

---

# Localization with Sampling-Argmax

## *Supplementary material*

---

Jiefeng Li   Tong Chen   Ruiqi Shi   Yujing Lou   Yong-Lu Li   Cewu Lu  
Shanghai Jiao Tong University

{ljf\_likit, chentong1023, gzfoxie, louyujing, yonglu\_li, lucewu}@sjtu.edu.cn

In the supplemental document, we elaborate on the training settings (Appendix A), the broader impact of our work (Appendix B), limitation and future work (Appendix C), descriptions of the utilized datasets (Appendix D), experiments on facial landmark localization (Appendix E), comparison between the learned distribution of soft-argmax and sampling-argmax (Appendix F), and qualitative results (Appendix G).

### A Training Details

**2D Human Pose Estimation from RGB** We adopt SimplePose [17] for experiments. The model is trained and evaluated on COCO Keypoint [13]. ResNet-50 [9] is adopted as the backbone network. The input image is resized to  $256 \times 192$ . The learning rate is set to  $1 \times 10^{-3}$  at first and reduced by a factor of 10 at the 90th epoch and the 120th epoch. We use the Adam solver and train for 140 epochs, with a mini-batch size of 32 per GPU and 8 1080Ti GPUs in total. For comparison with the auxiliary losses, we set the target variance  $\sigma_t^2$  to 4, the loss weight of variance regularization to 1, and the loss weight of distributions regularization to 0.1 to achieve the best results after tuning.

**3D Human Pose Estimation from RGB** We follow the model architecture of Integral Pose [16]. ResNet-50 [9] is adopted as the backbone network. The input image is resized to  $256 \times 256$ . The learning rate is set to  $1 \times 10^{-3}$  at first and reduced by a factor of 10 at the 90th and 120th epoch. We use the Adam solver and train for 140 epochs, with a mini-batch size of 16 per GPU and 8 1080Ti GPUs in total. Following the settings of previous works [16, 14], we mix Human3.6M and MPII [8] data for training. Each mini-batch consists of half 2D and half 3D samples. Five subjects (S1, S5, S6, S7, S8) are used for training and two subjects (S9, S11) for evaluation. We set the target variance  $\sigma_t^2$  to 4, the loss weight of variance regularization to 1, and the loss weight of distributions regularization to 0.1 to achieve the best results after tuning.

**Retina Segmentation from OCT** We follow the model architecture of [10]. The input image is resized to  $128 \times 1024$ . The learning rate is set to  $1 \times 10^{-4}$  at first and reduced by a factor of 10 at the 10th and the 20th epoch. We use the Adam solver and train for 30 epochs, with a mini-batch size of 2 and 1 GPU. The split of training, validation and test sets follows the settings of the previous method [10]. We set the target variance  $\sigma_t^2$  to 4, the loss weight of variance regularization to 1, and the loss weight of distributions regularization to 1 to achieve the best results after tuning.

**Supervised Object Keypoint Estimation from Point Clouds** We adopt PointNet++ [15] as the backbone network. The output of the last layer is a per-point probability map for each keypoint. The input point cloud consists of 2048 points represented by their Euclidean coordinates sampled from a normalized object, and the indexes of keypoints are given. The learning rate is set to  $1 \times 10^{-3}$  and halved every 10 epochs. We use Adam solver and train for 100 epochs with a mini-batch size of 8 on one GPU for each category. We set the target variance  $\sigma_t^2$  to 4, the loss weight of variance regularization to 1, and the loss weight of distributions regularization to 0.01 to achieve the best results after tuning.

**Unsupervised Object Keypoint Estimation from Point Clouds** The learning rate is set to  $1 \times 10^{-3}$  and halved every 10 epochs. We use the Adam solver and train for 50 epochs, with a mini-batch size of 8 and one GPU for each category. We set the target variance  $\sigma_t^2$  to 4, the loss weight of variance regularization to 1, and the loss weight of distributions regularization to 0.01 to achieve the best results after tuning.

