




Supplementary Material: Learning 2D Human Poses for Better 3D Lifting via Multi-Model 3D-Guidance

Sanghyeon Lee^{*}, Yoonho Hwang^{*}, and Jong Taek Lee^{†}

School of Computer Science and Engineering
Kyungpook National University, Daegu, South Korea
{hyeon1263,ghkddbshg99,jongtaeklee}@knu.ac.kr

A Dataset Details

For both Human3.6M [2] and Panoptic [4] datasets, we select longer data intervals for computational efficiency. This decision is due to the higher computational intensity and time required for 2D pose estimation, which processes image inputs, compared to training 3D lifting networks with 2D poses. In Human3.6M, every 40th frame is used for training and every 100th frame for testing. In Panoptic, every 15th frame is used for both training and testing. The Panoptic dataset training scenarios include “171204_pose1”, “171204_pose2”, “171204_pose3”, “171026_pose1”, “171026_pose2”, “171026_cello3”, “161029_piano3”, “1161029_piano4”, and “170407_office2”. Testing scenarios include “171026_pose3”, “161029_piano2”, and “170915_office1”.

B Additional Quantitative Results

B.1 Detailed Result of Single-Model Training Method

In this section, we present detailed results on the Human 3.6M dataset, showcasing the action-specific Mean Per Joint Position Error (MPJPE) for various 2D pose detectors and 3D lifting networks, as illustrated in Tab. 1. Our single-model training method generally improves 3D lifting performance for most actions and across all 2D pose detectors and 3D lifting networks. This comprehensive analysis underscores the efficacy of our 3D-guided training method in enhancing performance across various actions in 3D HPE tasks.

B.2 Evaluation for 2D Human Pose Estimation

Our training method primarily aims to improve 3D lifting performance by enhancing 2D pose detector training, focusing on reducing MPJPE. Interestingly, we also observed improvements in 2D pose accuracy. We use the Percentage of

^{*} Equal contributions

[†] Corresponding Author

Table 1: Detailed quantitative comparison results of our single-model 3D-guided training method on Human3.6M [2] in millimeters under MPJPE (mm). FT-2D represents results from traditional 2D pose detector training, while GT denotes 3D lifting using GT 2D poses. Ours represents results from our proposed 3D-guided training method. Lower is better, with the best results highlighted in bold.

