This ACCV 2024 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2024
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

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,ghkddbsgh99,jongtaeklee}@knu.ac.kr

Abstract. Recent advancements in 2D pose detectors have significantly improved 3D human pose estimation via the 2D-to-3D lifting approach. Despite these advancements, a substantial accuracy gap remains between using ground-truth 2D poses and detected 2D poses for 3D lifting. However, most methods focus solely on enhancing the 3D lifting network, using 2D pose detectors optimized for 2D accuracy without any refinement to better serve the 3D lifting process. To address this limitation, we propose a novel 3D-guided training method that leverages 3D loss to improve 2D pose estimation. Additionally, we introduce a multi-model training method to ensure robust generalization across various 3D lifting networks. Extensive experiments with three 2D pose detectors and four 3D lifting networks demonstrate our method's effectiveness. Our method achieves an average improvement of 4.6% in MPJPE on Human3.6M and 16.8% on Panoptic, enhancing 2D poses for accurate 3D lifting. The code is available at https://github.com/knu-vis/L2D-Pose.

Keywords: Human pose estimation · Training strategy

1 Introduction

Monocular 3D Human Pose Estimation (HPE) aims to estimate the 3D locations of body joints from a single-view image. This task captures significant details of body geometry and motion, making it widely applicable in fields such as action recognition [6, 27, 36], augmented reality [17, 23], and human-robot interaction [12, 40]. Monocular 3D HPE can be divided into two main approaches. The first approach is the direct estimation method [34, 38, 39, 55], which directly predicts 3D poses from images. The second approach is the 2D-to-3D lifting method [29, 30, 35, 44, 51, 56], where 2D poses are first estimated by an off-the-shelf 2D pose detector from images and then lifted to 3D space. Due to the reliable performance of 2D pose detectors [2, 15, 37, 47] and the availability of abundant 2D pose data, the majority of recent works are based on the 2D-to-3D lifting approach, which has demonstrated superior results compared to direct estimation methods [53].

⋆ Equal contributions
† Corresponding Author

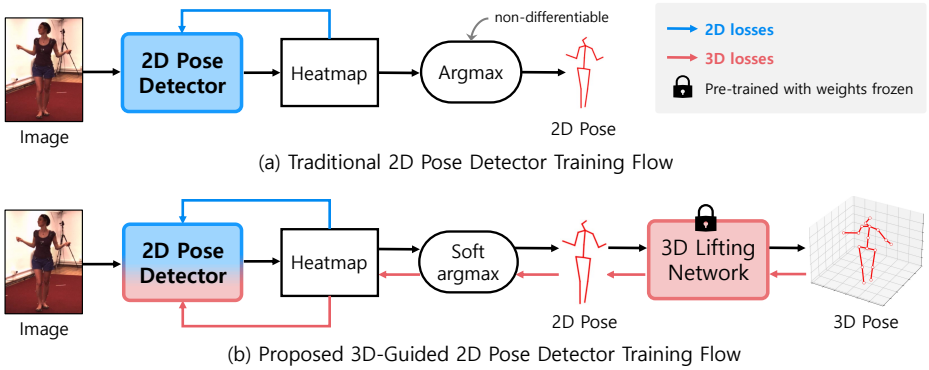

Fig. 1: Comparison of traditional and proposed 2D pose detector training method. Unlike traditional 2D pose detector training (a), which relies solely on 2D pose data, our 3D-guided training method (b) incorporates a differentiable soft-argmax, integrating 3D lifting networks and 3D losses. By leveraing both 2D and 3D information, our method produces 2D poses better for accurate 3D lifting.

Despite significant advancements leading to near-perfect performance in 2D pose estimation, a substantial gap still exists in 3D pose accuracy when detected 2D poses are used as inputs for 3D lifting networks compared to ground-truth (GT) 2D poses [10, 29, 45, 49, 51, 54, 56]. Since practical applications in 3D pose estimation [1, 41, 57] rely on detected 2D poses due to the unavailability of GT 2D poses, it is crucial to enhance the accuracy of 3D poses derived from detected 2D poses. However, existing 2D-to-3D lifting works focus on improving 3D lifting networks, using off-the-shelf 2D pose detectors, leaving the enhancement of 2D pose detection coupled with 2D-to-3D lifting largely unexplored.

These pre-trained 2D pose detectors are typically optimized for 2D pose accuracy using supervision on 2D pose data, rather than being designed for effective lifting to 3D poses. Consequently, while detected 2D poses perform well in 2D metrics, they result in significant discrepancies in 3D pose accuracy compared to GT 2D poses. This underscores the need for efforts to bridge the gap in 3D pose accuracy when using detected 2D poses, ensuring that the 2D poses are optimized not only for 2D accuracy but also for effective 3D lifting.

To address this gap, we propose a training method that trains the 2D pose detector for more accurate 3D lifting by combining the strengths of direct estimation and 2D-to-3D lifting. Since the 2D-to-3D lifting method involves two stages, end-to-end learning is infeasible. By replacing the non-differentiable argmax operation with a differentiable soft-argmax [26, 28, 33], we combine the 2D pose detector and 3D lifting network, allowing for integrated training. This integration allows us to leverage both 2D and 3D losses, as illustrated in Fig. 1. We introduce the Online Hard Keypoint for Lifting Mining (OHKLM) loss, which focuses on keypoints with the highest 3D lifting errors, ensuring the 2D pose detector benefits from the valuable knowledge in the 3D lifting network with the

3D data. Consequently, our adaptable approach improves 2D pose estimation, thereby enhancing the overall performance of 3D HPE.

