AugData Distillation for Monocular 3D Human Pose Estimation

Jiman Kim
Samsung Research
Seoul R&D Compus, Republic of Korea
jiman14.kim@samsung.com

Abstract

A large amount of data is necessary to lift the 2D human pose to the correct 3D pose, but the available public data is very limited. In particular, since monocular-based algorithms use only limited visual information acquired from one viewpoint, the amount of data is much smaller than that of multi-view. To overcome this problem, 2D-3D pair augmentation methods have been proposed, but they mainly focus on increasing the amount. However, recent research shows that quality rather than quantity significantly impacts performance improvement. This paper proposes AugData Distillation (ADD), which can dramatically reduce the 3D human pose estimation errors with only a small amount of augmentation by simultaneously considering the quality and quantity of training data. Quality distillation selects core data that significantly contributes to performance improvement among all augmented data. The total amount of augmentation is adjusted through scale distillation. These processes remove meaningless data and enable the 3D pose estimator to train core information. We selected TAG-Net [15] as the baseline model to verify the performance improvement in the data-centric method. Although it is not the top rank in all 3D HPEs, the algorithm achieved the highest accuracy in the monocular data-centric method. Experimental results show that our approach reduced a baseline method's 3D human pose estimation error by 22% with only 1.6 times augmentation. This means that most of the baseline model's augmented data used for training adversely affects performance improvement. A much lower estimation error can be expected if the ADD is combined with various latest network architectures.

1. Introduction

3D human pose estimation (HPE) is fundamental in applications such as action recognition, behavior analysis, human-computer interaction, image understanding, and context recognition. Monocular cameras have high ease of use, but it is difficult to obtain depth information, so self-occlusion and camera viewpoint change significantly reduce estimation accuracy. That means poor generalization for new poses or unknown environmental conditions. Recent works overcome these challenging problems through two approaches. One is a model-centric approach, which improves deep network architecture or learning strategy to extract and learn more efficient information from given training data. It introduced a Pictorial structure model (PSM) [1, 2, 12, 27, 30], a Graph natural network (GNN) [5, 22, 42], a transformer [8, 16, 23, 41, 43, 44], or a diffusion model [32]. Also, it uses attention mechanisms [18] and temporal information [18, 43] or combines top-down and bottom-up models. However, since the amount of information that can be extracted from specific data is limited, the degree of performance improvement is insignificant compared to improvement efforts. The other is a data-centric approach that increases the diversity of training data through data augmentation methods. 3D human pose estimation usually consists of estimating 2D poses from an RGB image and converting it into 3D poses, which is lifting. Unlike 2D pose estimation, available public data for

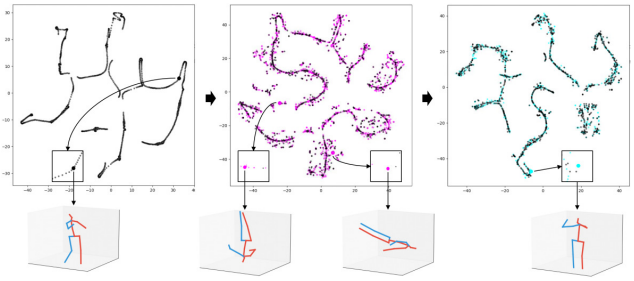

Figure 1. t-SNE [21] results of three datasets. As the data components change, the distribution shape of the feature space also changes. The original training data and the corresponding 3D pose to a particular point (left). Augmented 'magenta' data for the original 'black' data; 3D poses of meaningless data show a significant difference from the original data (center). Meaningful 'cyan' data remaining after AugData Distillation and the corresponding 3D pose (right).

the lifting is very limited. Therefore, data augmentation methods specialized in 2D-3D pose pair generation are essential to overcome this problem. TAG-Net [15] introduced evolutionary operators to combine initial poses and generate large training data. PoseAug [6] utilized additional information such as camera view and human-scale variation and alleviated the dependency between joints to expand the diversity of generated data. However, since these methods focus only on the quantity rather than the quality of the augmented training data, they include lots of meaningless data that is unnecessary for learning deep networks.

In this paper, we propose a novel method for selective data augmentation. Dataset distillation [39] compresses the entire data into a small amount of representative data. Meanwhile, our ADD carefully selects core data without any data transformation. It dramatically reduces the estimation error compared to the baseline methods. Our contributions are summarized as follows:

- We propose a novel data-centric method that selects only core data directly related to performance improvement among augmented data.
- The proposed method is a design that optimizes the quality and quantity of training data simultaneously, which increases overall training efficiency.
- Our method achieves meaningful improvement with a small amount of distilled data in monocular 3D human pose estimation.

2. Related Works

Model-centric 3D HPE. Given well-annotated data, 3D human pose estimation methods can be divided into end-to-end and two-stage manners. The end-to-end manner directly estimates the 3D pose from a monocular RGB image without intermediate 2D representation [16, 18, 28, 36, 37]. This manner consists of a single model, but it has a high computational cost and needs a considerable amount of 3D pose data for network learning. Two-stage manner estimates 2D pose in RGB image [4, 20, 35] and lifts them to 3D coordinate [5, 17, 19, 23, 32, 40, 42–44]. Recent works for 2D human pose estimation show very high accuracy. Since the two-stage manner leverages reliable 2D key points, its 3D pose estimation performance outperforms the end-to-end manner. However, like the end-to-end manner, there is not enough 3D human pose data for network learning. Eventually, in both manners, 3D human pose data became a critical bottleneck for performance improvement. In other words, if the 3D pose data problem is solved, significant improvement is possible.