2D input	LiftNet:	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
HRNet (FT-2D)	SB	45.6	54.0	49.8	53.6	58.4	64.2	50.4	47.9	62.5	76.0	53.7	54.8	57.2	43.8	47.7	55.1
	IGANet	47.9	56.0	48.6	53.7	56.7	61.7	53.0	47.5	61.0	76.0	55.3	55.3	56.2	46.3	48.2	55.3
	VPose	48.1	54.7	48.5	55.0	57.3	61.2	51.4	48.9	60.2	75.3	53.7	53.2	56.5	45.0	48.3	54.8
	MixSTE	54.1	61.4	52.3	59.4	60.8	63.3	60.0	52.1	63.8	77.2	58.1	60.2	56.6	51.7	51.1	59.3
HRNet (Ours)	SB	43.9	50.5	48.1	51.4	56.6	61.7	49.5	49.3	60.5	73.7	52.0	50.6	53.9	42.1	44.9	52.9
	IGANet	47.2	55.0	48.0	53.8	54.5	61.7	50.3	49.5	60.0	77.4	54.7	52.0	54.6	44.2	47.5	54.4
	VPose	45.1	52.2	47.0	51.5	55.6	57.9	48.9	48.0	58.7	75.3	52.2	49.4	53.9	41.7	44.5	52.5
	MixSTE	52.1	55.9	51.6	56.6	57.4	63.4	56.1	54.1	62.3	74.3	56.8	53.8	56.7	49.7	51.6	56.9
RTMPose (FT-2D)	SB	43.0	51.5	49.6	54.2	56.7	61.7	47.2	47.1	64.5	79.1	52.8	56.7	55.7	39.9	42.8	54.0
	IGANet	43.3	52.0	48.3	53.9	53.6	58.3	47.2	45.9	61.9	77.5	53.0	56.3	52.2	38.9	41.2	52.8
	VPose	44.7	51.9	47.8	55.3	56.0	59.1	46.8	47.6	61.8	77.9	52.7	55.3	54.8	40.2	43.4	53.5
	MixSTE	45.9	54.9	52.9	55.9	55.6	59.8	50.7	48.5	64.0	78.4	55.4	59.5	52.8	42.4	42.9	55.3
RTMPose (Ours)	SB	41.0	48.9	46.6	52.6	52.9	57.9	46.1	45.1	61.2	78.5	51.2	53.7	51.9	38.9	41.4	51.6
	IGANet	40.1	48.5	44.1	51.3	48.6	54.6	45.2	42.2	58.9	77.0	50.1	53.0	48.0	37.0	39.5	49.7
	VPose	43.1	49.7	44.6	54.0	52.1	56.1	46.8	44.8	58.5	76.3	49.9	52.4	51.2	38.0	42.5	51.0
	MixSTE	42.2	50.2	47.5	53.4	50.4	54.3	46.7	45.0	59.7	77.6	50.0	54.4	49.4	39.9	41.0	51.2
DWPose (FT-2D)	SB	43.7	52.2	49.2	56.6	56.4	62.3	48.1	47.8	64.0	86.6	53.6	56.4	55.5	40.4	45.2	54.9
	IGANet	44.8	52.8	46.9	56.1	52.8	59.5	47.9	45.8	62.5	86.4	53.1	56.7	52.4	39.7	44.1	53.9
	VPose	46.7	52.6	47.8	57.8	55.6	60.1	47.7	48.6	61.6	88.4	53.1	55.4	54.9	41.4	47.2	54.9
	MixSTE	47.4	56.8	50.6	58.5	56.1	62.4	52.6	50.3	62.8	88.5	56.5	60.3	53.4	43.6	45.9	56.8
DWPose (Ours)	SB	42.5	50.7	47.8	54.5	53.7	61.2	47.1	45.9	61.4	83.0	52.0	56.7	52.9	38.8	42.9	53.1
	IGANet	43.8	50.3	44.7	54.8	51.5	59.0	47.5	43.6	59.9	78.3	51.4	54.4	50.5	39.8	43.0	51.9
	VPose	45.3	50.8	45.3	56.4	53.3	59.0	47.3	45.6	59.1	80.4	51.2	54.8	52.7	39.8	44.5	52.7
	MixSTE	43.3	49.9	48.6	54.5	52.9	56.3	46.8	44.9	59.9	83.4	52.0	55.0	50.9	41.6	43.4	52.6
GT	SB	34.7	40.5	36.5	40.3	42.1	49.5	42.4	38.4	48.6	52.6	41.3	42.0	42.3	32.3	34.6	41.4
	IGANet	30.4	35.1	30.9	33.4	35.1	41.6	36.6	30.9	40.3	42.4	35.4	36.7	35.8	28.1	29.2	35.0
	VPose	35.8	40.1	35.5	39.6	39.8	43.7	42.1	38.7	45.8	48.8	39.3	40.6	41.1	31.8	34.3	39.9
	MixSTE	35.5	40.7	33.5	38.8	37.5	42.0	42.6	37.1	40.4	43.3	37.1	41.6	38.4	30.8	32.7	38.3

Table 2: Quantitative comparison results using our multi-model 3D-guided training method on Human3.6M and Panoptic under PCK (%) and MPJPE (mm). We train each 2D pose detector using four 3D lifting networks [6, 7, 9, 11] and report the average results across each network. The \uparrow indicates that the higher values are better, while the \downarrow indicates that lower values are better.

2D Detector	Method	Human3.6M		Panoptic	
		PCK@0.2 \uparrow	MPJPE \downarrow	PCK@0.2 \uparrow	MPJPE \downarrow
HRNet	FT-2D	94.93	56.13	94.32	49.83
	Ours	94.82	54.50	94.34	45.80
RTMPose	FT-2D	95.07	53.90	93.70	56.20
	Ours	95.22	50.58	94.50	45.75
DWPose	FT-2D	95.01	55.13	93.05	57.23
	Ours	95.16	52.30	93.84	46.20

Correct Keypoints (PCK) metric to evaluate 2D pose accuracy, following [1, 12], with a matching threshold set to 20% of the bounding box size at the pixel level. Table 2 presents the results of 2D pose accuracy and 3D lifting errors under PCK and MPJPE. This demonstrates that our training method outperforms the FT-2D for most 2D pose detectors. Notably, an exception is observed with HRNet [8] on Human3.6M, where FT-2D achieves a slightly better 2D accuracy with PCK of 94.93% compared to our method’s 94.82%. However, our method excels in 3D lifting performance, achieving an MPJPE improvement of 2.9%.