Furthermore, we propose a multi-model training method that trains a 2D pose detector with multiple 3D lifting networks to prevent overfitting to a single lifting network. Building on multi-task learning principles [9, 27, 46], which improves performance by sharing information across related tasks, our multi-model training ensures that the 2D pose detector generalizes well across different 3D lifting networks. Due to the computational efficiency of the 3D lifting networks compared to the 2D pose detector, we can utilize multiple 3D lifting networks without burden, improving generalization. We introduce consistency loss and model-wise dropout in our multi-model training. By demonstrating robust performance in cross-model validation, we prove the effectiveness and generalization capability of our training method.

We evaluate our method with three 2D pose detectors [15, 37, 47] and four 3D lifting networks [29, 35, 44, 51] on the Human3.6M [13] and Panoptic [16] datasets. Our 3D-guided training method for 2D pose detectors significantly outperforms existing training methods. Note that we did not fine-tune the 3D lifting networks during our experiments; the improvements in 3D HPE are solely attributed to the enhanced training of the 2D pose detectors. Furthermore, our multi-model training method demonstrates robust generalization capabilities, achieving strong performance across multiple 3D lifting networks. In summary, our contributions are threefold, as follows:

- We introduce a new perspective on training 2D pose detectors within a 2D-to-3D lifting approach, enhancing 3D lifting accuracy by integrating 2D pose detectors with 3D lifting networks.
- We propose a multi-model training method with multiple 3D lifting networks with model-wise dropout and consistency losses, showing effective cross-model generalization.
- With our training method, we improve 3D pose accuracy without fine-tuning the 3D lifting network. Extensive experiments demonstrate superior performance and generalization across various models.

2 Related Work

2.1 2D Human Pose Estimation

2D HPE aims to estimate the coordinates of human joints in images and can be broadly categorized into regression-based and heatmap-based approaches.

Unlike heatmap-based methods, regression-based approaches [3, 8, 21, 42] directly regress the keypoints instead of relying on heatmaps. This methodology has the advantage of being end-to-end trainable. However, due to its unsatisfactory performance compared to heatmap-based approaches, regression-based methods have received less attention.

In heatmap-based methods [15, 22, 32, 37, 47], the network generates probability heatmaps that indicate where keypoints are likely to exist among body

parts. Keypoints are then determined by selecting the locations with the highest probabilities on the heatmap. This approach offers several advantages, such as preserving spatial location information and providing richer supervision. Consequently, heatmap-based methods consistently achieve state-of-the-art performance and dominate the field of 2D HPE [15, 47].

2.2 Monocular 3D Human Pose Estimation

Monocular 3D HPE aims to estimate 3D body joint locations from a single-view image and includes two main approaches: direct estimation and 2D-to-3D lifting.

Direct estimation methods [33, 34, 38, 39] estimate 3D pose directly from images without intermediate 2D poses, allowing end-to-end learning and a simple one-stage structure. However, it suffers from performance degradation due to limited 3D data diversity. While some studies [31,55] leverage abundant 2D data to mitigate this, performance still lags behind the 2D-to-3D lifting approach.

To address the lack of 3D data, 2D-to-3D lifting methods [14,19,20,29,30,44, 49,51] utilize 2D pose detectors with excellent performance on abundant 2D data. These methods first estimate the 2D pose with a 2D pose detector, then predict the 3D pose from the estimated 2D pose using a 3D lifting network. Various 2D-to-3D lifting methods have been proposed, including simple fully connected network [29], skeleton-LSTM network [19], multiple hypotheses [14,20], graph-based networks [44, 49], and transformer-based networks [30, 51]. Due to the outstanding performance of 2D pose detectors [2, 15, 37, 47], 2D-to-3D lifting methods still achieve state-of-the-art performance in 3D pose estimation [30, 49, 56]. Despite these advancements, a significant gap remains remains between using GT 2D input and predicted 2D input for 3D lifting. However, most 2D-to-3D lifting methods focus solely on improving the 3D lifting network, utilizing 2D pose detectors optimized for 2D accuracy without refinement with 3D information.

In this study, we propose a training method that combines the advantages of direct estimation and 2D-to-3D lifting. By training a 2D pose detector with 3D information, we aim to produce 2D poses for more accurate 3D lifting.

2.3 Multi-task learning

Multi-task learning aims to improve the overall performance of all tasks by sharing useful information among various related tasks. It is commonly used to ensure data diversity and generalize task performance, and is extensively employed in the field of machine learning [4, 7, 18, 25, 52]. In the domain of pose estimation, multi-task learning is often utilized in conjunction with similar tasks such as action recognition and mesh reconstruction [9, 27, 46, 56]. The advantages of multi-task learning include improved performance when tasks are more similar and increased generalization performance, which helps avoid overfitting [4, 25].

Leveraging these benefits, we utilize multiple 3D lifting models from a multi-task perspective, focusing on the same task: 3D HPE. By training the 2D pose detector with multiple 3D lifting networks, we integrate the 3D loss from the multi-model to enhance the generalization performance of the 2D pose detector.

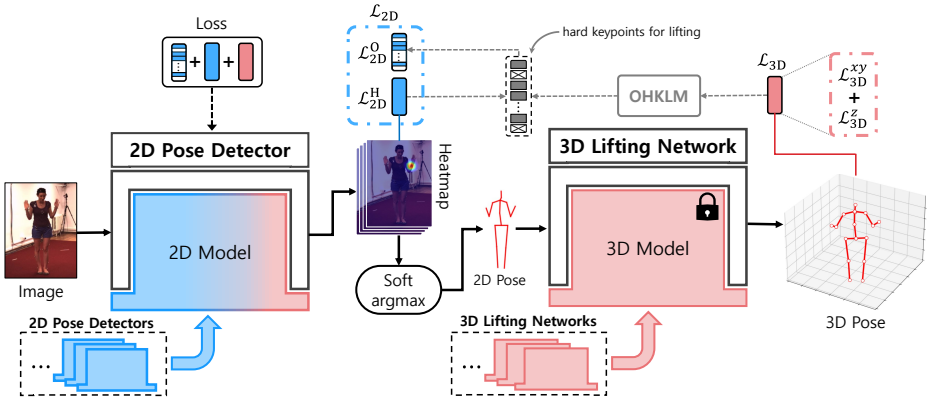

Fig. 2: Overview of our 3D-guided training method for 2D-to-3D lifting. The 3D loss $\mathcal{L}_{3D}$ from 3D lifting network, divided into $\mathcal{L}^{xy}_{3D}$ and $\mathcal{L}^{z}_{3D}$, is combined with the heatmap loss $\mathcal{L}^{H}_{2D}$ and OHKLM loss $\mathcal{L}^{O}_{2D}$ to train the 2D pose detector. This enables the 2D pose detector to learn from the 3D lifting network, improving its lifting performance. Note that the 2D pose detector and 3D lifting network can be replaced with any models.

