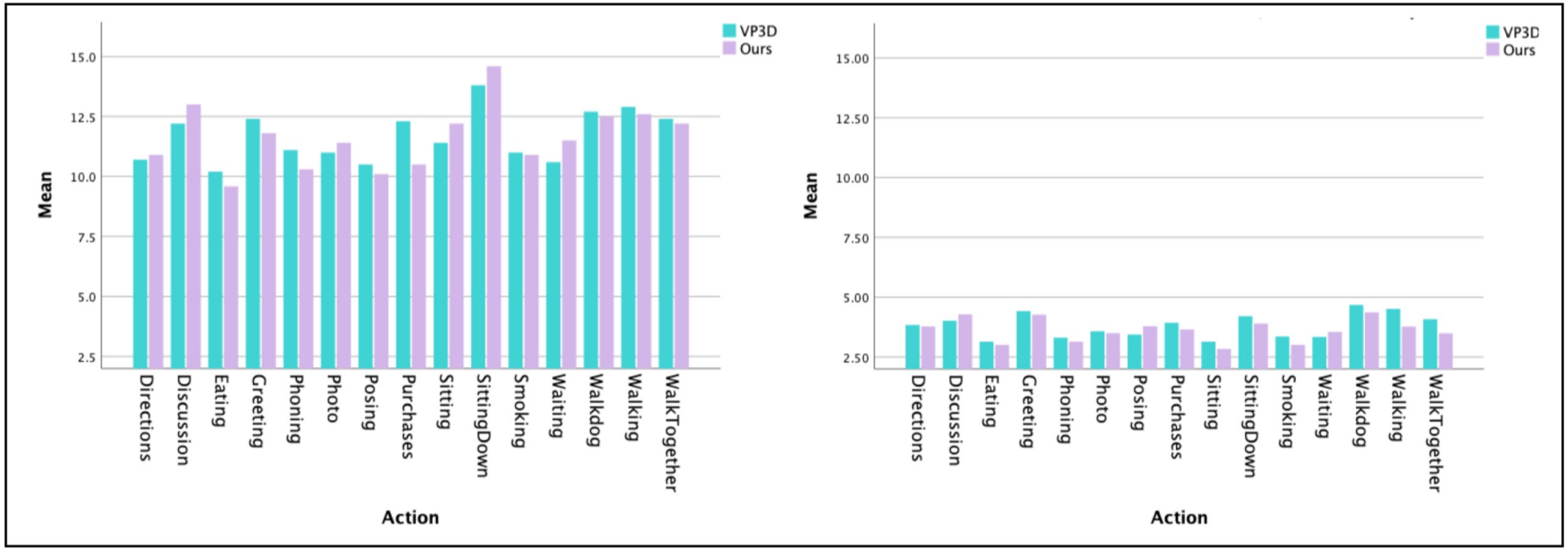

Fig. 5. Visualization for V-MPJPE error comparison. **Left**: error under 1F, 15UE. **Right**: error under 27F, 15UE.

Temporal Comparison

TABLE III. PROPERTIES COMPARISON WITH F=243, UE=1200.

Object	VP3D-Avg	Ours-Avg
TC (40 E)	84972 sec.	28848 sec.
MPJPE	50.27 mm	50.49 mm
Velocity-M	2.66 mm	3.02 mm
TC (80 E)	176976 sec.	56921 sec.
MPJPE	48.22 mm	50.54 mm
Velocity-M	2.49 mm	3.08 mm
MPJPE ϵ_0	60.46	
Velo-M ϵ_0	3.408	
MPJPE TPR	$6.92 \cdot 10^{-5}$	$17.4 \cdot 10^{-5}$
Velo-M TPR	$5.19 \cdot 10^{-6}$	$5.76 \cdot 10^{-6}$

TPR: Time-Precision Rate.
Higher TPR means higher performance.

Hao Bai(*). "A Training Method For VideoPose3D with Ideology of Action Recognition." 2021 The International Conference on Signal Processing and Machine Learning (CONF-SPML 2021).

TABLE IV. TEMPORAL COMPARISON WITH F=243, UE=1200

Epochs	VP3D-Avg	Ours-Avg	VP3D-Eat	Ours-Eat
1	59.11	57.52	54.22	51.14
2	56.26	52.06	52.21	48.72
3	61.73	51.28	51.17	47.14
4	54.24	51.41	49.22	45.28
5	51.17	51.11	49.74	45.17
6	50.16	50.11	51.12	46.28
7	50.28	50.37	50.71	45.82
8	50.19	50.25	50.23	47.14
9	50.24	50.34	49.17	47.22
10	50.14	50.94	49.82	46.32
30	49.56	50.56	47.71	45.64
40	50.27	50.49	46.14	45.18
50	48.17	50.29	45.61	45.12
80	48.22	50.54	45.22	45.14