These results suggest that our 3D-guided training method not only enhances the 3D lifting performance of the 2D pose detector but also improves 2D pose accuracy in most cases.

B.3 Impact of Our Method on 2D Pose Estimation Quality

We conducted additional experiments to evaluate the impact of our method on samples with both high and low 2D pose estimation accuracy. For samples with nearly perfect 2D pose estimations based on the PCK@0.2 metric. As shown in Tab. 3, our method reduced the 3D pose error (MPJPE) by 1.70mm, from 48.30mm to 46.60mm. However, the 2D pose accuracy experienced a slight decline 0.29% decreasing from 100.00% to 99.71%. Our primary goal is to improve 3D pose accuracy, so a minor decrease in 2D pose accuracy is acceptable if we can improve 3D pose performance.

Table 3: All samples and perfect 2D samples: Comparison of 2D and 3D performance on Human3.6M. FT-2D represents results from the traditional 2D pose detector training method. We use the single-model training method with RTMPose and SB models.

	FT-2D \rightarrow Ours	
	All Samples	Perfect 2D Samples
2D Acc (%) \uparrow	95.07 \rightarrow 95.22	100.00 \rightarrow 99.71
3D Error (mm) \downarrow	53.90 \rightarrow 50.58	48.30 \rightarrow 46.60

Table 4: Cross-model validation with HRNet [8] on Human3.6M [2] under MPJPE (mm). Single-model training uses one 3D lifting network for training and tests on others, while multi-model training excludes one 3D lifting network for training and tests on the excluded one.

Method	Trained with	Tested on			
		SB	IGANet	VPose	MixSTE
FT-2D	-	55.1	55.3	54.8	59.3
Single-Model Training (Ours)	SB	-	54.6	53.3	57.8
	IGANet	54.5	-	54.8	58.5
	VPose	53.4	55.0	-	58.1
	MixSTE	54.9	55.6	55.3	-
Multi-Model Training (Ours)	w/o SB	54.8	-	-	-
	w/o IGANet	-	54.4	-	-
	w/o VPose	-	-	53.8	-
	w/o MixSTE	-	-	-	56.4

Table 5: Cross-model validation with DWPose [10] on Human3.6M [2] under MPJPE (mm).

Method	Trained with	Tested on			
		SB	IGANet	VPose	MixSTE
FT-2D	-	54.9	53.9	54.9	56.8
Single-Model Training (Ours)	SB	-	52.0	52.8	55.2
	IGANet	54.2	-	53.3	56.5
	VPose	53.7	52.3	-	56.0
	MixSTE	53.5	51.7	53.3	-
Multi-Model Training (Ours)	w/o SB	54.3	-	-	-
	w/o IGANet	-	51.4	-	-
	w/o VPose	-	-	53.3	-
	w/o MixSTE	-	-	-	53.9

Table 6: Ablation study on the scaling factor β in soft-argmax using HRNet [8] and SB [6] under MPJPE (mm).

β	1	10	50	100	200	500
MPJPE	319.41	54.13	53.24	52.85	53.70	56.30

B.4 Additional Results Cross-Model Validation

In this section, we present additional cross-model validation results for HRNet [8] and DWPose [10] to further demonstrate the generalization capabilities of our multi-model training method. Table 4 and Tab. 5 illustrate the cross-model validation results for HRNet and DWPose, respectively. These results demonstrate that our multi-model training method generally exhibits robust generalization performance. The improvements observed in both HRNet [8] and DWPose [10] experiments further validate the robustness and versatility of our method in different architectural contexts.

B.5 Ablation Study on Soft-argamx

In our experiments, the soft-argmax operation utilizes a scaling factor β as a hyperparameter, which controls the sharpness of the distribution in the heatmap after applying softmax. A higher β value makes the distribution more peaked, closely approximating the argmax operation, while a lower β value results in a smoother distribution. To find the best β value, we conducted an ablation study across a range of values: 1, 10, 50, 100, 200, and 500. As shown in Tab. 6, the best performance is achieved when $\beta = 100$, with an MPJPE of 52.85mm. Consequently, we set β to 100 for our experiments.

Additionally, we explored the impact of replacing softmax with ReLU, following [5]. The results showed that softmax achieved an MPJPE of 51.6mm, while ReLU led to a slightly higher MPJPE of 53.1mm. These findings suggest

Table 7: Results of different hyperparameter settings for our **single-model** training method on Human3.6M. The ‘‘Proposed’’ row represents the baseline configuration we used, and each subsequent row shows results from varying one hyperparameter while keeping others at their baseline values.