3 Method

We introduce our approach to enhance 3D lifting performance by training 2D pose detectors. First, we present the basic 3D-guided training approach, integrating 2D and 3D loss functions (Sec. 3.1). Next, we propose a multi-model 3D-guided training method using multiple 3D lifting networks to improve generalization (Sec. 3.2). Finally, we detail our training strategy (Sec. 3.3).

3.1 Single-Model 3D-Guided Training Approach

The limitation of the 2D-to-3D lifting method is that the non-differentiable argmax operation used to extract 2D poses from 2D heatmaps prevents end-to-end learning. By replacing the argmax with a differentiable soft-argmax operation, we integrate 2D and 3D loss to train the 2D pose detector, forming the foundation of our approach. To utilize the richer information from the 3D loss, we introduce loss functions to train the 2D pose detector. This approach allows the 2D pose detector to better understand 3D structures, improving performance in 3D lifting tasks. As shown in Fig. 2, we use a single lifting network and learn comprehensively through a combination of 2D and 3D losses.

Soft-argmax. The soft-argmax operation [28] has been proposed to convert heatmaps directly into coordinates, making the entire process differentiable. While direct 3D pose estimation from images [26,33] has utilized the soft-argmax to learn from 3D volumetric heatmaps, there has been no attempt to apply the soft-argmax operation in 2D-to-3D lifting methods. We replace the argmax with

a differentiable soft-argmax operation. Given a pose heatmap $\mathbf{H} \in \mathbb{R}^{h \times w}$ of height h and width w, the soft-argmax computes the coordinates (x, y) as:

$$(x, y) = \sum_{i=1}^{h} \sum_{j=1}^{w} (j, i) \cdot \text{softmax}(\beta \mathbf{H})_{ij}, \tag{1}$$

where β is a scaling factor, set to 100. By incorporating the soft-argmax operation, we ensure the 2D-to-3D lifting process becomes fully differentiable, allowing the 3D loss to backpropagate through the 2D pose detector.

2D Loss. For the 2D loss, we primarily use heatmap loss $\mathcal{L}_{\text{2D}}^{\text{H}}$, computed between the GT and predicted heatmaps, following [15, 37, 47]. Additionally, we introduce the Online Hard Keypoints for Lifting Mining (OHKLM), inspired by Online Hard Keypoints Mining (OHKM) [5]. While OHKM targets K keypoints with the highest 2D loss, OHKLM focuses K keypoints with the highest 3D lifting errors. The OHKLM loss $\mathcal{L}_{\text{2D}}^{\text{O}}$ is defined as:

$$\mathcal{L}_{\text{2D}}^{\text{O}} = \frac{1}{K} \sum_{k \in \mathcal{H}} \mathcal{L}_{\text{2D}}^{\text{H}}(k), \tag{2}$$

where $\mathcal{H}$ represents the set of K keypoints with the highest 3D lifting errors, focusing on the top $K (K \leq N)$ losses out of the N annotated keypoints for each person. We incorporate both the heatmap and OHKLM loss in our 2D loss $\mathcal{L}_{\text{2D}}$, ensuring all keypoints are considered while prioritizing those challenging for 3D lifting, thus enhancing overall 3D lifting accuracy.

3D Loss. We observed higher 3D pose error in the z direction compared to x and y. To better capture depth information, we separate the 3D loss into xy and z components, calculating them independently and giving more weight to the depth loss. This approach enhances the learning of 3D lifting by allowing the model to more accurately represent and learn from depth variations. Here, $[X, Y, Z]^\top = (x_i, y_i, z_i)_{i=1}^{N} \in \mathbb{R}^{N \times 3}$, where (x_i, y_i, z_i) represents the 3D coordinates of the i-th keypoint. We decompose the 3D loss $\mathcal{L}_{\text{3D}}$ into two components, $\mathcal{L}_{\text{3D}}^{xy}$ and $\mathcal{L}_{\text{3D}}^{z}$, calculated using the Mean Squared Error (MSE) loss as follows:

$$\mathcal{L}_{\text{3D}}^{xy} = \|[X, Y]^\top - [\hat{X}, \hat{Y}]^\top\|_2^2, \;\; \mathcal{L}_{\text{3D}}^{z} = \|Z - \hat{Z}\|_2^2, \tag{3}$$

where $[X, Y, Z]^\top$ is the GT 3D pose, and $[\hat{X}, \hat{Y}, \hat{Z}]^\top$ is the predicted 3D pose.

We combine these 2D and 3D losses to improve the 3D lifting performance of the 2D pose detector. This integration helps the 2D pose detector learn more accurate poses for subsequent 3D lifting. The final loss function for single-model 3D-guided training is defined as follows:

$$\mathcal{L}_{\text{single}} = \underbrace{\lambda_{\text{H}} \mathcal{L}_{\text{2D}}^{\text{H}} + \lambda_{\text{O}} \mathcal{L}_{\text{2D}}^{\text{O}}}_{\text{2D loss } (\mathcal{L}_{\text{2D}})} + \underbrace{\lambda_{xy} \mathcal{L}_{\text{3D}}^{xy} + \lambda_{z} \mathcal{L}_{\text{3D}}^{z}}_{\text{3D loss } (\mathcal{L}_{\text{3D}})}, \tag{4}$$

where each λ is a hyperparameter that adjusts the weight of each loss term.

