Explicit Knowledge Distillation for 3D Hand Pose Estimation from Monocular RGB-Supplementary Document

Yumeng Zhang[1]
zhangyum17@mails.tsinghua.edu.cn
Li Chen[1]
chenlee@tsinghua.edu.cn
Yufeng Liu[2]
liuyufeng@kuaishou.com
Wen Zheng[2]
zhengwen@kuaishou.com
Junhai Yong[1]
yongjh@tsinghua.edu.cn

[1] School of Software, BNRist, Tsinghua University
Beijing, China

[2] Y-tech, Kuaishou Technology
Beijing, China

1 Introduction

This document mainly contains the following:

- The illustration of network architecture used in the main paper.
- The setup for training.
- The definition of evaluation metrics.
- Some qualitative results.
- Comparisons with the state-of-the-art knowledge distillation methods.

2 Illustration of Network Architecture

In the main paper, we evaluate the proposed method using five different networks-TS120G, TS20G, TS500M, Resnet50 and Squeezenet. These networks can be divided into two categories. TS120G is the network of [4], TS20G and TS500M are modified from TS120G. Thus, we record the three networks as TS-MAC network. The illustration of TS-MAC with EKD is shown in Figure 1. Resnet50 and Squeezenet are general networks. The illustration of them is shown in Figure 2. Note that the EKD is a framework. It can be easily adapted to various networks and tasks, and we leave this for future work.

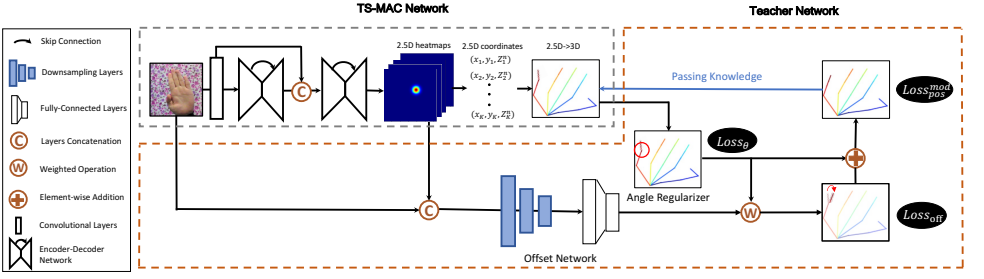

Figure 1: The illustration of TS-MAC network with EKD. $\{(x_i, y_i, Z_i^n)\}_{i=1}^K$ is the 2.5D coordinates [2] of the keypoints, where K=21. The 2.5D coordinates can be translated to 3D coordinates using intrinsic camera parameters and the depth of the root keypoint. The 3D coordinates are used to calculate the angles that are defined in our main paper.

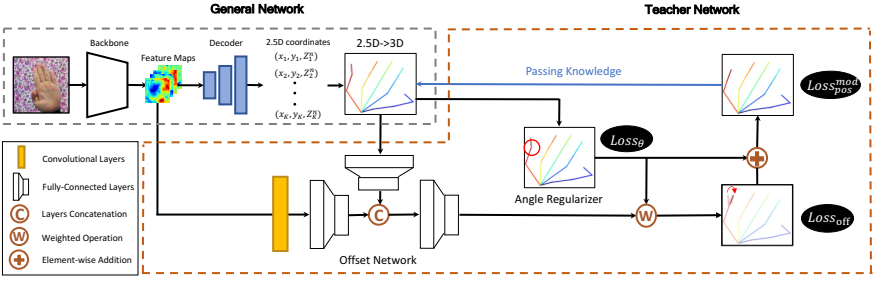

Figure 2: The illustration of general network with EKD.

3 Setup for Training

All the training images are cropped into 128×128 before feeding into the network. Online data augmentation is adopted to alleviate the overfitting problem. During the training, the samples are rotated (−45°,45°), translated (±20 pixels) and scaled (0.7,1.1) randomly. Adam optimizer is used. The batch size is 32 and the regularization strength is 5×10^{-4}. The learning rate is set to 10^{-4} for TS120G and 10^{-3} for other models and decreases to its $\frac{1}{10}$ every 30 epochs. The hyperparameter λ_1 is set to 20 and λ_2, λ_3 is set to 1 in the experiments. All the experiments are done on Geforce GTX 2080Ti GPU with CUDA 9.0.

4 Definition of Evaluation Metrics

In this paper, we use Area Under the Curve (AUC) and angle violation frequency to evaluate the proposed algorithm. AUC is the area under the PCK curve and the PCK curve is shown in Figure 4 of the main paper. The angle violation frequency is the proportion of physiologically invalid samples in the test dataset. A sample that has at least one invalid angle prediction is considered as a physiologically invalid sample. The invalid angle means that this angle is outside the predefined angle range.

