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

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

Abstract

RGB-based 3D hand pose estimation methods frequently produce physiologically invalid gestures due to depth ambiguity and self-occlusion. Existing methods typically adopt complex networks and a large amount of data to avoid invalid gestures by automatically mining the physical constraints of the hand. These networks exhibit high computational complexity and thus are difficult to be deployed into mobile devices. In consideration of this problem, a novel knowledge distillation framework, called Explicit Knowledge Distillation, is proposed to enhance the performance of small pose estimation networks. The proposed teacher network has interpretable knowledge, explicitly passing the physical constraints to the student network. Experimental results on three benchmark datasets with five different sized models demonstrate the potential of our approach.

1 Introduction

Gestures are among the most natural interactive movements. With the application of virtual reality and augmented reality, 3D hand pose estimation technologies have been increasingly applied to enhance the user experience. Existing methods can be divided into two categories, depth based and RGB based. Depth-based methods have been studied extensively in recent years and their performance has been improved considerably because of the development of deep learning and the emergence of large training datasets. However, the popularization of this technology is limited by the strict requirements for image acquisition equipment. Therefore, an increasing number of studies are focusing on RGB-based 3D pose estimation.

However, 3D pose estimation from an RGB image poses a more challenging problem, because several poses should be guessed on the basis of prior knowledge due to serious depth ambiguity and inherent self-occlusion. Prior knowledge (e.g., physical constraints) requires a network to be aware of the hand structure. Thus, existing methods have adopted multi-stage regression [12, 31] to predict a pose progressively or have applied a pre-defined hand model [4, 15, 29] to obtain physical valid pose predictions. Unfortunately, considering the complex relationship between the RGB images and poses, these methods utilize deep and sophisticated networks to obtain valid results. For example, the number of multiply-accumulate operations (MAC) of Iqbal [10] reaches more than 120G, hindering the transformation of most studies on mobile devices.

Small models are easy to deploy. However, they frequently fail to learn the prior knowledge, such as physical constraints, due to their weak learning capability, resulting in a significant decline in performance. A promising solution is to enable the small networks, also called the student networks, to mimic a large, well-trained teacher network; this process is known as knowledge distillation (KD). However, the majority of recent knowledge distillation studies have focused on image classification. Two major drawbacks exist when adapting this solution to 3D hand pose estimation.

The first drawback is knowledge uncertainty. Inferencing the 3D pose from RGB images is a complicated reasoning process due to the depth ambiguity. Consequently, the correlations between features are considerably more complex than that of the classification models. Although current knowledge distillation methods are reported to enhance the performance of student networks by using strategies, such as similarity preserving [22], correlation congruence [17] or relations [16], whether the knowledge transferred from a teacher network is beneficial or comprehensive for the student networks to mimic remains unclear. The second drawback is knowledge omission. Prior knowledge of hand structure is crucial to ensure the validity of the poses. However, transferring such knowledge completely by learning only the characteristics or relationships of certain layers is difficult. Hence, important knowledge may be overlooked during the transferring process.

On the basis of the preceding considerations, we provide a solution from a novel perspective, where any pose estimation network can benefit from our approach. In practice, teachers frequently teach by discovering the mistakes. First, teachers allow students give their own answers; then, they teach students knowledge by correcting the mistakes in their answers. Inspired by this real-life teaching activity, we propose an explicit knowledge distillation (EKD) framework to introduce the physical constraints, passing the human knowledge to the training process. The teacher network of EKD initially reviews the predictions of a student and then points out the errors in them by artificially defined knowledge. Furthermore, the teacher network provides the modification to help the student avoid similar mistakes in the future. Conventional KD methods transfer knowledge from big networks, whose knowledge is implicit. The proposed EKD passes the explicit knowledge defined by human, thus the knowledge is definite and complete.