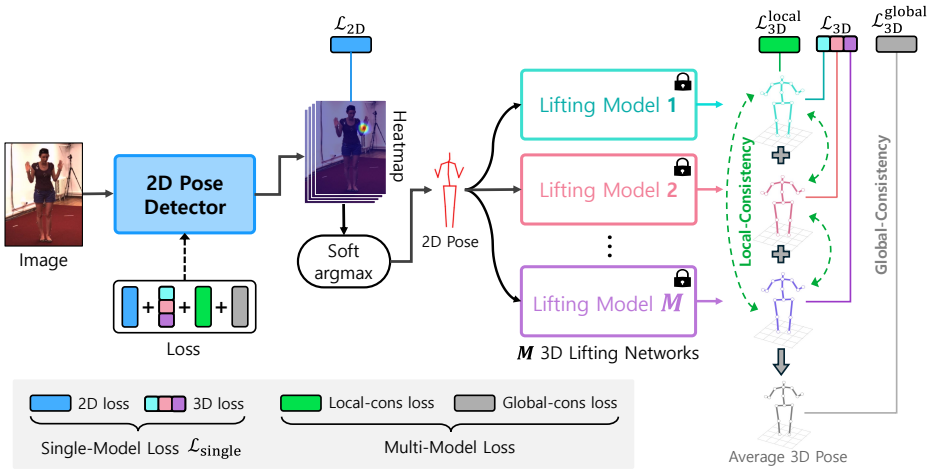

Fig. 3: Overview of our **multi-model** 3D-guided training method with M 3D lifting networks. This approach trains the 2D pose detector using diverse 3D losses and two consistency losses to ensure consistent predictions. By leveraging the collective knowledge of various 3D models, it enhances the 2D pose detector's generalization ability.

3.2 Multi-Model 3D-Guided Training Approach

In Sec. 3.1, we discussed training the 2D pose detector with a single 3D lifting network. However, this approach can lead to overfitting and poor performance on new 3D lifting networks. To address this, we propose a multi-model 3D guided training method, illustrated in Fig. 3. This approach uses multiple 3D lifting networks to guide the 2D pose detector. By propagating multi-model 3D loss, the 2D detector learns poses that generalize well across different lifting networks. During inference, only a single 3D network is used.

Multi-Model Training. To adaptively adjust the contribution of each 3D lifting network, we adopt the Dynamic Weight Average (DWA) [25], which assigns dynamic weights to the 3D losses $\mathcal{L}_{3D}$ from each network for balanced training. The weight λ_k for the k-th 3D lifting network is computed as:

$$\lambda_k(t) := \frac{\exp(w_k(t-1)/T)}{\sum_i \exp(w_i(t-1)/T)}, \quad w_k(t-1) = \frac{\mathcal{L}_k(t-1)}{\mathcal{L}_k(t-2)}, \tag{5}$$

where t is an iteration index and T is a temperature parameter controlling the softness of each weighting, set to $T = 2$. $\mathcal{L}_k(t-1)$ and $\mathcal{L}_k(t-2)$ are the losses for the k-th 3D lifting network at the $(t-1)$-th and $(t-2)$-th iterations, respectively.

Additionally, we incorporate model-wise dropout, inspired by dropout techniques [11]. During training, it randomly excludes certain 3D lifting networks, preventing the 2D pose detector from overfitting to a specific 3D lifting network. This enhances the generalization capability of the 2D pose detector.

Multi-Model 3D Loss. To enhance generalization by ensuring consistent predictions from multiple 3D lifting networks, we introduce two types of consistency loss: local consistency loss and global consistency loss.

Local consistency loss compares 3D pose predictions from each 3D lifting network in pairs. By minimizing the differences between these predictions, the 2D pose predictor learns the detailed differences between the results of each network and helps improve local consistency. Given $\hat{\mathbf{J}}_m$ as the predicted 3D pose from the m-th 3D lifting network, with M representing the total number of lifting networks, the local consistency loss $\mathcal{L}_{\text{3D}}^{\text{local}}$ is computed as follows:

$$\mathcal{L}_{\text{3D}}^{\text{local}} = \sum_{m=1}^{M} \sum_{n=1, n\neq m}^{M} \|\hat{\mathbf{J}}_m - \hat{\mathbf{J}}_n\|_2^2. \tag{6}$$

Global consistency loss averages the predictions from all 3D lifting networks and computes the loss based on this average. Unlike local consistency loss, it integrates the overall pose to learn broad and comprehensive information, thereby encouraging the model to produce robust and reliable poses. The global consistency loss $\mathcal{L}_{\text{3D}}^{\text{global}}$ is computed as follows:

$$\mathcal{L}_{\text{3D}}^{\text{global}} = \left\| \frac{1}{M} \sum_{m=1}^{M} \hat{\mathbf{J}}_m - \mathbf{J} \right\|_2^2, \tag{7}$$

where $\mathbf{J}$ is the GT 3D pose. The final multi-model loss $\mathcal{L}_{\text{multi}}$ is defined as:

$$\mathcal{L}_{\text{multi}} = \lambda_{\text{single}}\mathcal{L}_{\text{single}} + \underbrace{\lambda_{\text{local}}\mathcal{L}_{\text{3D}}^{\text{local}} + \lambda_{\text{global}}\mathcal{L}_{\text{3D}}^{\text{global}}}_{\text{multi-model 3D loss}}, \tag{8}$$

where each λ is a hyperparameter adjusting the ratio of each loss term. Integrating these losses ensures consistent and accurate 3D poses, leveraging the collective knowledge of multiple 3D lifting networks.

