Enhancing Human Pose Estimation with Privileged Learning

Alexander Marusov[*], Mariam Kaprielova[*†], Radoslav Neychev [*‡]
[*]Moscow Institute Of Physics And Technology, Moscow, Russia
[†]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia
[‡]Moscow Institute Of Physics And Technology, Moscow, Russia
[‡]Harbour.Space University, Barcelona, Spain
{marusov.ae, kaprielova.ms, neychev}@phystech.edu

***Abstract*—Transformer architecture shows significant improvements in different applications, such as Natural Language Processing, Computer Vision and even Graph Machine Learning. Recent advances in the Human Pose Estimation (HPE) show that Vision Transformers are a great choice for this problem as well. But even state of the art architectures require additional enhancements to the training process to achieve the best results. In this paper we propose the privileged learning approach to HPE by incorporating the information about body proportions into the training pipeline. We quantitatively and qualitatively evaluate our method on the standard benchmark dataset Human3.6M. The proposed method shows stable improvements using the same model architecture as [1].**

I. Introduction

The Human Pose Estimation (HPE) task is one of the significant computer vision tasks along with such topics as object detection, object classification, face recognition, image segmentation. HPE has many practical applications. For example, modern video surveillance systems, virtual reality, sport, and medicine and a lot of other fields. In addition, HPE is used to solve the problems of human action recognition [2-3] and tracking [4-5].

The task of 3D-HPE is to determine the keypoints of the body in 3D space. The formal statement of the problem is to find the positions of K key points J = {Jk|k = 1,...,K} from the input image.

In the current work, we put forward a hypothesis on the improvement of the quality of prediction when taking into account characteristics invariant to rotation and point of view change. Moreover, we study the dependence between the constraints strictness and the quality of prediction.

II. Related works

Human 3D pose estimation has a rich history of research. The classical approach to the Human Pose Estimation problem is to use frameworks which use a predefined pose, regardless of the image data supplied to the algorithm [6,7]. Of course, such approaches are very limited, as they use predefined pose templates. With the evolution of deep neural networks (and especially CNNs), this limitation was removed. Rapid development of deep learning approaches made it possible to extract informative feature representation straightforward from the provided visual data.

After the introduction of "DeepPose"[8], many started using CNN as a backbone. Deep learning approaches can be divided into two groups. The first one includes solutions which predict key points directly, e.g. “DeepPose”, Fan et al. [9] and other works [10-11]. Another group first generates a heat map with scores for all possible points, and then predicts keypoints, e.g. works [12-15].

Many works use the refinement method. The main idea is to refine the position of keypoints after receiving some of their initial position. One of the ways to refine the position of keypoints is to take into account their relative position. So, for example, in [1], [16-18] the position of a particular keypoint is determined by the positions of its neighbors. In [1] this accounting is done using a spatial transformer, which extracts features, taking into account the correlation of the joints.

It is known that incorporation of prior knowledge in the appropriate form can significantly improve the stability and quality of the solution and increase the number of its potential applications [19]. We focus on the LUPI (Learning Using Privileged Information) paradigm proposed by Vladimir Vapnik and Rauf Izmailov [20]. 3D representation of the human body should be the same for all positions of the camera (ignoring the HPE error), only the relative sizes should change. That’s the reason why body proportions are used as invariants in this case. Hence, we use the relative positions of the specific human body keypoints as privileged information during the training stage. This constraint is incorporated into the loss function in the form of an additional penalty for the difference in the proportions of the predicted and correct poses.

III. Method

As noted above, the main idea of the proposed method is to use the prior assumptions and incorporate the privileged information on the body proportions of the subject. To do so, we add the regularization term to the main loss function. This term represents the difference between the size characteristics of the predicted and ground truth poses. The greater the discrepancy between the prediction and ground truth, the greater the penalty. We use Euclidean distance between specific keypoints (for example, shoulder width) as dimensional characteristics. In our method we took into account following properties:

- Shoulder width.
- Distance from heel to knee for both feets.

The penalty is the MSE (Mean Squared Error) between the predicted and correct size characteristics:

$$ProportionsPenalty = \frac{1}{K}\sum_{i=1}^{K} \|E_i - E_i^*\|_2,$$

where E_i stands for real size of the selected edge, and E_i^* for the predicted one. This regularization term incorporates the information about the body proportions, but also naturally takes into account the certainty of the prediction. With the increase of the distance to the tested subject the uncertainty of the prediction increases, and this regularization term reduces the penalty corssespondingly.