**Facial Landmark Localization from RGB** ResNet-18 [9] is adopted as the backbone network. The head network consists of 3 deconvolution layers and a  $1 \times 1$  convolution layer. The input image is resized to  $256 \times 256$ . The learning rate is set to  $1 \times 10^{-3}$  at first and reduced by a factor of 10 at the 10th and 20th epoch. We use the Adam solver and train for 30 epochs, with a mini-batch size of 32 and 4 GPUs in total. We set the target variance  $\sigma_t^2$  to 4, the loss weight of variance regularization to 1, and the loss weight of distributions regularization to 0.1 to achieve the best results after tuning.

## B Broader Impact

In this work, we propose sampling-argmax to improve the ability of machines to understand target positions in input data. Current methods usually adopt computationally expensive models to improve the localization accuracy, which could cost many financial and environmental resources. We partly alleviate this issue by presenting a simple yet effective method.

Furthermore, our method is an improvement of existing capabilities but does not introduce a radically new capability in machine learning. Thus our contribution is unlikely to facilitate misuse of technology that is already available to anyone.

## C Limitation and Future Work

In our method, the underlying density function of the target position is approximated by a mixture of sub-distributions. By comparing the performance of the three proposed bases, we see that a more accurate reconstruction of the underlying function leads to better results. Theoretically, the underlying density function cannot be perfectly reconstructed since the proposed basis distributions are fixed. To address this limitation, learnable sub-distributions could be adopted in future works. For example, *normalizing flow* models can be leveraged to predict sub-distribution at each position according to the corresponding features. In this way, the sub-distributions are no longer fixed, and the mixture distribution has the potential to precisely reconstruct the underlying distribution and further improve the model performance.

## D Data Acquisition

In our experiments, we use five different datasets, including COCO Keypoint [13], Human3.6M [12], MSHC [11], KeypointNet [18] and MTFL [19]. These public datasets do not contain personally identifiable information or offensive content.

**COCO Keypoint** COCO Keypoint dataset is licensed under the Creative Commons Attribution 4.0 License [2]. The images and annotations are publicly available. We download the images and annotations from its official website [1].

**Human3.6M** Human3.6M dataset is licensed under [5]. To obtain the data, we register and download it from its official website [4].

**MSHC** MSHC dataset is publicly available, and no license is specified. We download the data from its official website [7].

**KeypointNet** KeypointNet dataset is publicly available, and no license is specified. We download the data from its official website [6].

**MTFL** MTFL dataset is publicly available, and no license is specified. We download the data from its official website [3].

## E Facial Landmark Localization from RGB

We further evaluate the proposed sampling-argmax on the facial landmark localization dataset MTFL [19]. Absolute error and relative error (normalized by the two-eye distance) are adopted as evaluation metrics. Quantitative results are reported in Table 1. Consistent with the experiments on other tasks, sampling-argmax provides performance improvement to facial landmark localization.

Table 1: Quantitative results on MTFL dataset.

	Soft	Soft w/ V.R.	Soft w/ D.R.	Samp. Uni.	Samp. Tri.	Samp. Gau.
Abs. Err ↓	3.18	3.16	3.15	3.00	2.98	<b>2.94</b>
Rel. Err ↓	7.25	7.22	7.20	6.86	<b>6.82</b>	6.96

## F Visualization of learned probability maps

We show the predicted probability maps of soft-argmax and sampling-argmax in Figure 1. It shows that soft-argmax is prone to predict multi-modal distribution, while the proposed sampling-argmax predicts better-calibrated probability maps.

## G Qualitative Results

Qualitative results on six tasks are shown in Figure 2, 3, 4, 5, 6 and 7.