To summarize, our contributions are as follows: **(1)** We introduce a constraints-enforcing framework for pose estimation networks that can improve the accuracy without changing the size of the original network. **(2)** We design a physical constraints regularizer for the keypoints-based 3D hand pose estimation method. **(3)** We propose a teacher network with interpretable knowledge, explicitly passing human knowledge to the student network.

Extensive experiments on three benchmark datasets using different sized networks are conducted to demonstrate the effectiveness of EKD in transferring interpretable knowledge. The EKD method is compared with state-of-the-art hand pose estimation algorithms. The results show that the proposed method achieves the best accuracy among the compared algorithms.

2 Related Work

2.1 Hand Pose Estimation

Depth-based 3D hand pose estimation has recently achieved rapid development. By contrast, studies on RGB-based 3D pose estimation remain few. Although differences exist between these types of images, methods for extracting 3D poses have much in common. Thus, a brief review of 3D hand pose estimation methods is provided from a unified perspective. The two major strategies in this field are summarized as: 1) model-based and 2) joint-based pose estimation.

1) Model-based pose estimation Model-based methods define a hand model using specific parameters obtained in advance. Physical constraints are reflected within the range of these parameters; thus, constraints can be easily embed into the algorithm to ensure the validity of the results. Previous methods such as [14, 21] obtain the optimal model parameters by minimizing the handcrafted energy functions. However, the optimization process is intricate making it difficult to be adopted in resource-limited applications. Several researchers [2, 26] have recently proposed the use of deep learning methods to regress these parameters. For example, Mueller et al. [11] obtained the parameters of a hand model via a post process of the network. Zhang et al. [26] introduced an end-to-end method for recovering the hand parameters. However, considering the complex relations between input image and these parameters, the sophisticated networks should be used in their methods. Such requirement becomes an obstacle to transform their research into practice.

2) Joint-based pose estimation Joint-based methods learn the mapping between images and the 3D joint positions. A recent study by Zimmermann and Brox [31] used the PoseNet and PosePrior networks to predict 3D poses in canonical coordinates. Thereafter, a two-stack hourglass with latent 2.5D heatmaps was introduced by Iqbal et al. [10]. Their approach enhanced the performance of the joint-based methods by a considerable margin. Then, Zhang et al. [27] improved the performance by learning an adaptive latent space of synthetic and real-world datasets. More recently, Zhao et al. [28] proposed a knowledge distillation and generalization framework for RGB based pose estimation. They distilled knowledge from depth images and then transferred this cross-modal knowledge to those datasets without depth images. However, whether this knowledge is beneficial or complete for students to mimic is unclear. Thus, the physiologically invalid predictions may not be reduced. Keypoint position is not as abstract as hand parameters; thus, joint-based methods can achieve relatively higher accuracy with small networks than model-based methods. However, joint-based methods do not utilize hand geometry, such as physical constraints; hence, they produce many physiologically incorrect hand poses when predicting with small networks.

2.2 Knowledge Distillation

Knowledge distillation [1, 8, 16] refers to use a complex and superior teacher network to guide the training of low-complexity student networks. Hinton et al. [7] used the soft labels from the teacher network as part of the optimization goals of the student network to provide the interclass information for the first time. Then, Tung et al. [22] suggested that the knowledge transferring process benefited from a similarity metric. At the same time, Peng et al. [17] discovered that the correlation between instances is also crucial to boost the performance of the student. Zhang et al. [24] proposed a knowledge distillation strategy in