3.3 Training Strategies

We use 2D pose detectors pre-trained on the COCO dataset [24] and fine-tune them on 3D datasets [13,16] using our method. The 3D lifting networks are pre-trained on each 3D dataset following the respective methods [29,35,44,51]. Our method does not involve fine-tuning the 3D lifting networks; instead, we focus on enhancing the 2D pose detectors for improved 3D lifting performance.

For training, we followed the optimizer and learning rate specified for each 2D pose detector (HRNet [37], RTMPose [15], DWPose [47]), training each for 50 epochs. All input images are resized to 288×384 for consistency across experiments. Since each 2D pose detector uses different loss functions and learning rates, we adjusted the absolute magnitude of the weights for each loss term while maintaining their ratios. For all detectors, 2D loss terms, λ_{H} and λ_{O} are set to 0.5 and 0.3, with λ_{single} at 0.8. For 3D loss terms, HRNet uses $\lambda_{xy} = 0.02$, $\lambda_z = 0.04$, $\lambda_{\text{local}} = 0.03$, and $\lambda_{\text{global}} = 0.01$, while RTMPose and DWPose use $\lambda_{xy} = 10.0$, $\lambda_z = 20.0$, $\lambda_{\text{local}} = 15$, and $\lambda_{\text{global}} = 5$.

Table 1: Quantitative comparison results of 3D lifting errors using our **single-model** 3D-guided training method on **Human3.6M** [13] in millimeters under MPJPE and P-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. For the video-based models VPose [35] and MixSTE [51], we configured the experiments with the number of input frames set to 1.

Metric	LiftNet.	HR-Net		RTMPose		DWPose		GT
		FT-2D	Ours	FT-2D	Ours	FT-2D	Ours	
MPJPE	SB	55.1	**52.9**	54.0	**51.6**	54.9	**53.1**	41.4
	IGANet	55.3	**54.4**	52.8	**49.7**	53.9	**51.9**	35.0
	VPose	54.8	**52.5**	53.5	**51.0**	54.9	**52.7**	39.9
	MixSTE	59.3	**56.9**	55.3	**51.2**	56.8	**52.6**	38.3
P-MPJPE	SB	42.8	**41.3**	43.0	**41.3**	43.6	**42.3**	32.6
	IGANet	43.3	**41.7**	42.0	**39.6**	42.5	**41.0**	27.9
	VPose	42.6	**41.2**	42.5	**40.8**	43.2	**41.8**	31.0
	MixSTE	45.9	**43.7**	43.6	**41.2**	44.4	**42.1**	29.3

4 Experiments

4.1 Datasets and Evaluation Metrics

Human3.6M [13] is the most commonly used benchmark for 3D HPE, consisting of 3.6 million images captured from four cameras, where 15 daily activities are performed by 11 subjects. We follow previous works [29, 35, 44, 51] to train on five subjects (S1, S5, S6, S7, S8) and test on two subjects (S9, S11). More details are in supplementary materials.

Panoptic [16] is a comprehensive multi-camera dataset for 3D HPE. It provides 30 Hz Full HD video streams of 40 subjects from up to 31 synchronized cameras. Following [43, 48, 50], we use five HD cameras (3, 6, 12, 13, 23). Only single-person scenarios are used in our experiments. We divide it into training and test sets without any overlap. More details are in supplementary materials.

Evaluation Metrics. We employ two commonly used evaluation metrics in 3D HPE for both datasets: Mean Per Joint Position Error (MPJPE) and Procrustes MPJPE (P-MPJPE). MPJPE measures the Euclidean distance between the predicted joint positions and GT joint positions. P-MPJPE is the MPJPE after aligning the predicted poses to the GT via a rigid transformation.

4.2 Single-Model 3D-Guided Training Results

In this section, we present the results of our single-model training method as discussed in Sec. 3.1. Three 2D pose detectors, pre-trained on the COCO dataset,

Table 2: Quantitative comparison results of 3D lifting errors using our **single-model** 3D-guided training method on **Panoptic** [16] in millimeters under MPJPE and P-MPJPE (mm). Lower is better, with the best results highlighted in bold.

Metric	LiftNet.	HR-Net		RTMPose		DWPose		GT
		FT-2D	Ours	FT-2D	Ours	FT-2D	Ours	
MPJPE	SB	49.9	**46.7**	57.2	**46.8**	57.8	**46.8**	34.5
	IGANet	49.0	**45.4**	53.3	**45.1**	55.3	**45.3**	29.7
	VPose	50.2	**45.8**	59.7	**45.4**	59.9	**46.2**	31.4
	MixSTE	50.2	**45.3**	54.6	**45.7**	55.9	**46.5**	27.5
P-MPJPE	SB	36.0	**34.3**	42.0	**34.3**	41.6	**35.0**	24.4
	IGANet	36.6	**34.3**	41.7	**33.6**	42.5	**34.6**	20.0
	VPose	36.0	**33.8**	41.9	**33.1**	42.4	**35.0**	22.4
	MixSTE	36.7	**33.5**	42.5	**33.6**	43.0	**35.0**	19.5

are evaluated alongside four different networks. For each 2D pose detector, we compare the standard fine-tuning with 2D pose supervision methods [15, 37, 47] to fine-tuning using our training method on Human3.6M and Panoptic.

Table 1 shows the results of 3D lifting errors on Human3.6M for various models. Our single-model training method improves 3D lifting performance across all 2D pose detectors and 3D lifting networks. Specifically, our method achieves average improvements of 1.9mm (**3.5%**) for HRNet, 3.0mm (**5.6%**) for RTMPose, and 2.6mm (**4.6%**) for DWPose under MPJPE. Table 2 presents the results on Panoptic where more substantial improvements are observed: 4.0mm (**8.1%**) for HRNet, 10.4mm (**18.6%**) for RTMPose, 11.0mm (**19.3%**) for DWPose under MPJPE. These results demonstrate that our method significantly enhances 3D lifting performance across various 2D pose detectors and 3D lifting networks. This underscores the effectiveness of our method in improving overall 3D HPE performance.