Then mathematically the new loss function looks like this:

$$ProposedLoss = OriginalLoss + \alpha \cdot ProportionsPenalty$$

- Here, the variable α stands for a positive real factor multiplied by the $ProportionsPenalty$ to vary the impact of the regularization term. Large alpha coefficient forces the algorithm to focus on the proportions preservation.
- $ProportionsPenalty$ – penalty for violation of proportions

Since the ground truth size characteristics are taken from the ground truth key points on the current photo, the proportions are used as prior during the training stage. There is no need for additional information on the inference stage.

This approach is similar to the LUPI paradigm [20], where the training set contains i.i.d. triplets: $(x, x^\star, y)$.

- First term x stands for the original data representation, available both during training and inference. In the selected problem this is the visual information on the human subject.
- Second term $x^\star$ stands for the privileged information, available only during the training stage. Here we use the body proportions, which are available for every subject in the training set.
- Last term y stand for the labels: human body keypoint locations.

A. OriginalLoss details

Our experiments are based on the model presented in [1]. Mean Per Joint Position Error (MPJPE) is used as loss function accordingly. MPJPE is the sum of Euclidean distances averaged over the number of keypoints between predicted J_i^* and ground truth J_i key points:

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N} \|J_i - J_i^*\|_2$$

So the new loss is:

$$ProposedLoss = MPJPE + \alpha \cdot ProportionsPenalty$$

B. Evaluation metrics details

For evaluation were used two metrics – MPJPE and P-MPJPE.

P-MPJPE aligns the estimated 3D pose to the ground-truth by a rigid transformation before computing the MPJPE:

P-MPJPE = $MPJPE(aligned_{pose}, gt_{pose})$, where

- $aligned_{pose}$ – is the result of performing scaling, rotation and transformation to the predicted pose,
- gt_{pose} – is the ground truth pose.

IV. Experiments

A. Dataset

To evaluate the proposed method we used Human3.6M dataset [21]. This dataset consists of 3.6 million video frames with ground truth annotation. Since Human3.6M has 11 professional actors in 17 different actions (like discussion, smoking, taking photos, talking on the phone etc.), then this dataset is truly diverse. Training and testing experiment settings are the same as in works [1], [22-24]. In more details for training and testing all 15 actions were used. Five subjects (S1, S5, S6, S7, S8) were selected for training and two subjects (S9, S11) – for testing.

B. Computational experiments

Two NVIDIA GeForce RTX 2080 Ti GPUs were used for model training and inference. One epoch took approximately 100 minutes, and in every experiment the model was trained until convergence (which took 12 epochs on average). Batch size was set to 256 due to the hardware limitations to achieve more stable convergence.

PoseFormer architecture was used as baseline. Hyperparameters (except alpha coefficient and initial learning rate) were inherited from the original paper.

Before training the model with the novel loss function, the best initial learning rate was selected as shown in Table I.

TABLE I. Selection of the required initial rate

Alpha	Initial learning rate	MPJPE	P-MPJPE
0	$2 \cdot 10^{-6}$	45.6	35.5
0	$2 \cdot 10^{-7}$	45.1	35.1
0	$2 \cdot 10^{-8}$	44.5	34.8
0	$2 \cdot 10^{-9}$	44.3	34.6

The initial learning rate was fixed equal to $2 \cdot 10^{-9}$. Alpha coefficient was varied.

To select the appropriate coefficient the scale of MPJPE and ProportionsPenalty was estimated. MPJPE values are about 100 times bigger than ProportionsPenalty. So the initial coefficient was 100 (and it already has shown the improvement). In further experiments we have analyzed lower and higher coefficient values. Coefficient increase led us to the almost monotonous improvement of the results. The results are shown in Table II.

TABLE II. EXPERIMENTS WITH A FIXED INITIAL LEARNING RATE AND THE PROPOSED LOSS FUNCTION. MPJPE AND P-MPJPE ARE CALCULATED OVER THE ENTIRE TEST DATA SET

Alpha	Initial learning rate	MPJPE	P-MPJPE
80	$2 \cdot 10^{-9}$	44.24	34.62
100	$2 \cdot 10^{-9}$	44.24	34.62
200	$2 \cdot 10^{-9}$	44.22	34.61
1000	$2 \cdot 10^{-9}$	44.20	34.60
5000	$2 \cdot 10^{-9}$	44.20	34.60
10000	$2 \cdot 10^{-9}$	44.20	34.60
100000	$2 \cdot 10^{-9}$	44.20	34.60

The table above shows the positive dynamics of changes in metrics with an increase in the coefficient.