Data-centric 3D HPE. Data augmentation is a representative data-centric approach. It improves the generalization ability of the 3D pose estimator by increasing the diversity of training data instead of focusing on designing complex network architectures or learning schemes. Previous works have deformed original training images [31] or created new images through synthesis [3, 29, 38] to obtain the diversity of images. Recent works have secured the diversity of 3D pose by modifying the 3D skeleton information, not the image. TAG-Net [15] generates a large amount of training data from original 2D-3D pairs by introducing evolutionary operators performing partial skeleton recombination and joint angle perturbation. Moreover, it limits changes in joint angle and viewpoint to ensure the plausibility of data. PoseAug [6] jointly optimizes the 3D pose estimator and the data augmentor. The augmentor utilizes position, body size, and viewpoint information for new pose generation and uses the pose estimator's error as a feedback signal. Also, it reduces the mutual dependency between joints to expand the diversity of the pose. Both TAG-Net [15] and PoseAug [6] mainly focus on increasing the diversity and amount of training data. Thus, the meaningless data generated in the augmentation process limits performance improvement. Multiple 3D pose hypotheses [11, 13, 33] represent a similar effect to data augmentation. It computes the full posterior distribution of the feasible 3D poses for one 2D pose and generates various 3D hypotheses from the distribution. However, it does not directly generate new 2D-3D pair data to learn the lifting network.

From this analysis, we can see that the performance of both two approaches depends heavily on 3D human pose data. Expanding the diversity of original training data is undoubtedly helpful in solving the data limitations. However, we should remove the meaningless data generated in the augmentation process to maximize performance improvement. Our proposed method efficiently selects only core data directly related to estimation performance among augmented data. Furthermore, the joint optimization of the quality and quantity of training data drastically reduces 3D human pose estimation errors.

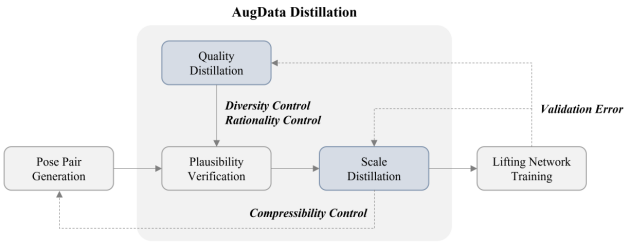

Figure 2. Overview of the ADD process. It maximizes the quality of training data and minimizes the total amount of augmentation.

3. AugData Distillation

3.1. Overall Structure

Knowledge distillation [7, 9, 26] generates a lightweight student network that imitates the core characteristics of a

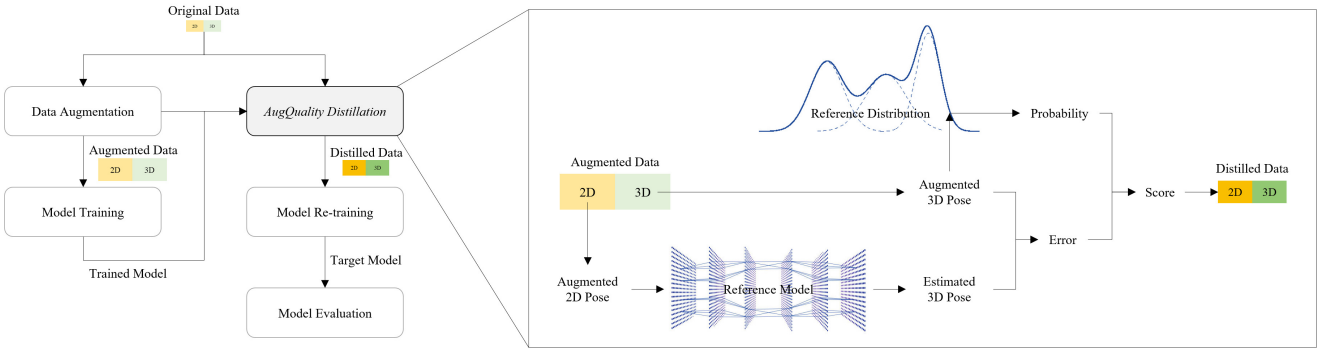

Figure 3. Changes in the lifting learning process. The model learning process with existing data augmentation (left). After applying the proposed ADD, the model learning process creates refined augmented data and re-learns models using pre-trained models and original data (right).

teacher network. On the other hand, our ADD generates a high-quality training dataset. The high-quality dataset consists of only meaningful data that is directly helpful in improving performance. ADD is entirely independent of other processes for 3D pose estimation, as shown in Figure 2. Therefore, the proposed idea can be easily combined with various lifting network architectures and pose augmentation methods. Figure 2 shows the overall process of our method. *Pose Pair Generation* augments original pose data, and *Plausibility Verification* determines whether the joint angles of the generated pose are valid. *Quality Distillation* intensively develops this verification part to select only meaningful data. In other words, the goal of the proposed method is to maximize the quality of the training dataset. *Lifting Network Training* learns the deep network using the distilled dataset. *Scale Distillation* adjusts the number of iterations for pose generation to minimize the total amount of augmentation.