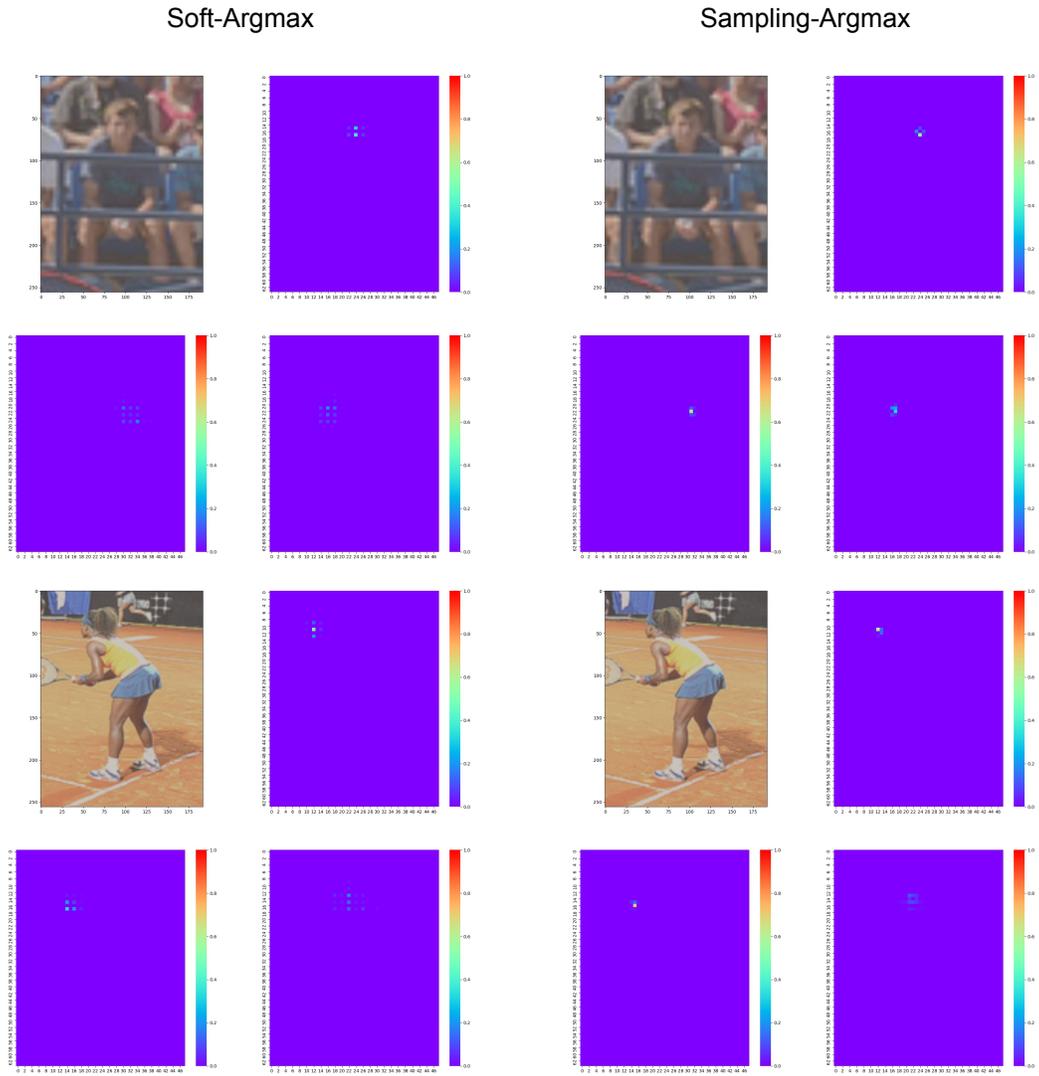


Figure 1: **Visualization** of the learned distribution. **Left:** Soft-Argmax. **Right:** Sampling-Argmax.