Moreover, additional experiments were carried out to test the stability of the algorithm. The stabilities verification consisted in carrying out 5 experiments with different seeds. The obtained results showed that the algorithm is stable. The training pipeline was restarted five times with different random seeds. The standard deviation for final results does not exceed 10^{-3}, so we assume the convergence is stable. Further stability analysis requires additional hardware resources.

C. Results

Before training the model with the novel loss function, the best hyperparameters (learning rate and batch size) were selected.

1) Dependency of the quality of the model on the value of the coefficient: In this section we represent graphs that illustrate the dependency of the quality on the coefficient alpha. The names of the following paragraphs are built according to the following principle. First, the metric is given, then the data set, which is used to evaluate the quality of the model. For all the graphs below, the y-axis is the metric, and the x-axis is coefficients on a logarithmic scale.

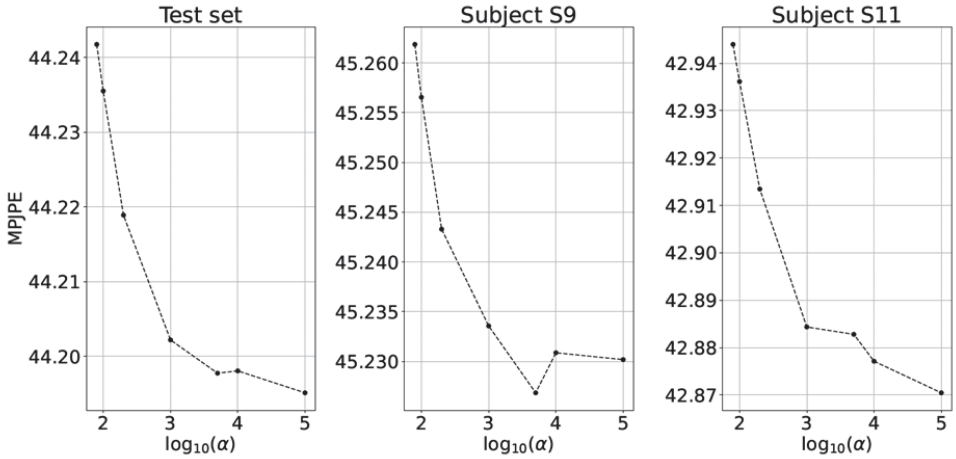

Fig. 1. Dependency of MPJPE metrics on α coefficient

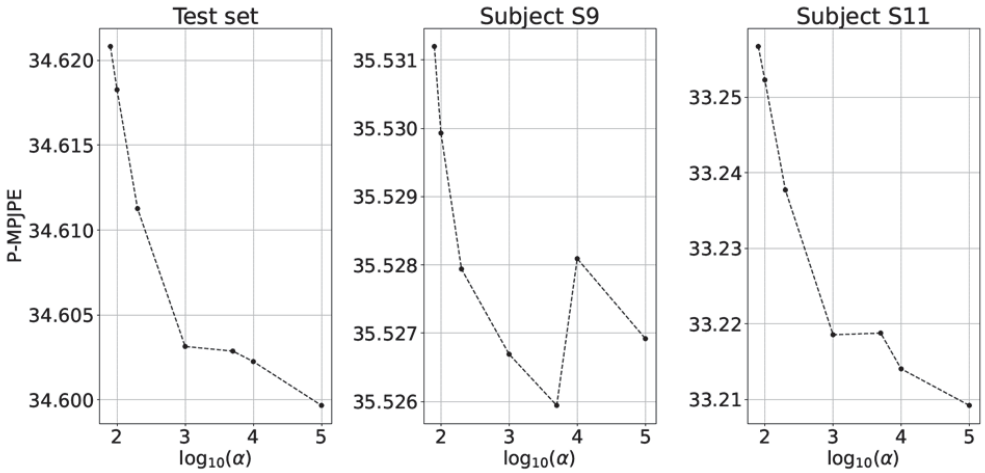

Fig. 2. Dependency of P-MPJPE metrics on α coefficient

Experimental results show that the value of the target metric improves as the coefficient increases. Moreover, the majority of experiments show monotonous improvement. Improvement occurs both on the train set and on the test set. Thus, this kind of improvement shows that increasing the contribution of the penalty has a positive effect on the performance of the model. We observethis trend for both metrics (MPJPE, P-MPJPE) and all test sets (full test set, S11, S9). Though, for some actions (e.g. SittingDown) target metric slightly increases. It might happen due to the occlusion and requires further investigation.