3.2. Quality Distillation

Diversity. If data diversity increases, then the generalization ability of a lifting network for new poses also improves. To evaluate the diversity of augmented 3D poses, we generate a reference distribution from the original training data (Figure 3). The reference distribution represents how the original training data is scattered in the feature space. Therefore, we can evaluate how well the augmented new pose data fits into that distribution to assess the characteristics of the new data. In other words, the reference distribution serves as a criterion for judging how similar or new the new pose data is compared to the existing data. The reference distribution is created only once at the beginning and is not updated. The detailed process for implementing it is as follows: Each pose data consists of a 51-dimensional vector; 17 joints $\times$ three coordinates. For efficient calculation, we reduce its dimension to 2-dimensions using PCA [34]. Moreover, we use a Variational Bayesian Gaussian Mixture Model (VBGMM) to approximate the reference distribution. Since the model automatically sets the number of components based on the Dirichlet process prior, it is suitable for dealing with dynamic distribution. Data diversity is the probability that the augmented 3D pose belongs to the reference distribution and is defined as

$$prob_i = p(V_i|\theta_D), \tag{1}$$

where V_i is a 2-dimensional vector of ith augmented pose and θ_D is the parameters of the reference distribution. The low probability value means that the augmented pose is less similar to the poses of the original training data.

Rationality. Data augmentation generates new 2D-3D pose pairs. To evaluate the rationality of the augmented 3D pose, a lifting model already trained is used as a reference model (Figure 3). Since obtaining a model with perfect performance is unrealistic, we define the lifting model trained on the original data as the reference model. The reference model is a criterion for judging how well the augmented data's 2D pose matches the corresponding 3D pose. In other words, the reference model can be used to approximately validate errors that may occur when projecting augmented 3D poses into 2D poses while acquiring new training data. Like the reference distribution, the reference model is created only once at the beginning and is not updated. The detailed process for implementing it is as follows: We use the augmented 2D pose to input the reference model and perform the lifting. We compute the error between the reference model's output and the augmented 3D pose that acts as a virtual ground truth as

$$error_i = |f(X_i^{2d}, \theta_M) - X_i^{3d}|, \tag{2}$$

where (X_i^{2d}, X_i^{3d}) is the ith augmented pose pair, f is the function of a reference model, and θ_M is the parameters of

Table 1. Comparison of 3D human pose estimation performance in human3.6M data on 2D pose detector input.

Protocol #1	**Dir.**	**Disc.**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Pur.**	**Sit**	**SitD.**	**Smoke**	**Wait**	**WalkD.**	**Walk**	**WalkT.**	**Avg.**
TAG-Net [15]	45.6	44.6	49.3	49.3	52.5	58.5	46.4	44.3	53.8	67.5	49.4	46.1	52.5	41.4	44.4	49.7
ADD	37.2	45.3	38.5	43.6	45.1	54.6	45.0	45.4	47.1	59.9	44.4	46.4	50.9	38.8	41.1	45.5
	(↓18.4%)	(↑1.6%)	(↓21.9%)	(↓11.6%)	(↓14.1%)	(↓6.7%)	(↓3.0%)	(↑2.5%)	(↓12.5%)	(↓11.3%)	(↓10.1%)	(↑0.7%)	(↓3.0%)	(↓6.3%)	(↓7.4%)	(↓8.5%)
Protocol #2	**Dir.**	**Disc.**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Pur.**	**Sit**	**SitD.**	**Smoke**	**Wait**	**WalkD.**	**Walk**	**WalkT.**	**Avg.**
TAG-Net [15]	34.2	34.6	37.3	39.3	38.5	45.6	34.5	32.7	40.5	51.3	37.7	35.4	39.9	29.9	34.5	37.7
ADD	25.8	32.2	28.2	32.5	31.4	39.9	30.1	32.5	34.0	43.7	32.8	31.9	37.4	28.3	30.7	32.8
	(↓24.6%)	(↓6.9%)	(↓24.4%)	(↓17.3%)	(↓18.4%)	(↓12.5%)	(↓12.8%)	(↓0.6%)	(↓16.0%)	(↓14.8%)	(↓13.0%)	(↓9.9%)	(↓6.3%)	(↓5.4%)	(↓11.0%)	(↓13.0%)

Table 2. Comparison of results when ground truth 2D keypoints are used as inputs for the lifting model.

Protocol #1 (2D-GT)	**Dir.**	**Disc.**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Pur.**	**Sit**	**SitD.**	**Smoke**	**Wait**	**WalkD.**	**Walk**	**WalkT.**	**Avg.**
TAG-Net [15]	39.7	47.1	43.8	46.3	49.3	57.2	45.3	50.3	50.3	81.3	49.6	47.3	54.4	40.3	43.3	49.7
ADD	34.0	37.4	32.3	38.0	38.1	48.3	38.6	34.4	37.3	45.6	37.4	37.8	39.7	29.9	33.0	37.5
	(↓14.4%)	(↓20.6%)	(↓26.3%)	(↓17.9%)	(↓22.7%)	(↓15.6%)	(↓14.8%)	(↓31.6%)	(↓25.8%)	(↓43.9%)	(↓24.6%)	(↓20.1%)	(↓27.0%)	(↓25.8%)	(↓23.8%)	(↓24.5%)
Protocol #2 (2D-GT)	**Dir.**	**Disc.**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Pur.**	**Sit**	**SitD.**	**Smoke**	**Wait**	**WalkD.**	**Walk**	**WalkT.**	**Avg.**
TAG-Net [15]	30.2	35.4	31.1	35.0	34.6	41.9	32.9	34.3	36.5	50.9	36.0	34.9	40.0	30.7	34.1	35.9
ADD	24.1	27.5	24.7	28.9	28.0	36.8	27.2	25.4	28.4	34.6	28.7	27.2	30.6	21.7	25.9	28.0
	(↓20.2%)	(↓22.3%)	(↓20.6%)	(↓17.4%)	(↓19.1%)	(↓12.2%)	(↓17.3%)	(↓25.9%)	(↓22.2%)	(↓32.0%)	(↓20.3%)	(↓22.1%)	(↓23.5%)	(↓29.3%)	(↓24.0%)	(↓22.0%)

the model. A high error means that the augmented 2D-3D pose pair is unreasonable from the lifting model's perspective.