Single-Model Training Method					
Experiment	λ_{2D}	λ_O	λ_{xy}	λ_z	MPJPE
Proposed	0.5	0.3	10.0	20.0	51.6
Vary λ_{2D}	0.3 / 0.7	-	-	-	51.6 / 52.0
Vary λ_O	-	0.2 / 0.4	-	-	51.6 / 51.9
Vary λ_{xy}	-	-	5.0 / 15.0	-	51.8 / 51.8
Vary λ_z	-	-	-	10.0 / 30.0	52.1 / 51.8

Table 8: Results of different hyperparameter settings for our **multi-model** training method on Human3.6M.

Multi-Model Training Method				
Experiment	λ_{single}	λ_{local}	λ_{global}	MPJPE
Proposed	0.8	15.0	5.0	50.6
Vary λ_{single}	0.6 / 1.0	-	-	50.9 / 50.9
Vary λ_{local}	-	10.0 / 20.0	-	51.0 / 51.1
Vary λ_{global}	-	-	3.0 / 7.0	50.7 / 50.7

that softmax remains a better choice for minimizing pose estimation errors in our framework.

B.6 Effect of Hyperparameters in Our Method

We conducted experiments with different hyperparameter settings used in both single-model and multi-model training methods. The results for our single-model training method are presented in Tab. 7, while the results for the multi-model training method are shown in Tab. 8.

B.7 Comparison with Voting-Based Baseline

We conducted additional experiments comparing our method to a baseline that utilizes a voting approach among multiple 2D pose detectors. In this approach, the most accurate detector (HRNet [8], DWPose [10], RTMPose [3]) is selected for each sample. This voting method achieved an MPJPE of 53.1mm. Despite its strong performance, our method, which uses only RTMPose, outperformed the voting approach with an MPJPE of 51.6mm, demonstrating its superior effectiveness.