2) Comparison with PoseFormer: In this section we represent the comparison between the original PoseFormer and our solution. We obtained improvement of the performance for the majority of the actions represented in the dataset. We report all actions for both metrics (MPJPE and P-MPJPE) in Table III and Table IV respectively. The last row is the average metric on the test set.

TABLE III. COMPARISON OF THE RESULTS OF THE BEST MODEL WITH POSEFORMER BY MPJPE METRIC ON THE ENTIRE TEST DATASET AND FOR ALL ACTIONS

	PoseFormer	**Our**
Dir	41.5	41.4
Disc	44.8	44.7
Eat	39.8	39.7
Greet	42.5	42.5
Phone	46.5	46.4
Photo	51.6	51.6
Pose	42.1	42.1
Purch	42.0	41.8
Sit	53.3	53.3
SitD	60.7	60.8
Smoke	45.5	45.4
Wait	43.3	43.3
WalkD	46.1	45.9
Walk	31.8	31.7
WalkT	32.2	32.4
Aver	44.3	44.2

Table III shows that model quality is improved on eight actions, worsened on two actions and is not affected on the other five actions. Moreover, we were able to improve the metric on such difficult actions as *WalkDog* and *Smoking*. The model's quality worsened on *SittingDown* and *WalkTogether*.

As a result, we were able to outperform average performance of the model from [1] on the entire test dataset (average value) and for the majority of actions (calculated individually) and yield the best result 44.2mm as shown in Table III.

Table IV shows that the proposed model improved on 4 actions, stayed the same on 9 actions and worsened on 2 actions. In terms of P-MJPE we didn't obtain any improvement of average performance of the model from [1] on the entire test dataset. Nevertheless, the quality increased for actions *Directions, Discussion, Phoning* and *Purchases*, as shown in Table IV.

More specific details (e.g the metric changes for different actions) are presented in Tables V-X in the Appendix section.

TABLE IV. COMPARISON OF THE RESULTS OF THE BEST MODEL WITH POSEFORMER BY P-MPJPE METRIC ON THE ENTIRE TEST DATASET AND FOR ALL ACTIONS

	PoseFormer	**Our**
Dir	32.5	32.4
Disc	34.8	34.7
Eat	32.6	32.6
Greet	34.6	34.6
Phone	35.3	35.2
Photo	39.5	39.5
Pose	32.1	32.2
Purch	32.0	31.9
Sit	42.8	42.8
SitD	48.5	48.5
Smoke	34.8	36.4
Wait	32.4	32.4
WalkD	35.3	35.3
Walk	24.5	24.5
WalkT	26.0	26.0
Aver	34.6	34.6

V. CONCLUSION

Qualitative analysis shows that presence of the privileged information leads to stable improvements of MPJPE and P-MPJPE metrics. The proposed approach does not require any additional information during the inference stage and does not increase the complexity of the HPE solution. Moreover, it is flexible and can be used with any baseline model, including future state-of-the-art approaches.

VI. FURTHER RESEARCH

The analysis of the proposed solution shows potential in additional analysis of the solution, especially in presence of additional information. Human body proportions might be passed as additional information to the model input. This hypothesis requires additional study. Also benchmark several models with the proposed LUPI approach and compare results.

REFERENCES

[1] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers", *in Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 11656-11665.

[2] C. Wang, Y. Wang, A.L. Yuille, "An approach to pose-based action recognition", CVPR, 2013, pp.915-922.C. Wang, Y. Wang, A.L. Yuille, "An approach to pose-based action recognition", *CVPR*, 2013, pp.915-922.

[3] Z. Liang, X. Wang, R. Huang, L. Lin, "An expressive deep model for human action parsing from a single image", IEEE International Conference on Multimedia and Expo (ICME), 2014, pp.1-6.