Score. Finally, we compute the score to measure the meaningless degree of augmented poses. The score of ith augmented pose considers the rationality and diversity of each augmented pose simultaneously. It is defined as a weighted linear combination of $(1-prob_i)$ and $error'_i$, where $error'_i$ means a normalized value between 0 and 1. The high score value means there is a high probability that it is meaningless data. Through multi-scale search and evaluation, we select 80% of augmented data with low scores.

3.3. Scale Distillation

The data augmentation is repeated until the total amount of the training dataset reaches the user-defined target value. For example, TAG-Net secures 5 times the initial training data. Each iteration generates a new augmented distribution by applying the augmentation algorithm, and the output is set as a new initial distribution. In our process, Quality distillation is added after the augmentation algorithm, and the distilled distribution becomes a new initial distribution in the next iteration. We assumed that the iterations might again generate redundant data after our distillation.

We performed a multi-scale search to find an optimal multiplier showing the best result, with Quality distillation fixed at 80%. The initial search scale is one, which is reduced to 1/10. This means we compared augmented results from 1.0 to 10.0 times and performed the search from 0.1 to 0.9, giving the best N.0 time. In our experiment, we performed the search in 3 steps up to a 1/100 scale. The best performance improvement was achieved at 1.60 times the original dataset. It proves that a small number of high-quality data is much more effective in improving performance than a large amount of low-quality data. If we introduce Bayesian optimization to optimize the quality and scale simultaneously, we can find more precise values of two hyper-parameters.

4. Experiments

4.1. Dataset and Evaluation Metrics

We used the Human3.6M [10] dataset for quantitative performance comparison with previous methods, which is most commonly used in 3D human pose estimation. We followed the same data configuration and testing processes as the previous methods [6, 13–15, 22, 24, 25, 33]. Human3.6M consists of 3.6 million 3D human poses and corresponding images taken in an indoor environment, and there are 17 scenarios of 11 professional actors. In addition, accurate 2D and 3D joint coordinate values obtained from the four calibrated cameras are provided. Skeleton topology, representing the 3D human body, consists of 17 joints. S1, S5, S6, S7, and S8 are used for the training, and S9 and S11 are used for the test. All estimation errors are measured in two ways: Protocol #1 and Protocol #2. Protocol #1 is calculated in mm units through Mean Per Joint Position Error (MPJPE) without a separate rigid alignment. Protocol #2 is described as Procrustes Analysis MPJPE (P-MPJPE, PA-MPJPE) because it includes the rigid alignment based on a root joint. Each Protocol uses two inputs: ground truth 2D keypoints and the results of 2D joint detection.

4.2. 3D HPE Results

We compared the estimation error (mm) and computed the error reduction degree (%). As shown in Table 1 and Table 2, the proposed method reduced the estimation error of the baseline method to a considerable extent, from 8.5% to

Table 3. Comparison with various data-centric approaches. Items that do not contain the detailed value in the paper are marked as '–'.

Protocol #1	**Dir.**	**Disc.**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Pur.**	**Sit**	**SitD.**	**Smoke**	**Wait**	**WalkD.**	**Walk**	**WalkT.**	**Avg.**
VNect [24]	62.6	78.1	63.4	72.5	88.3	63.1	74.8	106.6	138.7	78.8	93.8	73.9	55.8	82.0	59.6	80.5
SIM [22]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
WSGAN [14]	62.0	69.7	64.3	73.6	75.1	84.8	68.7	75.0	81.2	104.3	70.2	72.0	75.0	67.0	69.0	73.9
MDN [13]	43.8	48.6	49.1	49.8	57.6	61.5	45.9	48.3	62.0	73.4	54.8	50.6	56.0	43.4	45.5	52.7
MultiPoseNet with Oracle [33]	48.6	54.5	54.2	55.7	62.6	72.0	50.5	54.3	70.0	78.3	58.1	55.4	61.4	45.2	49.7	58.0
PoseAug [6]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	50.2
TAG-Net [15]	45.6	44.6	49.3	49.3	52.5	58.5	46.4	44.3	53.8	67.5	49.4	46.1	52.5	41.4	44.4	49.7
GraphMDN [25]	40.0	43.2	41.0	43.4	50.0	53.6	40.1	41.4	52.6	67.3	48.1	44.2	44.9	39.5	40.2	46.2
ADD	37.2	45.3	38.5	43.6	45.1	54.6	45.0	45.4	47.1	59.9	44.4	46.4	50.9	38.8	41.1	45.5
Protocol #2	**Dir.**	**Disc.**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Pur.**	**Sit**	**SitD.**	**Smoke**	**Wait**	**WalkD.**	**Walk**	**WalkT.**	**Avg.**
SIM [22]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
WSGAN [14]	38.5	41.7	39.6	45.2	45.8	46.5	37.8	42.7	52.4	62.9	45.3	40.9	45.3	38.6	38.4	44.3
MDN [13]	35.5	39.8	41.3	42.3	46.0	48.9	36.9	37.3	51.0	60.6	44.9	40.2	44.1	33.1	36.9	42.6
MultiPoseNet with Oracle [33]	35.3	35.9	45.8	42.0	40.9	52.6	36.9	35.8	43.5	51.9	44.3	38.8	45.5	29.4	34.3	40.9
PoseAug [6]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	39.1
TAG-Net [15]	34.2	34.6	37.3	39.3	38.5	45.6	34.5	32.7	40.5	51.3	37.7	35.4	39.9	29.9	34.5	37.7
GraphMDN [25]	30.8	34.7	33.6	34.2	39.6	42.2	31.0	31.9	42.9	53.5	38.1	34.1	38.0	29.6	31.1	36.3
ADD	25.8	32.2	28.2	32.5	31.4	39.9	30.1	32.5	34.0	43.7	32.8	31.9	37.4	28.3	30.7	32.8