4.3 Multi-Model 3D-Guided Training Results

In Sec. 4.2, we demonstrated the superiority of our single-model training with each 3D lifting network. However, this approach may overfit the specific 3D lifting network used, limiting generalization. In this section, we evaluate our multi-model 3D-guided training method with multiple 3D lifting networks introduced in Sec. 3.2.

Table 3 presents the results of each 2D pose detector trained with all four 3D lifting networks and tested on each network individually on Human3.6M and Panoptic. Our multi-model method ensures the 2D pose detector consistently outperforms across all 3D lifting networks by leveraging the collective knowledge of multiple 3D lifting networks. Compared to the single-model training results (Tab. 1 and Tab. 2), the multi-model training method demonstrates comparable performance on Human3.6M and generally superior performance on Panoptic, as shown in Tab. 3. These results indicate that our multi-model training method effectively enhances both generalization and overall performance.

Table 3: Quantitative comparison results of 3D lifting errors using our **multi-model** 3D-guided training method on Human3.6M and Panoptic under MPJPE (mm). We train a 2D pose detector using four 3D lifting networks and report the results for each lifting network individually.

		Human3.6M		Panoptic	
2D Detector	LiftNet.	FT-2D	Ours	FT-2D	Ours
HRNet	SB	55.1	**54.2**	49.9	**46.2**
	IGANet	55.3	**53.5**	49.0	**45.4**
	VPose	54.8	**54.3**	50.2	**45.2**
	MixSTE	59.3	**56.0**	50.2	**44.9**
RTMPose	SB	54.0	**51.5**	57.2	**45.5**
	IGANet	52.8	**49.3**	53.3	**44.6**
	VPose	53.5	**51.0**	59.7	**44.8**
	MixSTE	55.3	**50.5**	54.6	**44.6**
DWPose	SB	54.9	**53.2**	57.8	**46.3**
	IGANet	53.9	**50.7**	55.3	**45.4**
	VPose	54.9	**52.9**	59.9	**45.5**
	MixSTE	56.8	**52.4**	55.9	**45.2**

Table 4: Cross-model validation with RTMPose [15] on Human3.6M [13] 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	-	54.0	52.8	53.5	55.3
Single-Model Training (Ours)	SB	-	50.8	51.3	52.7
	IGANet	52.1	-	**51.1**	53.0
	VPose	52.2	50.9	-	53.7
	MixSTE	52.2	50.4	51.7	-
Multi-Model Training (Ours)	w/o SB	**51.8**	-	-	-
	w/o IGANet	-	**50.3**	-	-
	w/o VPose	-	-	51.2	-
	w/o MixSTE	-	-	-	**52.5**

Additionally, we conduct cross-model validation to further evaluate the generalization capabilities of our multi-model training method using RTMPose [15]. Single-model training involves training on one 3D lifting network and testing on others, while multi-model training involves training on multiple 3D lifting networks minus one and testing on the excluded one. Table 4 shows that our multi-model training method generally outperforms single-model training in cross-model validation on unseen lifting networks. Specifically, compared to the single-model training average, it achieves improvements of 0.4mm MPJPE for

Table 5: Ablation study of OHKLM under MPJPE (mm) on Human3.6M. K represents the number of hard keypoints for lifting.

K	6	8	10	12	14
MPJPE	51.70	51.78	**51.65**	51.85	51.76

Table 6: Ablation study each component used in our **single-model** training method under MPJPE and P-MPJPE (mm) on Human3.6M. Split xy/z represents separating the 3D loss into xy and z components.

Method	OHKLM	Split xy/z	MPJPE	P-MPJPE
A	-	-	52.2	41.8
B	✓	-	51.7	41.4
C	-	✓	51.9	41.4
D	✓	✓	**51.6**	**41.3**

SimpleBaseline (SB) [29], 0.4mm for IGANet [44], 0.2mm for VPose [35], and 0.6mm for MixSTE [51], demonstrating effectiveness in enhancing robustness and generalization.

4.4 Ablation Study

To evaluate the impact and performance of each component in our training method, we assess their effectiveness in this section. The ablation study uses RTMPose [15] as the 2D pose detector on Human3.6M.

Single-Model. In our ablation study, we evaluate the single-model training method using SB [29] as the 3D lifting network. Table 5 presents results for different numbers of hard keypoints for lifting (K) in OHKLM, showing that the best performance is achieved with $K = 10$. This configuration reaches an MPJPE of 51.65mm, which we use in our experiments. Table 6 summarizes the effectiveness of OHKLM and the 3D loss divided into xy and z-directions (Split xy/z). OHKLM and Split xy/z improve MPJPE by 0.5mm and 0.3mm, respectively, with a combined improvement of 0.6mm when both are used, demonstrating their effectiveness in enhancing 3D lifting accuracy.

Multi-Model. In the ablation study using our multi-model training method, we utilize four 3D lifting networks during training and report the average performance across all networks. The results are summarized in Tab. 7. We investigate the contributions of model-wise dropout, local consistency loss, and global consistency loss. Applying model-wise dropout prevents overfitting by randomly excluding lifting networks during training, improving the average MPJPE by

Table 7: Ablation study each component used in our **multi-model** training method under MPJPE (mm) on Human3.6M.