[4] N.G. Cho, A.L. Yuille, S.W. Lee, "Adaptive occlusion state estimation for human pose tracking under self-occlusions", Pattern Recognition, 2013, pp. 649–661.

[5] B. Xiao, H. Wu, Y. Wei, "Simple baselines for human pose estimation and tracking", ECCV, 2018.

[6] L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele, "Poselet conditioned pictorial structures", CVPR, 2013, pp. 588-595.

[7] Y. Yang, D. Ramanan, "Articulated human detection with flexible mixtures of parts", IEEE transactions on pattern analysis and machine intelligence, 2012, pp. 2878–2890.

[8] A. Toshev, C. Szegedy, "Deeppose: Human pose estimation via deep neural networks", *CVPR*, 2014, pp. 1653-1660.

[9] X. Fan, K. Zheng, Y. Lin, S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation", CVPR, 2015, pp. 1347-1355.

[10] J. Carreira, P. Agrawal, K. Fragkiadaki, J. Malik, “Human pose estimation with iterative error feedback”, CVPR, 2016, pp. 4733-4742.
[11] X. Sun, B. Xiao, F. Wei, S. Liang, Y. Wei, “Integral human pose regression”, ECCV, 2018, pp. 529-545.
[12] J.J. Tompson, A. Jain, Y. LeCun, C. Bregler, “Joint training of a convolutional network and a graphical model for human pose estimation”, NeurIPS, 2014.
[13] X. Chen, A.L. Yuille, “Articulated pose estimation by a graphical model with image dependent pairwise relations”, NeurIPS, 2014.
[14] W. Yang, W. Ouyang, H. Li, X. Wang, “End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation”, CVPR, 2016, pp. 3073-3082.
[15] S.E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, “Convolutional pose machines”, CVPR, 2016, pp. 4724-4732.
[16] H. Isack, C. Haene, C. Keskin, S. Bouaziz, Y. Boykov, S. Izadi, S. Khamis, “Repose: Learning deep kinematic priors for fast human pose estimation”, arXiv:2002.03933, 2020.
[17] J. Wang, X. Long, Y. Gao, E. Ding, and S. Wen, “Graph-pcnn: Two stage human pose estimation with graph pose refinement,” arXiv preprint arXiv:2007.10599, 2020.
[18] H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, J. Jia, “Human pose estimation with spatial contextual information”, arXiv preprint arXiv:1901.01760, 2019.
[19] A. M. Lehrmann, P. V. Gehler and S. Nowozin, "A Non-parametric Bayesian Network Prior of Human Pose," 2013 IEEE International Conference on Computer Vision, 2013, pp. 1281-1288.
[20] V.Vapnik, R.Izmailov, “Learning using privileged information: similarity control and knowledge transfer”, The Journal of Machine Learning Research, vol. 16, pp. 2023–2049.
[21] C. Ionescu, D. Papava, V. Olaru and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1325-1339.
[22] D. Pavllo, C. Feichtenhofer, D. Grangier and M. Auli, "3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training," CVPR, 2019, pp. 7753-7762.
[23] R. Liu, J. Shen, H. Wang, C. Chen, S. -c. Cheung and V. Asari, "Attention Mechanism Exploits Temporal Contexts: Real-Time 3D Human Pose Reconstruction," CVPR, 2020, pp. 5064-5073.
[24] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen and J. Luo, "Anatomy-Aware 3D Human Pose Estimation With Bone-Based Pose Decomposition," IEEE Transactions on Circuits and Systems for Video Technology, 2021, vol. 32, no. 1, pp. 198-209.