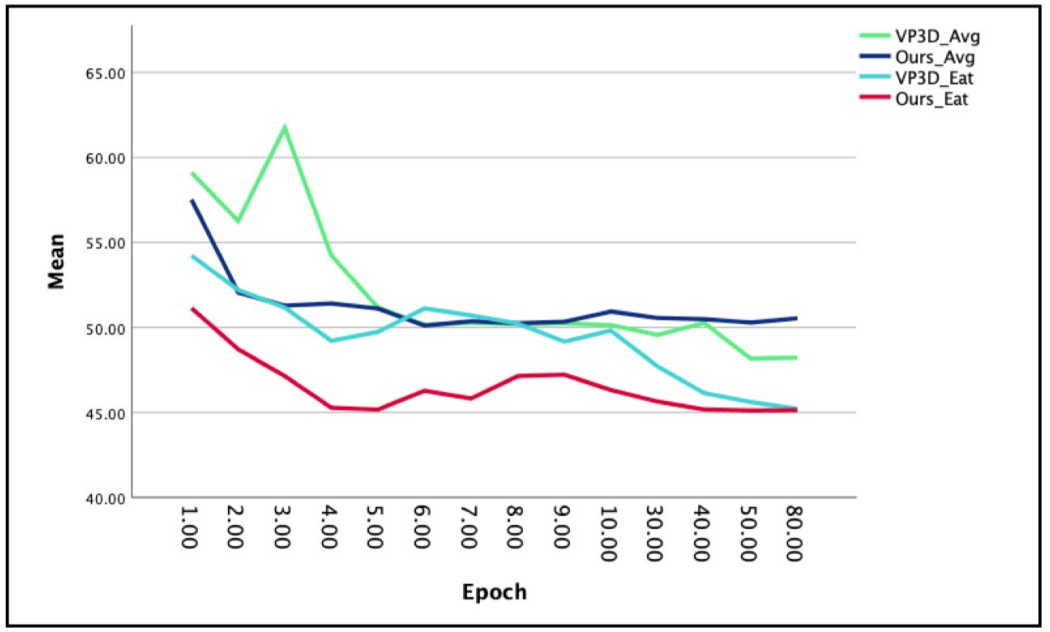

Fig. 6. Visualization of the temporal comparison with F=243, UE=1200.

Hao Bai(*). "A Training Method For VideoPose3D with Ideology of Action Recognition." 2021 The International Conference on Signal Processing and Machine Learning (CONF-SPML 2021).

Statistics and Visualization

General Thoughts

- Looking at apple then banana, you can estimate apple better, but slower than simply looking at apple.
- Looking at more things uses more time.

Hao Bai(∗). "A Training Method For VideoPose3D with Ideology of Action Recognition." 2021 The International Conference on Signal Processing and Machine Learning (CONF-SPML 2021).

REFERENCES

[1] K. Lee, I. Lee, and S. Lee, “Propagating lstm: 3d pose estimation based on joint interdependency,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 119–135.

[2] Jhuang, Hueihan, et al. "Towards understanding action recognition." *Proceedings of the IEEE international conference on computer vision*. 2013.

[3] Yao, Angela, et al. "Does human action recognition benefit from pose estimation?"." *Proceedings of the 22nd British machine vision conference-BMVC 2011*. BMV press, 2011.

[4] Yao, Angela, Juergen Gall, and Luc Van Gool. "Coupled action recognition and pose estimation from multiple views." *International journal of computer vision* 100.1 (2012): 16-37.

[5] Iqbal, Umar, Martin Garbade, and Juergen Gall. "Pose for action-action for pose." *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017.

[6] Gall, Juergen, Angela Yao, and Luc Van Gool. "2d action recognition serves 3d human pose estimation." *European Conference on Computer Vision*. Springer, Berlin, Heidelberg, 2010.

[7] Xiaohan Nie, Bruce, Caiming Xiong, and Song-Chun Zhu. "Joint action recognition and pose estimation from video." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[8] Luvizon, Diogo C., David Picard, and Hedi Tabia. "2d/3d pose estimation and action recognition using multitask deep learning." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[9] Li, Sijin, Weichen Zhang, and Antoni B. Chan. "Maximum-margin structured learning with deep networks for 3d human pose estimation." *Proceedings of the IEEE international conference on computer vision*. 2015.

[10] Oberweger, Markus, Paul Wohlhart, and Vincent Lepetit. "Training a feedback loop for hand pose estimation." *Proceedings of the IEEE international conference on computer vision*. 2015.

[11] Pavllo, Dario, et al. "3d human pose estimation in video with temporal convolutions and semi-supervised training." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[12] Sigal, Leonid, Alexandru O. Balan, and Michael J. Black. "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion." *International journal of computer vision* 87.1-2 (2010): 4.

[13] Ionescu, Catalin, et al. "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 (2013): 1325-1339.

[14] Lee, Kyoungoh, Inwoong Lee, and Sanghoon Lee. "Propagating lstm: 3d pose estimation based on joint interdependency." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[15] Miao, Yunqi, et al. "ST-CNN: Spatial-Temporal Convolutional Neural Network for crowd counting in videos." *Pattern Recognition Letters* 125 (2019): 113-118.

[16] Xu, Zhenqi, Shan Li, and Weihong Deng. "Learning temporal features using LSTM-CNN architecture for face anti-spoofing." *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015.

[17] Shin, Hoo-Chang, et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning." *IEEE transactions on medical imaging* 35.5 (2016): 1285-1298.

[18] Pavlakos, Georgios, et al. "Coarse-to-fine volumetric prediction for single-image 3D human pose." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[19] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems* 28 (2015): 91-99.

Thanks for listening!

Hao Bai(*). "A Training Method For VideoPose3D with Ideology of Action Recognition." 2021 The International Conference on Signal Processing and Machine Learning (CONF-SPML 2021).