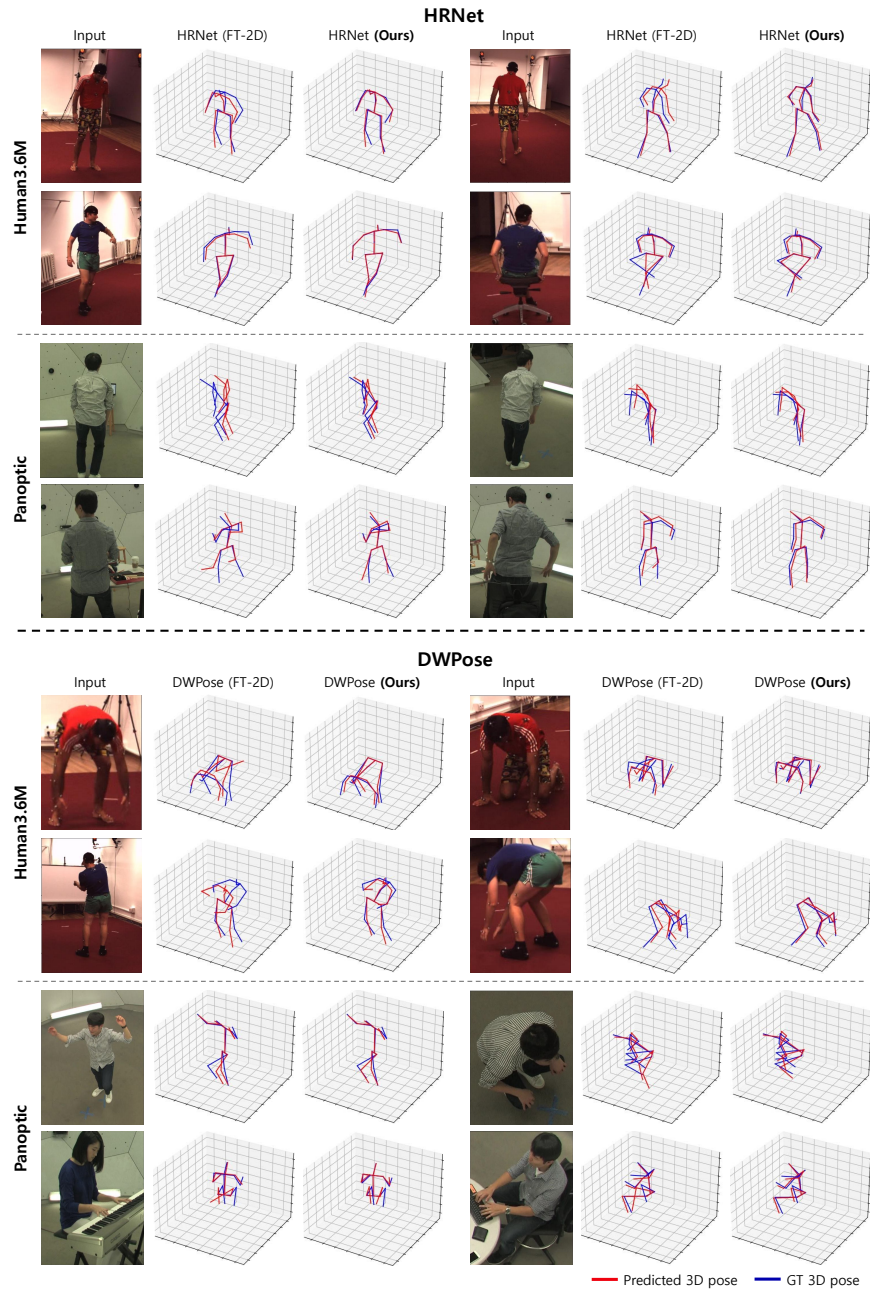


Fig. 1: Qualitative comparisons with the traditional training method (FT-2D) with HRNet [8] (top) and DWPose [10] (bottom) on Human3.6M and Panoptic. The blue lines represent GT 3D poses, while the red lines indicate 3D poses lifting from the predicted 2D poses.

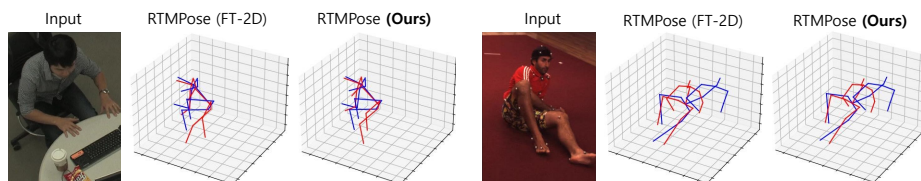


Fig. 2: Failure cases of our multi-model training methods using RTMPose [3] and SB [6]. Both FT-2D and Ours approach show suboptimal results. Panoptic example (left) demonstrates a case with severe occlusion, while Human3.6M (right) presents a scenario where the subject’s clothing color is similar to the background.

C Additional Qualitative Results

In this section, we present additional qualitative results and failure cases based on our multi-model training method with with four 3D lifting networks, tested on SB [6]. Figure 1 shows qualitative comparisons with the traditional training method using HRNet [8] and DWPose [10] on Human3.6M and Panoptic. The 2D poses estimated by HRNet and DWPose trained with our multi-model training method show higher accuracy and improved depth estimation compared to the traditional training method. This improvement is particularly evident in scenarios with complex poses and occlusions, underscoring our method’s robustness and effectiveness in diverse conditions.

In addition, we present failure cases of RTMPose [3] visualized in Fig. 2, including examples that involve severe occlusions and subjects with clothing colors matching the background. These challenging scenarios highlight areas where our method still struggles, providing insight for further improvements.

References

1. Fan, Z., Liu, J., Wang, Y.: Motion adaptive pose estimation from compressed videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11719–11728 (2021) 2
2. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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), 1325–1339 (2013) 1, 2, 3, 4
3. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: Rtm-pose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399* (2023) 5, 7
4. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE international conference on computer vision. pp. 3334–3342 (2015) 1
5. Kumar, A., Marks, T.K., Mou, W., Wang, Y., Jones, M., Cherian, A., Koike-Akino, T., Liu, X., Feng, C.: Luvli face alignment: Estimating landmarks’ location, uncertainty, and visibility likelihood. In: CVPR. pp. 8236–8246 (2020) 4

6. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 2640–2649 (2017) [2](#), [4](#), [7](#)
7. Pavlo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 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 (2019) [2](#)
8. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019) [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
9. Wang, T., Liu, H., Ding, R., Li, W., You, Y., Li, X.: Interweaved graph and attention network for 3d human pose estimation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) [2](#)
10. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023) [4](#), [5](#), [6](#), [7](#)
11. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13232–13242 (2022) [2](#)
12. Zhang, Y., Wang, Y., Camps, O., Sznajder, M.: Key frame proposal network for efficient pose estimation in videos. In: European Conference on Computer Vision. pp. 609–625. Springer (2020) [2](#)