TABLE V. COMPARISON OF OUR MODEL’S RESULTS BY MPJPE METRIC ON THE ENTIRE TEST DATASET AND FOR ALL ACTIONS

	Alpha						
	80	**100**	**200**	**1000**	**5000**	**10000**	**100000**
Photo	51.58	51.57	51.56	51.55	51.55	51.55	51.56
Purch	41.97	41.95	41.92	41.86	41.87	41.85	41.84
Smoke	45.49	45.48	45.47	45.45	45.45	45.45	45.45
Phone	46.45	46.45	46.43	46.42	46.42	46.42	46.41
Pose	42.10	42.10	42.10	42.10	42.09	42.10	42.10
Sit	53.30	53.30	53.30	53.32	53.31	53.32	53.33
Wait	43.32	43.31	43.29	43.27	43.26	43.26	43.26
WalkD	46.05	46.04	46.00	45.96	45.96	45.95	45.94
Disc	44.80	44.79	44.77	44.74	44.74	44.73	44.73
WalkT	32.35	32.35	32.35	32.36	32.35	32.36	32.37
SitD	60.72	60.72	60.73	60.74	60.74	60.75	60.76
Eat	39.75	39.74	39.72	39.71	39.70	39.70	39.69
Greet	42.51	42.50	42.49	42.47	42.47	42.47	42.47
Dir	41.46	41.45	41.43	41.40	41.39	41.39	41.39
Walk	31.77	31.76	31.72	31.68	31.67	31.66	31.65
Aver	44.24	44.24	44.22	44.20	44.20	44.20	44.20

TABLE VI. COMPARISON OF OUR MODEL’S RESULTS BY P-MPJPE METRIC ON THE ENTIRE TEST DATASET AND FOR ALL ACTIONS

	Alpha						
	80	**100**	**200**	**1000**	**5000**	**10000**	**100000**
Photo	39.49	39.48	39.48	39.48	39.48	39.49	39.49
Purch	31.97	31.96	31.95	31.92	31.92	31.91	31.90
Smoke	36.44	36.44	36.42	36.41	36.41	36.40	36.40
Phone	35.28	35.27	35.25	35.23	35.23	35.22	35.21
Pose	32.17	32.17	32.17	32.16	32.15	32.15	32.15
Sit	42.80	42.79	42.79	42.79	42.79	42.79	42.79
Wait	32.42	32.42	32.40	32.39	32.39	32.38	32.38
WalkD	35.31	35.31	35.31	35.30	35.30	35.30	35.29
Disc	34.80	34.79	34.78	34.76	34.76	34.75	34.75
WalkT	25.97	25.98	25.99	26.01	26.01	26.02	26.02
SitD	48.51	48.51	48.52	48.52	48.52	48.53	48.53
Eat	32.56	32.55	32.55	32.56	32.56	32.56	32.56
Greet	34.62	34.62	34.61	34.59	34.59	34.59	34.59
Dir	32.47	32.47	32.46	32.45	32.45	32.45	32.45
Walk	24.51	24.51	24.50	24.48	24.48	24.48	24.47
Aver	34.62	34.62	34.61	34.60	34.60	34.60	34.60

TABLE VII. COMPARISON OF OUR MODEL'S RESULTS BY MPJPE METRIC ON THE S9 SUBJECT AND FOR ALL ACTIONS

	Alpha						
	80	**100**	**200**	**1000**	**5000**	**10000**	**100000**
Photo	52.58	52.57	52.55	52.55	52.54	52.55	52.56
Purch	43.12	43.11	43.09	43.07	43.07	43.07	43.06
Smoke	46.31	46.31	46.29	46.28	46.27	46.28	46.27
Phone	49.48	49.47	49.46	49.46	49.45	49.45	49.45
Pose	39.33	39.33	39.32	39.33	39.32	39.33	39.33
Sit	58.95	58.94	58.95	58.97	58.96	58.97	58.98
Wait	42.46	42.46	42.46	42.46	42.45	42.46	42.46
WalkD	43.95	43.95	43.94	43.93	43.92	43.93	43.93
Disc	46.06	46.06	46.05	46.03	46.03	46.03	46.02
WalkT	35.12	35.11	35.10	35.10	35.09	35.10	35.10
SitD	64.07	64.07	64.07	64.07	64.06	64.07	64.07
Eat	36.47	36.45	36.41	36.35	36.35	36.33	36.32
Greet	45.20	45.21	45.22	45.23	45.22	45.23	45.24
Dir	41.40	41.39	41.38	41.36	41.35	41.35	41.35
Walk	34.43	34.42	34.38	34.33	34.32	34.31	34.30
Aver	45.26	45.26	45.24	45.23	45.23	45.23	45.23