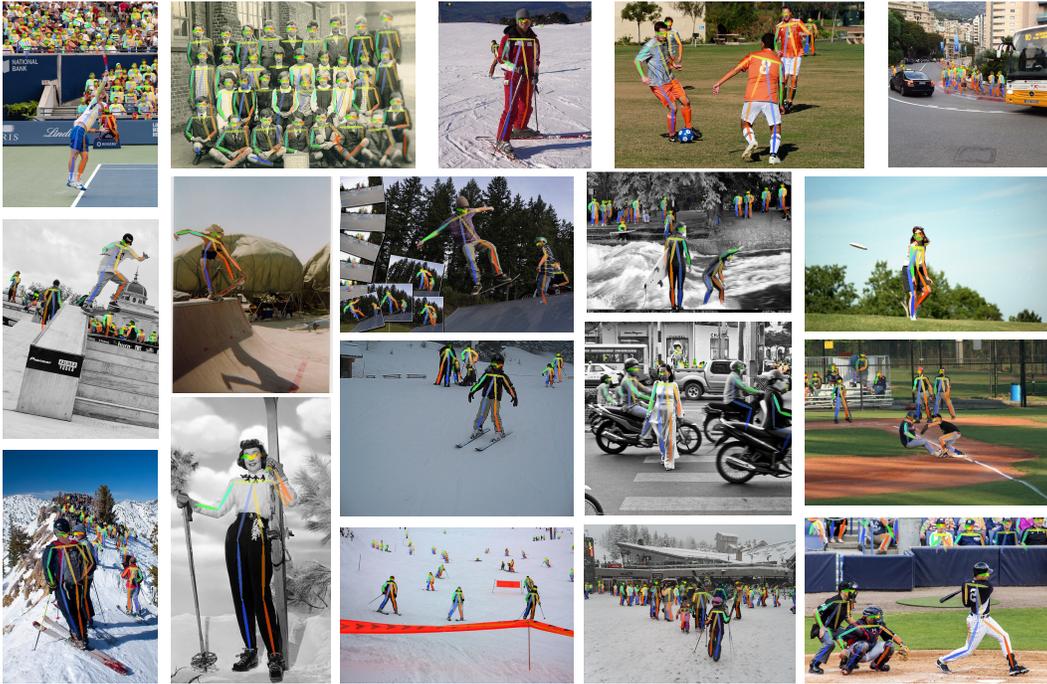


Figure 2: **Qualitative** results of 2D human pose estimation on COCO Keypoint.

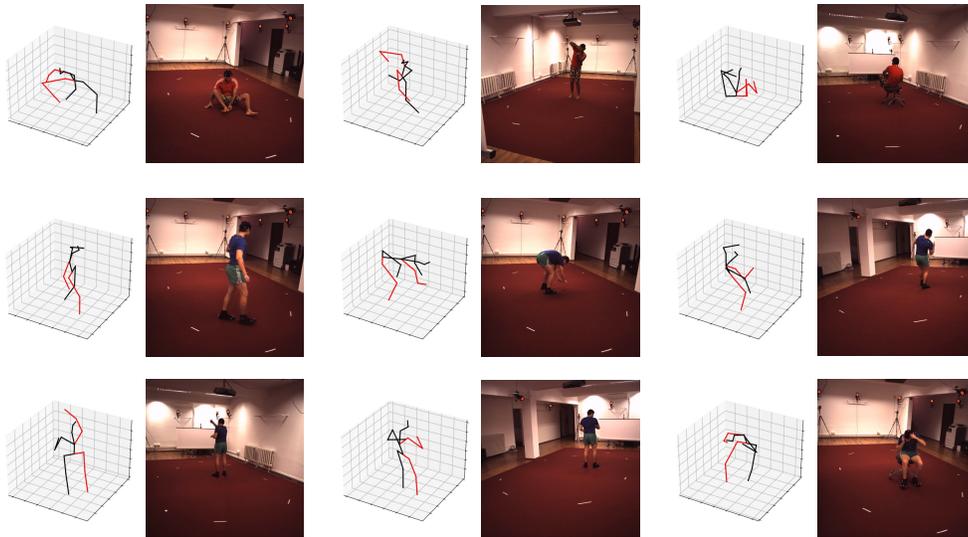


Figure 3: **Qualitative** results of 3D human pose estimation on Human3.6M.

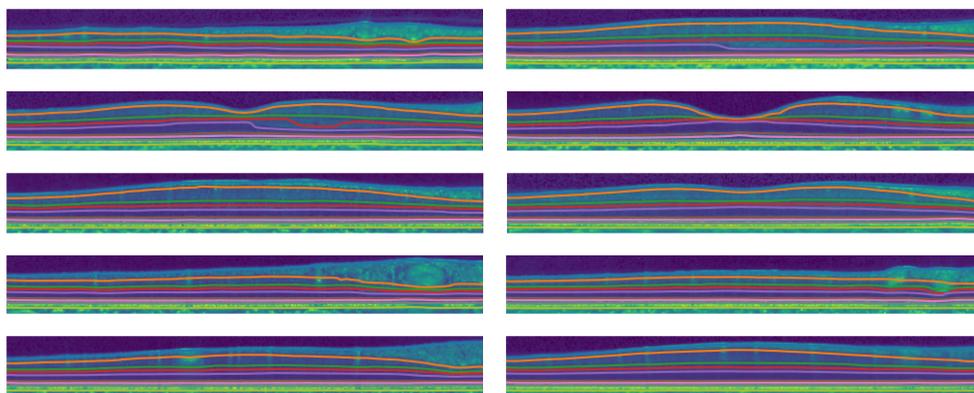


Figure 4: **Qualitative** results of retina segmentation on MSHC.