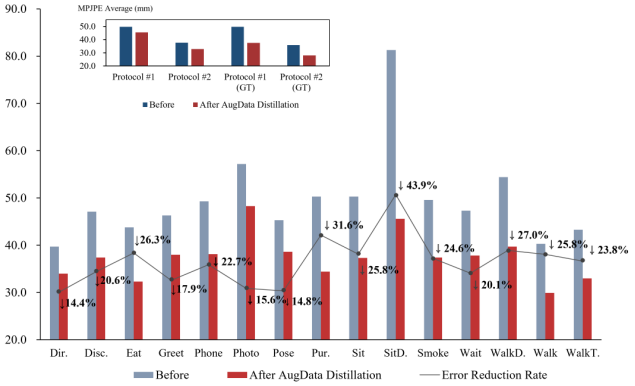

Figure 4. 3D human pose estimation error before and after ADD. The small figure on the upper left shows the average estimation errors, and the large figure shows the detailed estimation error for each joint of the last Protocol. A solid line means a degree of reduction in estimation error. The larger the error, the greater the degree of improvement.

24.5%. When using the results of a 2D pose detector as input, the improvement effect varies depending on the type of pose. Specifically, based on protocol #1 in Table 1, there was a significant improvement in most poses, but in some poses where the original 3D pose estimation accuracy was not high, the lifting results were worse. Protocol #2 did not show any cases where the results were worse, but overall, it showed a similar pattern to protocol #1. On the other hand, when using ground-truth values as input, as shown in Table 2 and Figure 4, significant performance improvements were observed in all poses regardless of the protocol, with a much greater improvement than in Table 1. Combining these results, we can conclude that (1) the higher the original accuracy of the pose, the greater the improvement effect, and (2) the higher the accuracy of the input 2D pose, the greater the performance improvement effect. From a different perspective, end-to-end approaches have the advantage that the final 3D pose estimation performance is independent of the 2D pose results, while two-stage approaches have a limitation in that the final 3D pose estimation performance depends on the performance of the 2D pose detector. The performance comparison with the existing data-centric approaches is shown in Table 3. The results are arranged in the order in which the average error of Protocol #2 is large. Our method of applying ADD shows the lowest estimation error.

Only the results of protocols presented in the paper were included in the comparison. Some methods showed different sorting results between the two protocols. The group at the top of Table 3 with large estimation errors showed a different performance order depending on the protocol, but the order was consistent regardless of the protocol for the lower group with smaller estimation errors. When the proposed method was applied, it was found that the performance improvement was greater when using ground-truth values as input. Similar to the previous results, this means that the effectiveness of the proposed method increases as the accuracy of the input 2D pose increases. From Table 3, the following conclusions can be drawn; (1) For poses with high estimation accuracy, the improvement depends on the type of algorithm rather than the measurement method, and for poses with low estimation accuracy, the improvement changes depending on the measurement method. In other words, difficult poses must be intensively improved to control the overall performance pattern. (2) To improve the results of a model consisting of 2 stages, maximizing the lifting input, that is, the 2D pose accuracy is helpful. End-to-end methods that directly estimate 3D pose from images do not have this issue but may have difficulty finding cor-

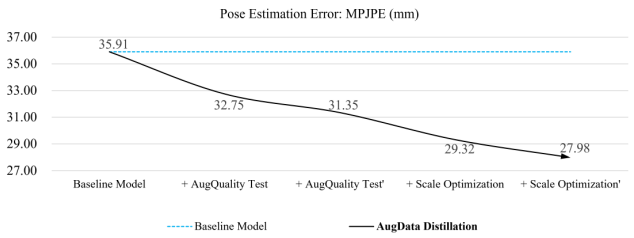

Figure 5. Changes in performance improvement before and after the quality and scale distillation. Even if a random hyper-parameter was used, applying each factor greatly improved performance. Additional performance improvements were possible by applying optimal parameters.

relations with the input when backward analyzing the final result.

Because the lifting network architecture of the baseline method is simple, we expect better results if we change it to the latest network architectures among model-centric approaches. Also, data generation efficiency and network training efficiency greatly increase when our method is applied. The number of iterations required to generate augmented data decreased to 1/3, the total amount of augmentation decreased to 1/2, and the training time decreased to 1/2.

4.3. Ablation Study

We performed the ablation study based on Protocol #2 with 2D ground truth input (Figure 5). Through the test, we can separate the effect of each factor's contribution to performance improvement. When the AugQuaility distillation was applied with a random hyper-parameter value, the estimation error decreased by 8.8% to the baseline model. The error decreased by an additional 4.3% when we optimized the hyper-parameter value. When scale distillation is applied using a random hyper-parameter value, the estimation error further decreases by 6.5%. Furthermore, the error was reduced by an additional 4.6% after the hyper-parameter optimization. In summary, the quality distillation resulted in a performance improvement of 12.7% compared to the baseline model, and the scale distillation provided an additional improvement of 10.7% for the quality distillation. If we apply only scale distillation, the estimated error naturally increases. This is because it is the same as simply reducing the amount of data without considering the quality of the data used for learning. If you randomly select 1.6 times the amount of original data from the entire augmented training data, in the worst case, only a very small number of original and augmented data that are meaningless in improving performance will be selected. As a result, performance may decrease compared to the baseline model. In other words, quality and scale distillation create synergy when used together, and quality should be considered first.