TABLE VIII. COMPARISON OF OUR MODEL'S RESULTS BY P-MPJPE METRIC ON THE S9 SUBJECT AND FOR ALL ACTIONS

	Alpha						
	80	**100**	**200**	**1000**	**5000**	**10000**	**100000**
Photo	39.53	39.53	39.53	39.54	39.54	39.54	39.55
Purch	32.63	32.62	32.61	32.60	32.61	32.60	32.60
Smoke	37.12	37.11	37.10	37.08	37.08	37.08	37.07
Phone	37.07	37.07	37.06	37.06	37.05	37.05	37.05
Pose	31.34	31.34	31.35	31.35	31.35	31.35	31.35
Sit	49.17	49.17	49.17	49.17	49.18	49.18	49.18
Wait	32.63	32.63	32.64	32.64	32.64	32.64	32.64
WalkD	34.24	34.24	34.25	34.26	34.26	34.26	34.27
Disc	36.16	36.16	36.15	36.15	36.15	36.15	36.15
WalkT	28.02	28.03	28.05	28.07	28.06	28.07	28.08
SitD	48.84	48.85	48.85	48.85	48.85	48.85	48.85
Eat	29.93	29.92	29.90	29.87	29.87	29.87	29.85
Greet	36.55	36.56	36.57	36.58	36.57	36.58	36.58
Dir	33.19	33.19	33.18	33.18	33.18	33.18	33.18
Walk	26.53	26.53	26.52	26.50	26.50	26.50	26.50
Aver	35.53	35.53	35.53	35.53	35.53	35.53	35.53

TABLE IX. COMPARISON OF OUR MODEL'S RESULTS BY MPJPE METRIC ON THE S11 SUBJECT AND FOR ALL ACTIONS

	Alpha						
	80	**100**	**200**	**1000**	**5000**	**10000**	**100000**
Photo	50.51	50.51	50.50	50.48	50.49	50.49	50.48
Purch	40.43	40.41	40.35	40.26	40.27	40.24	40.21
Smoke	44.11	44.10	44.08	44.06	44.06	44.06	44.06
Phone	43.32	43.31	43.29	43.27	43.27	43.27	43.27
Pose	45.88	45.88	45.88	45.87	45.87	45.87	45.87
Sit	44.85	44.85	44.86	44.87	44.87	44.88	44.89
Wait	44.25	44.24	44.19	44.14	44.14	44.13	44.12
WalkD	49.61	49.59	49.51	49.40	49.41	49.38	49.34
Disc	41.91	41.90	41.86	41.78	41.78	41.77	41.75
WalkT	29.38	29.38	29.39	29.42	29.41	29.42	29.43
SitD	56.81	56.81	56.84	56.87	56.86	56.88	56.89
Eat	43.68	43.67	43.68	43.71	43.70	43.72	43.72
Greet	39.31	39.30	39.26	39.20	39.20	39.19	39.17
Walk	28.46	28.44	28.40	28.37	28.37	28.36	28.35
Dir	41.65	41.64	41.61	41.55	41.54	41.53	41.51
Aver	42.94	42.94	42.91	42.88	42.88	42.88	42.87

TABLE X. COMPARISON OF OUR MODEL'S RESULTS BY P-MPJPE METRIC ON THE S11 SUBJECT AND FOR ALL ACTIONS

	Alpha						
	80	**100**	**200**	**1000**	**5000**	**10000**	**100000**
Photo	39.44	39.44	39.43	39.42	39.42	39.42	39.42
Purch	31.08	31.08	31.05	31.00	31.00	30.99	30.98
Smoke	35.30	35.30	35.29	35.27	35.27	35.27	35.27
Phone	33.42	33.40	33.37	33.33	33.33	33.32	33.31
Pose	33.29	33.29	33.28	33.26	33.25	33.25	33.24
Sit	33.27	33.26	33.25	33.25	33.25	33.25	33.25
Wait	32.19	32.18	32.15	32.12	32.12	32.11	32.10
WalkD	37.14	37.13	37.10	37.06	37.06	37.05	37.04
Disc	31.67	31.66	31.63	31.57	31.58	31.56	31.55
WalkT	23.76	23.77	23.79	23.80	23.80	23.81	23.81
SitD	48.11	48.11	48.13	48.14	48.14	48.14	48.15
Eat	35.70	35.70	35.73	35.77	35.77	35.79	35.80
Greet	32.33	32.32	32.29	32.24	32.25	32.23	32.22
Walk	22.00	21.99	21.97	21.96	21.96	21.96	21.95
Dir	30.14	30.14	30.11	30.08	30.07	30.07	30.06
Aver	33.26	33.25	33.24	33.22	33.22	33.21	33.21