Method	Model-wise Dropout	Local Cons Loss	Global Cons Loss	MPJPE	P-MPJPE
A	-	-	-	51.2	40.8
B	-	✓	-	50.8	40.5
C	-	-	✓	51.1	40.9
D	-	✓	✓	50.8	40.5
E	✓	-	-	51.0	40.7
F	✓	✓	-	**50.6**	40.6
G	✓	-	✓	50.9	40.8
H	✓	✓	✓	**50.6**	**40.3**

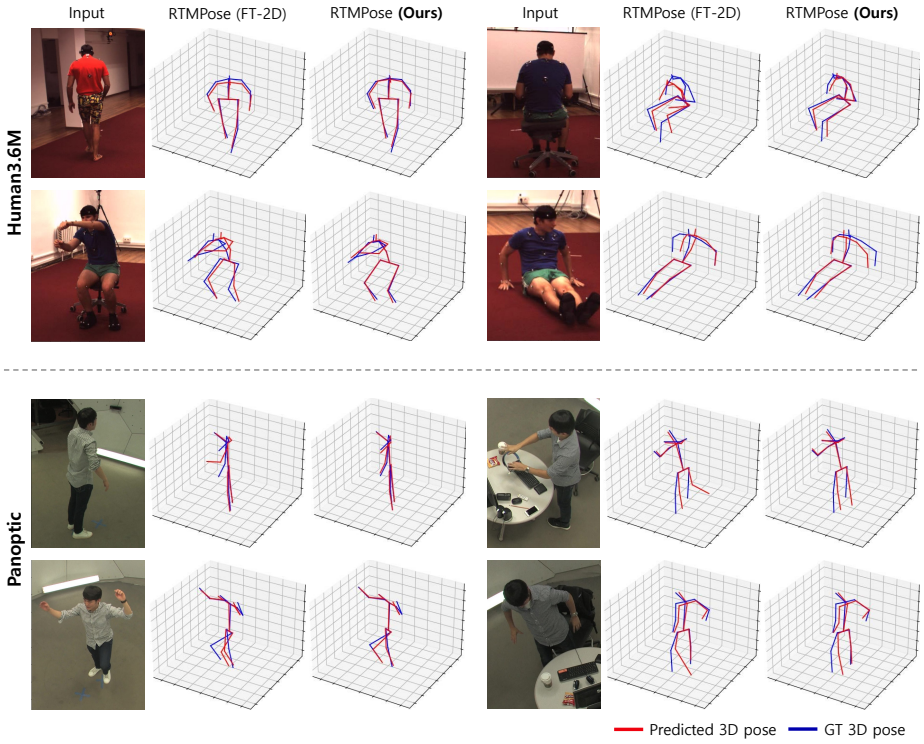

Fig. 4: Qualitative comparisons with the traditional training method (FT-2D) with RTMPose [15] on Human3.6M (top) and Panoptic (bottom). The blue lines represent GT 3D poses, while the red lines indicate 3D poses lifting from the predicted 2D poses.

0.2mm compared to the multi-model training without model-wise dropout. Incorporating both local and global consistency losses demonstrates their effectiveness in enhancing generalization and performance, improving the average MPJPE by 0.6mm.

4.5 Qualitative Results

Finally, we show the qualitative comparisons with the traditional training method with RTMPose [15] on Human3.6M and Panoptic in Fig. 4. The presented results are based on a multi-model training method with four 3D lifting networks and testing on SB [29]. Our training method demonstrates superior performance in producing more accurate and reasonable 3D poses when lifting from the predicted 2D poses, especially in challenging scenarios. For example, in the first row and second column of Human3.6M, our method predicts a more accurate 3D pose for ambiguous depth, such as upper body tilt, compared to FT-2D. In the second column of Panoptic, our method shows greater robustness to self-occlusion than FT-2D.

5 Conclusion

In this paper, we proposed a novel 3D-guided training method for the 2D pose detector to enhance 3D HPE. By integrating 2D and 3D losses and introducing the OHKLM loss, we combine the 2D pose detector and 3D lifting network, allowing for integrated training. Our training method enables the 2D pose detector to leverage the rich information from 3D lifting networks and 3D data, thereby improving the accuracy of 3D lifting. Additionally, our multi-model training approach, incorporating model-wise dropout and consistency loss, demonstrated superior generalization across various 3D lifting networks. The experimental results on the Human3.6M and Panoptic datasets validate the robustness and effectiveness of our approach, highlighting its potential for broader application in 3D HPE tasks. For future work, we plan to extend our 3D-guided training method to incorporate temporal information, making it applicable to video-based models. We hope this work can inspire future research in the 2D-to-3D lifting approach for 3D HPE.

Acknowledgments. This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [24ZD1140, Regional Industry ICT Convergence Technology Advancement and Support Project in Daegu-Gyeonguk (Medical)] and by the Commercialization Promotion Agency for R&D Outcomes (COMPA) grant funded by the Korean Government (Ministry of Science and ICT) (RS-2023-00304695).

References

1. Baumgartner, T., Klatt, S.: Monocular 3d human pose estimation for sports broadcasts using partial sports field registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5109–5118 (2023) 2
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017) 1, 4
3. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4733–4742 (2016) 3
4. Caruana, R.: Multitask learning. Machine learning **28**, 41–75 (1997) 4
5. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018) 6
6. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015) 1
7. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 109–117 (2004) 4
8. Fan, X., Zheng, K., Lin, Y., Wang, S.: Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1347–1355 (2015) 3
9. Gkioxari, G., Hariharan, B., Girshick, R., Malik, J.: R-cnns for pose estimation and action detection. arXiv preprint arXiv:1406.5212 (2014) 3, 4
10. Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13041–13051 (2023) 2
11. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012) 7
12. Huo, R., Gao, Q., Qi, J., Ju, Z.: 3d human pose estimation in video for human-computer/robot interaction. In: Yang, H., Liu, H., Zou, J., Yin, Z., Liu, L., Yang, G., Ouyang, X., Wang, Z. (eds.) Intelligent Robotics and Applications. pp. 176–187. Springer Nature Singapore, Singapore (2023) 1
13. 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) 3, 8, 9, 11
14. Jahangiri, E., Yuille, A.L.: Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 805–814 (2017) 4
15. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: Rtmpose: Real-time multi-person pose estimation based on mmpose. arXiv preprint arXiv:2303.07399 (2023) 1, 3, 4, 6, 8, 10, 11, 12, 13, 14
16. 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) 3, 8, 9, 10