4.4. Qualitative Analysis

We performed t-SNE [21] on training data samples to analyze the distribution change in feature space. We compared the distributions before and after ADD (Figure 6, 7). The difference is revealed in the 3D feature space. Data points are more concentrated around the center point of each cluster after ADD. We found that the meaningful data that helps improve the lifting model's performance has a geometrically close relationship in the feature space. This means that filling in missing data while maintaining the shape or pattern of the existing distribution is more helpful in increasing model accuracy or reducing estimation errors than randomly increasing the diversity of original data.

Also, we compared three groups of 3D human poses (Figure 8). The first group is pose samples randomly selected from the original data. The second group is meaningful pose samples selected as core data through ADD. Most of them slightly modify the poses of the original data. The third group is samples judged as meaningless poses through ADD and excluded from the final network training. It can be seen that the ratio of samples showing a different appearance from the original data is large. As with feature space, ADD does not ignore the increase in diversity in determining the final training data, but most of it contributes to maintaining the consistency of the pose distribution included in the existing training data.

5. Conclusion

Data augmentation is useful, especially when training data is insufficient. Previous studies were proposed mainly from a quantitative point of view. Recent studies have proved that data quality is more important than the amount of data. Our proposed AugData Distillation is the first study in the field of 3D human pose estimation to maximize the quality of augmented data while minimizing the total amount of augmentation. Our method significantly reduced 3D pose estimation error. Training efficiency also increased, including data generation time and training time. The experimental results prove that small but high-quality data is a good choice for considerable performance improvement.

Limitations. While ADD selects only data directly related to performance improvement, it depends on the reference model's architecture. Therefore, if the backbone model changes, the distilled dataset and the amount of performance improvement may vary. This means it is not an optimal dataset utterly independent of the model.

Future directions. Further verification is needed to determine whether optimal subset generation is possible by combining various SOTA backbone architectures. Additionally, ADD can be applied to end-to-end methods to eliminate the dependence on the 2D pose given as input for lifting.

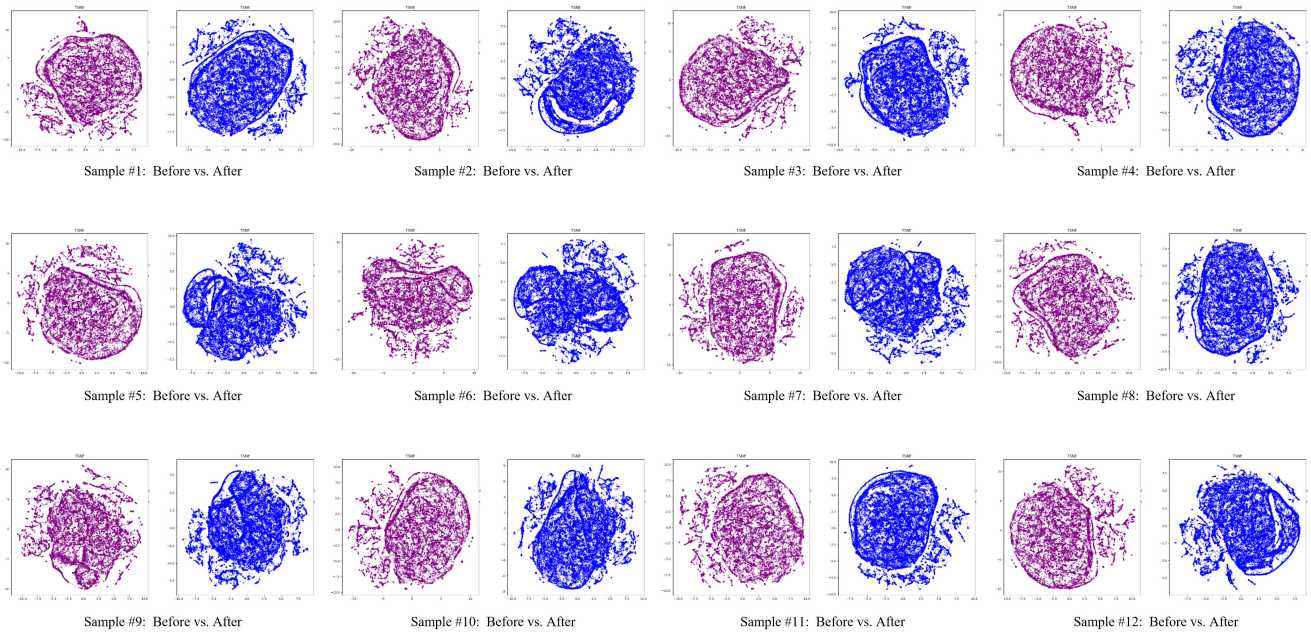

Figure 6. Training data distribution represented on the 2D feature space through t-SNE [21]. Purple means augmented training data, and blue means distilled training data after ADD. We randomly selected 5% of the samples from each dataset. We performed t-SNE by repeatedly sampling several times to analyze the shape pattern more objectively.

