Adaptive Computationally Efficient Network for Monocular 3D Hand Pose Estimation (Supplementary Material)

Zhipeng Fan[1], Jun Liu[2⋆], and Yao Wang[1]

[1] Tandon School of Engineering, New York University, Brooklyn NY, USA
{zf606, yw523}@nyu.edu
[2] Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore
jun_liu@sutd.edu.sg

1 Implementation details

We employ Hourglass network as the base pose encoder. For the coarse pose encoder, we set the number of channels to 32 for both the STB dataset and the FPHA dataset. The coarse pose encoder works at the resolution of 64×64 and 128×128 for STB and FPHA, respectively, which greatly reduce the amount of computation. We introduce two separate models with different complexities since the poses in STB are more stable and therefore are less challenging compared to FPHA. For the fine pose encoders, we use 64 channels and the input resolution is 256×256 on both datasets, which is the typical setting of the original Hourglass module. For both pose encoders, we only use one stack instead of multiple ones, as stacking multiple hourglass modules only brings marginal accuracy improvement while almost doubles the computational cost.

Based on the original Hourglass module, we further directly estimate the 3D coordinates of the hand joints from the summation of intermediate feature maps and the 2D heatmaps $\mathbf{H}^t$, as illustrated in Fig. 2(a) (from the main manuscript). Specifically, we apply 1×1 convolution to map the feature maps and heatmaps to the same dimension to facilitate the summation. The down-sampled feature maps before the Hourglass modules are also skip-connected to the summation, which helps to recover the information from earlier layers and ease the learning process. The combined feature maps are fed into four consecutive convolution layers with stride 2 followed by batch normalization. The down-sampled feature maps are then input into 2 linear layers to directly regress the 3D coordinates, which are parameterized following [1, 2]. We generate the ground truth 2D heatmaps $\mathbf{H}^t$ with $std = 1$.

During training, we anneal the temperature τ from 5 to 0.5 with a factor of 0.9 for each epoch, which yields good performance empirically. We set the weight term δ of the activation regularization of the fine pose encoder in Eq. (13) to 10 during training.

⋆ Corresponding author.

2 Visualization

We further visualize the decision of the gate as well as the 3D pose estimation results. As shown by the attached video, the proposed model dynamically decides whether to opt out the computation of fine features via the fine pose encoder and derives accurate hand pose estimations.

References

1. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2272–2281 (2019)
2. Liu, J., Ding, H., Shahroudy, A., Duan, L.Y., Jiang, X., Wang, G., Chichung, A.K.: Feature boosting network for 3d pose estimation. IEEE transactions on pattern analysis and machine intelligence (2019)