17. Karthikeyan, A., Ren, R., Kant, Y., Gilitschenski, I.: Avatarone: Monocular 3d human animation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3647–3657 (2024) 1
18. Kumar, A., Daume III, H.: Learning task grouping and overlap in multi-task learning. arXiv preprint arXiv:1206.6417 (2012) 4
19. Lee, K., Lee, I., Lee, S.: Propagating lstm: 3d pose estimation based on joint interdependency. In: Proceedings of the European conference on computer vision (ECCV). pp. 119–135 (2018) 4
20. Li, C., Lee, G.H.: Generating multiple hypotheses for 3d human pose estimation with mixture density network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9887–9895 (2019) 4
21. Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., Lu, C.: Human pose regression with residual log-likelihood estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11025–11034 (2021) 3
22. Li, Y., Yang, S., Liu, P., Zhang, S., Wang, Y., Wang, Z., Yang, W., Xia, S.T.: Simcc: A simple coordinate classification perspective for human pose estimation. In: European Conference on Computer Vision. pp. 89–106. Springer (2022) 3
23. Lin, H.Y., Chen, T.W.: Augmented reality with human body interaction based on monocular 3d pose estimation. In: Blanc-Talon, J., Bone, D., Philips, W., Popescu, D., Scheunders, P. (eds.) Advanced Concepts for Intelligent Vision Systems. pp. 321–331. Springer Berlin Heidelberg, Berlin, Heidelberg (2010) 1
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 8
25. Liu, S., Johns, E., Davison, A.J.: End-to-end multi-task learning with attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1871–1880 (2019) 4, 7
26. Luvizon, D.C., Picard, D., Tabia, H.: 2d/3d pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5137–5146 (2018) 2, 5
27. Luvizon, D.C., Picard, D., Tabia, H.: Multi-task deep learning for real-time 3d human pose estimation and action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(8), 2752–2764 (2021) 1, 3, 4
28. Luvizon, D.C., Tabia, H., Picard, D.: Human pose regression by combining indirect part detection and contextual information. Computers & Graphics **85**, 15–22 (2019) 2, 5
29. 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) 1, 2, 3, 4, 8, 9, 12, 14
30. Mehraban, S., Adeli, V., Taati, B.: Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6920–6930 (2024) 1, 4
31. Mehta, D., Rhodin, H., Casas, D., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation using transfer learning and improved cnn supervision. arXiv preprint arXiv:1611.09813 **1**(3), 5 (2016) 4
32. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 483–499. Springer (2016) 3

33. Nibali, A., He, Z., Morgan, S., Prendergast, L.: 3d human pose estimation with 2d marginal heatmaps. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1477–1485. IEEE (2019) 2, 4, 5
34. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7025–7034 (2017) 1, 4
35. Pavllo, 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) 1, 3, 8, 9, 12
36. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Feichtenhofer, C., Malik, J.: On the benefits of 3d pose and tracking for human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 640–649 (June 2023) 1
37. 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) 1, 3, 4, 6, 8, 10
38. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE international conference on computer vision. pp. 2602–2611 (2017) 1, 4
39. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European conference on computer vision (ECCV). pp. 529–545 (2018) 1, 4
40. Svenstrup, M., Tranberg, S., Andersen, H.J., Bak, T.: Pose estimation and adaptive robot behaviour for human-robot interaction. In: 2009 IEEE International Conference on Robotics and Automation. pp. 3571–3576 (2009) 1
41. Takahashi, K., Mikami, D., Isogawa, M., Kimata, H.: Human pose as calibration pattern; 3d human pose estimation with multiple unsynchronized and uncalibrated cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1775–1782 (2018) 2
42. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1653–1660 (2014) 3
43. Tu, H., Wang, C., Zeng, W.: Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 197–212. Springer (2020) 9
44. 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) 1, 3, 4, 8, 9, 12
45. Wang, Y., Wang, Z., Li, M., Yan, H.: 3d human pose estimation with two-step mixed-training strategy. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3332–3341 (2024) 2
46. Yan, Y., Ricci, E., Subramanian, R., Liu, G., Lanz, O., Sebe, N.: A multi-task learning framework for head pose estimation under target motion. IEEE transactions on pattern analysis and machine intelligence **38**(6), 1070–1083 (2015) 3, 4
47. 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) 1, 3, 4, 6, 8, 10

48. Ye, H., Zhu, W., Wang, C., Wu, R., Wang, Y.: Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In: European Conference on Computer Vision. pp. 142–159. Springer (2022) 9
49. Yu, B.X., Zhang, Z., Liu, Y., Zhong, S.h., Liu, Y., Chen, C.W.: Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8818–8829 (2023) 2, 4
50. Zhang, J., Cai, Y., Yan, S., Feng, J., et al.: Direct multi-view multi-person 3d pose estimation. Advances in Neural Information Processing Systems **34**, 13153–13164 (2021) 9
51. 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) 1, 2, 3, 4, 8, 9, 12
52. Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering **34**(12), 5586–5609 (2021) 4
53. Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M.: Deep learning-based human pose estimation: A survey. ACM Computing Surveys **56**(1), 1–37 (2023) 1
54. Zhou, F., Yin, J., Li, P.: Lifting by image–leveraging image cues for accurate 3d human pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7632–7640 (2024) 2
55. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: Proceedings of the IEEE international conference on computer vision. pp. 398–407 (2017) 1, 4
56. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15085–15099 (2023) 1, 2, 4
57. Zimmermann, C., Welschehold, T., Dornhege, C., Burgard, W., Brox, T.: 3d human pose estimation in rgbd images for robotic task learning. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 1986–1992. IEEE (2018) 2