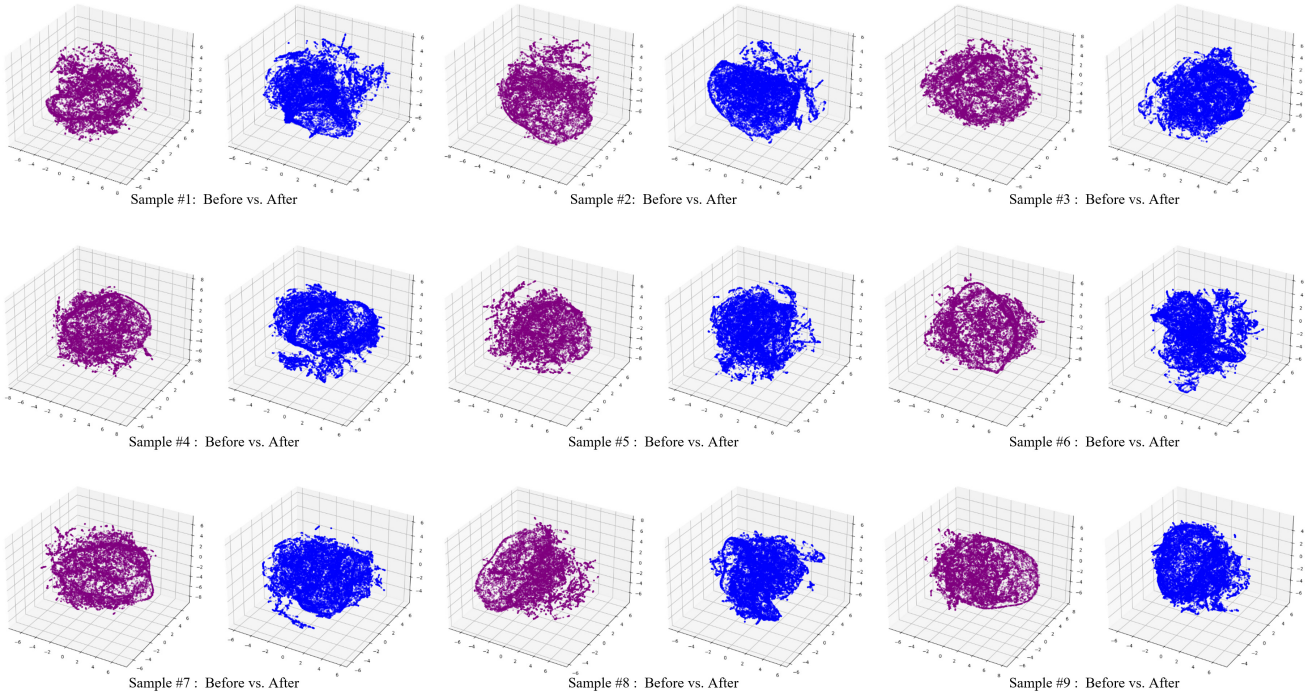

Figure 7. 3D distribution representation of two different training datasets through t-SNE [21]. Samples were selected under the same conditions as in Figure 6. The range of each axis was automatically adjusted according to the scale of the cluster. After ADD, it can be seen that the cluster density around the center has increased.