Methods	2DAUC@30	3DAUC	V.F.
Student	59.9%	61.6%	6.22%
KD	49.9%	53.1%	11.3%
NST	56.1%	61.4%	7.62%
VID	54.9%	58.3%	6.83%
RKD	56.0%	59.4%	6.67%
CC	55.7%	60.3%	7.46%
SP	57.9%	60.1%	8.44%
EKD	**61.2%**	**64.3%**	**4.74%**

Table 1: Comparisons with state-of-the-art knowledge distillation methods on RHP dataset. V.F. is the abbreviation of violation frequency. 2DAUC@30 means that the area under 2D PCK curve between 0-30pixel. 3DAUC is the area under 3D PCK curve between 20-50mm.

5 Comparisons with State-of-the-art Knowledge Distillation Methods

In this section, the proposed method is compared with state-of-the-art knowledge distillation methods-KD [2], NST [3], VID [1], RKD [5], CC [6] and SP [7] to demonstrate the effectiveness of the method. Note that existing KD methods are mainly designed for classification task. Thus, it is necessary to explore the way to apply these methods to pose estimation task. From our experimental results, the high-level features of the teacher network are not always helpful for students to mimic. For example, when the loss of CC is added at the high-level features, the network fails to converge. Therefore, we let the student network learn the knowledge from low-level features of the teacher network. As existing knowledge distillation methods require a sophisticated teacher network, the TS120G model is adopted as the teacher for the TS500M network in this task, and the low-level features are the features obtained by downsampling operations in the first stack.

The results are shown in Table 1 and 2. The proposed method surpasses existing methods by a large margin on RHP dataset and achieves best results on LM dataset when combining with the loss of the CC method. On the RHP dataset, existing KD methods obtained a lower performance than the baseline. However, this phenomenon is not surprising. RHP is synthetic dataset with many difficult poses, thus it requires a complex reasoning process during the inference, resulting in complicate relations between features. When transferring knowledge from the features of such a teacher, the students is more likely to be impacted by knowledge uncertainty and knowledge omission, which are introduced in the Introduction section of the main paper. By contrast, the proposed EKD method works well in this situation as it transfers human knowledge.

6 Qualitative Results

Figure 3 shows some visual results of proposed approach and corresponding baselines. As the teacher network has limited guiding effect on the TS120G model and there is no obvious improvement visually, we only show the results of the other four models. It can be seen that the models trained with our teacher network tend to output results that conform to the angle

Methods	2DAUC@30	3DAUC	V.F.
Student	69.1%	79.3%	4.83%
KD	65.8%	78.9%	7.33%
NST	72.0%	83.7%	3.42%
VID	71.8%	85.1%	2.67%
RKD	70.0%	82.5%	4.43%
CC	73.3%	84.9%	3.15%
SP	71.7%	82.9%	2.92%
EKD	72.7%	85.1%	**1.74%**
EKD+CC	**75.8%**	**86.8%**	2.67%

Table 2: Comparisons with state-of-the-art knowledge distillation methods on LM dataset.

constraints.

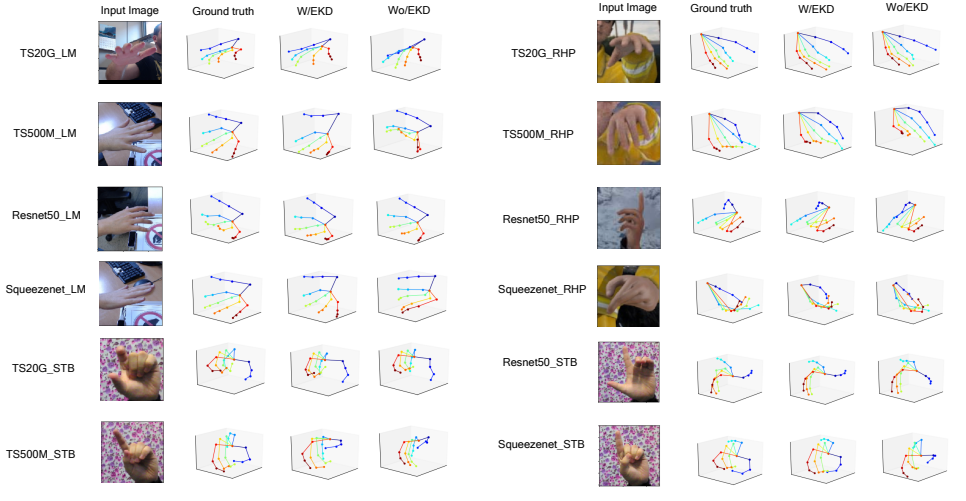

Figure 3: Visual results of different baselines and proposed method. The abnormal hand poses are corrected after using the proposed EKD method.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019.

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[3] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[4] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision*, pages 118–134, 2018.

[5] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[6] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5007–5016, 2019.

[7] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019.