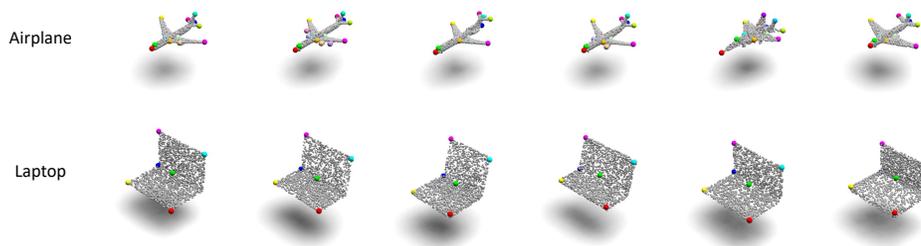


Figure 5: **Qualitative** results of supervised model on KeypointNet.

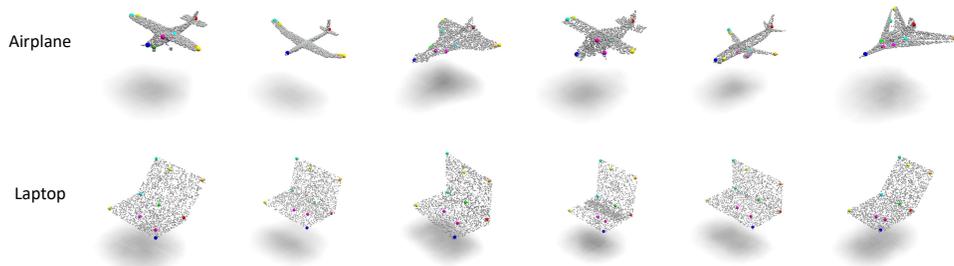


Figure 6: **Qualitative** results of unsupervised model on KeypointNet.



Figure 7: **Qualitative** results of facial landmark localization on MTLF.

## References

- [1] Coco - common objects in context. <https://cocodataset.org/#home>.
- [2] Coco license agreement. <https://creativecommons.org/licenses/by/4.0/legalcode>.
- [3] Facial landmark detection by deep multi-task learning. <http://mmlab.ie.cuhk.edu.hk/projects/TCDCN.html>.
- [4] Human3.6m dataset. <http://vision.imar.ro/human3.6m/description.php>.
- [5] Human3.6m license agreement. <http://vision.imar.ro/human3.6m/eula.php>.
- [6] Keypointnet. <https://github.com/qq456cvb/KeypointNet>.
- [7] Resources - iacl. <http://iacl.ece.jhu.edu/index.php?title=Resources>.
- [8] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [10] Yufan He, Aaron Carass, Yihao Liu, Bruno M Jedynek, Sharon D Solomon, Shiv Saidha, Peter A Calabresi, and Jerry L Prince. Fully convolutional boundary regression for retina oct segmentation. In *MICCAI*, 2019.
- [11] Yufan He, Aaron Carass, Sharon D Solomon, Shiv Saidha, Peter A Calabresi, and Jerry L Prince. Retinal layer parcellation of optical coherence tomography images: Data resource for multiple sclerosis and healthy controls. *Data in brief*, 2019.
- [12] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 2014.
- [13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.
- [14] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single rgb image. In *ICCV*, 2019.
- [15] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017.
- [16] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018.
- [17] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018.
- [18] Yang You, Yujing Lou, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Cewu Lu, and Weiming Wang. Keypointnet: A large-scale 3d keypoint dataset aggregated from numerous human annotations. In *CVPR*, 2020.
- [19] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014.