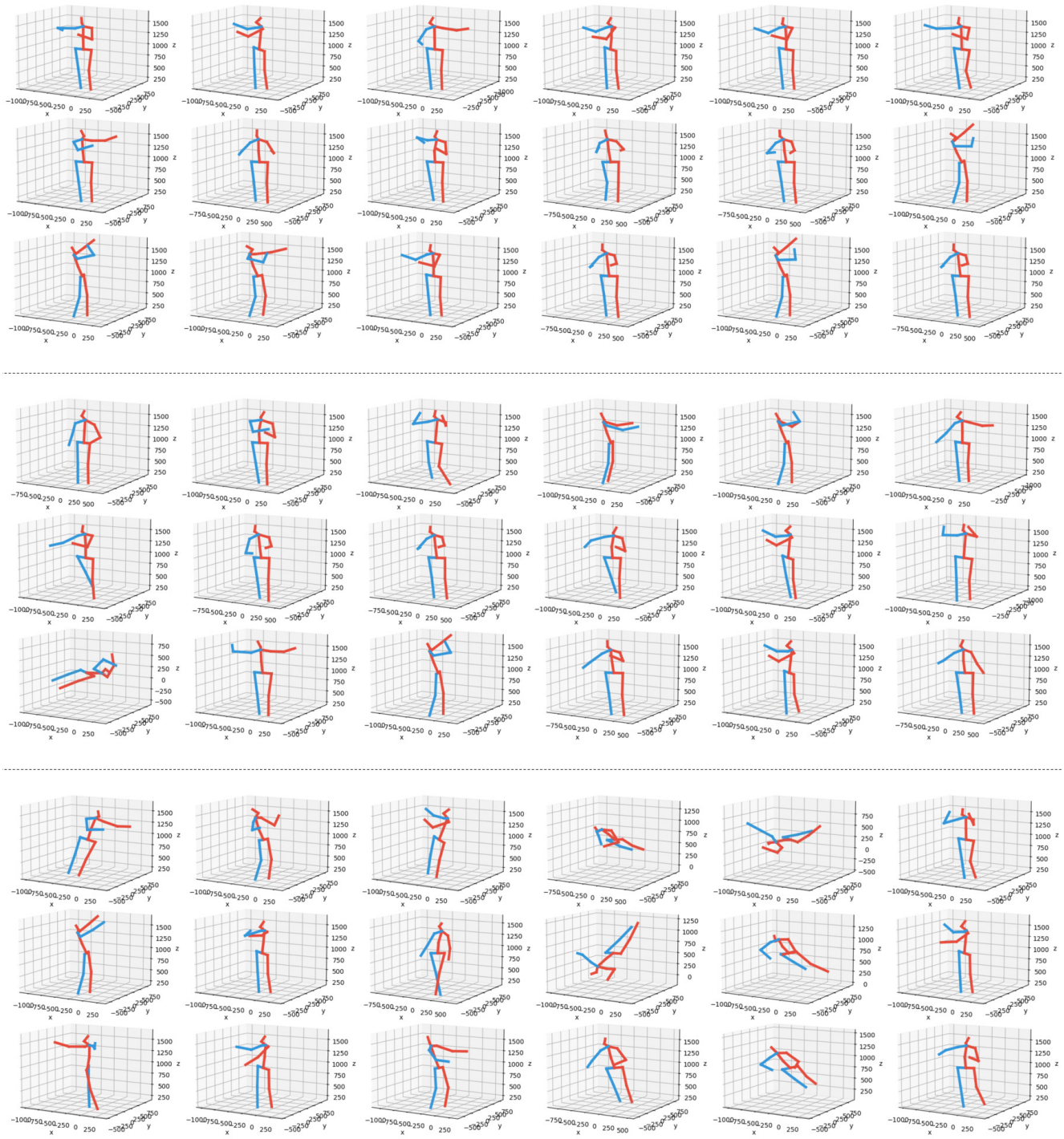

Figure 8. Examples of pose data for three groups: original data, core data after ADD, and meanless data. The poses of original data usually show a pattern in which only arms and legs move while standing upright (top). Meaningful poses judged to help improve performance show a pattern similar to the original poses (center). Meaningless poses judged by redundant data unrelated to performance improvement often show abnormal patterns: handstands and extreme joint deformation (bottom). Some of them are poses that people can take. However, if the pattern does not exist in the validation and test data, it is classified as a meaningless pose.

References

[1] Sikandar Amin, Mykhaylo Andriluka, Marcus Rohrbach, and Bernt Schiele. Multi-view pictorial structures for 3d human pose estimation. In *BMVC*, 2013. 1

[2] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *CVPR*, 2014. 1

[3] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *3DV*, 2016. 2

[4] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 2

[5] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *ICCV*, 2019. 1, 2

[6] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *CVPR*, 2021. 2, 4, 5

[7] Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey. In *IJCV*, 2021. 2

[8] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *CVPR*, 2020. 1

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPSW*, 2014. 2

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

[11] Ehsan Jahangiri and Alan L. Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *ICCVW*, 2017. 2

[12] Ilya Kostrikov and Juergen Gall. Depth sweep regression forests for estimating 3d human pose from images. In *BMVC*, 2014. 1

[13] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *CVPR*, 2019. 2, 4, 5

[14] Chen Li and Gim Hee Lee. Weakly supervised generative network for multiple 3d human pose hypotheses. In *BMVC*, 2020. 5

[15] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *CVPR*, 2020. 1, 2, 4, 5

[16] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 1, 2

[17] Kenkun Liu, Rongqi Ding, Zhiming Zou, Le Wang, and Wei Tang. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In *ECCV*, 2020. 2

[18] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *CVPR*, 2020. 1, 2

[19] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Hai Ci, and Yizhou Wang. Context modeling in 3d human pose estimation: A unified perspective. In *CVPR*, 2021. 2

[20] Xianzheng Ma, Hossein Rahmani, Zhipeng Fan, Bin Yang, Jun Chen, and Jun Liu. Remote: Reinforced motion transformation network for semi-supervised 2d pose estimation in videos. In *AAAI*, 2022. 2

[21] Laurens van der Maaten and Geoffrey Hinton. Visualizingddata using t-sne. In *Journal of Machine Learning Research*, 2008. 1, 6, 7

[22] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 1, 4, 5

[23] Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *WACV*, 2024. 1, 2

[24] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhondin, Mohammad Shafiei, Hans-peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. In *SIGGRAPH*, 2017. 4, 5

[25] Tuomas P. Oikarinen, Daniel C. Hannah, and Sohrob Kazerounian. Graphmdn: Leveraging graph structure and deep learning to solve inverse problems. In *arXiv preprint arXiv:2010.13668*, 2020. 4, 5

[26] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 2

[27] Georgios Pavlakos, Xiaowei Zhou, and Konstantinos G. Derpanis. Harvesting multiple views for markerless 3d human pose annotations. In *CVPR*, 2017. 1

[28] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017. 2

[29] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *CVPR*, 2018. 2

[30] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *ICCV*, 2019. 1

[31] Gregory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *NeurIPS*, 2016. 2

[32] Wennkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *ICCV*, 2023. 1, 2

[33] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *ICCV*, 2019. 2, 4, 5

[34] Jonathon Shlens. A tutorial on principal component analysis. In *arXiv:1404.1100*, 2014. 3

[35] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2

[36] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 2

[37] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Direct prediction of 3d body poses from motion compensated sequences. In *CVPR*, 2016. 2

[38] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2

[39] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. In *arXiv:1811.10959v3*, 2020. 2

[40] Tianhan Xu and Wataru Takano. Graph stacked hourglass networks for 3d human pose estimation. In *CVPR*, 2021. 2

[41] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Towards explainable human pose estimation by transformer. In *arXiv preprint arXiv:2012.14214*, 2020. 1

[42] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, 2019. 1, 2

[43] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *ICCV*, 2021. 1

[44] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, 2023. 1, 2