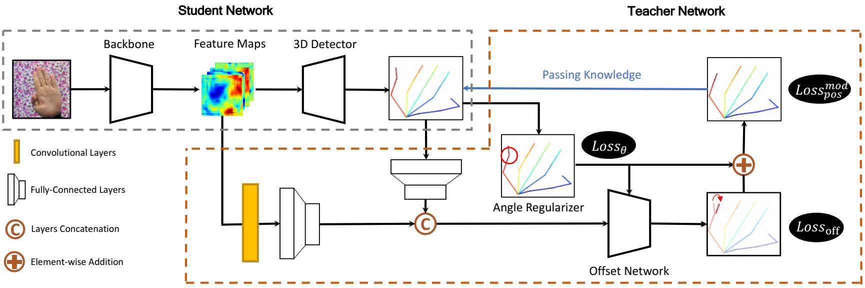


Figure 1: The overview of our method. A small pose estimation network (grey box) are regarded as the student network. The proposed teacher (brown box) analyzes the pose predictions of a student network and gives a demonstration to correct the angle-invalid pose. The information is passed to the student network by the back propagation of $Loss_{pos}^{mod}$. Note that the teacher network is only used during the training, thus it would not change the size of the student network during the inference.

human pose estimation, which aims to reduce the computational cost of Stacked Hourglass [12]. Although these approaches have achieved success in certain areas, we cannot identify what knowledge has transferred to the student due to uninterpretability of neural network. Thus, we have no idea if the physical constraints are passed to the student network. In view of this problem, we design a teacher model with explicit knowledge of physical constraints to help the student networks obtain valid hand poses.

3 Method

Given a cropped RGB image, the network predicts K=21 joint locations in the camera coordinate system. The physical prior knowledge is important to obtain physiologically valid poses and it mainly consists of two constraints, bone length and joint angle constraints. Owing to the powerful deep learning methods, bone length constraint can be easily satisfied as it directly associates with the accuracy of joint positions. However, the joint angle constraint may be neglected by the small networks, as it has complex nonlinear relationship with more than two joint positions. Therefore, we propose the EKD to help them recognize the validness of hand poses and avoid incorrect predictions.

The overview of our algorithm is shown in the Figure 1. The teacher of EKD consists of the angle regularizer and the offset network. The two parts are designed to mimic the behavior of a real teacher. In reality, the teacher helps students learn the knowledge by discovering errors in the behavior of students and giving them the right demonstration. Similarly, the angle regularizer recognizes the invalid pose predictions of student network and the offset network corrects them. Note that the teacher network is only adopted during training and the performance of the student can be improved with the guidance of the teacher. Next, we will introduce these modules in detail.

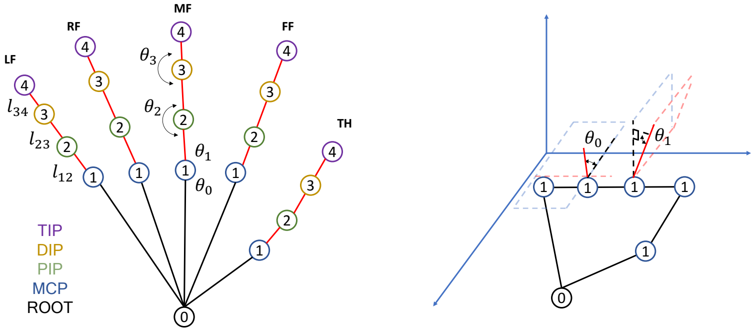


Figure 2: The illustration of the hand model and joint angles. TH, FF, MF, RF and LF are the abbreviation of thumb, forefinger, middle finger, ring finger and little finger. TIP, DIP, PIP, MCP and ROOT represent different knuckles.

3.1 Student Network

Small pose estimation networks are also called the student networks in knowledge distillation framework. In theory, the students can be networks of any arbitrary size, but usually those small and streamlined networks need the guidance of a teacher network. In this paper, we adopt modified networks of [10], which achieved optimal results in several datasets among joint-based methods, and some general networks, such as resnet [6] and squeezenet [9], as the student networks to prove the effectiveness of our algorithm.

3.2 Teacher Network

The proposed teacher network is to pass physical constraints to the student network. For better understanding of our method, we first introduce the definitions of joint angles and then the key modules of EKD–angle regularizer and offset network.

3.2.1 Angle Definition

The hand model has K=21 keypoints and 20 joint angles are defined with them. Each finger involves 4 angles $\{\theta_0, \theta_1, \theta_2, \theta_3\}$, and a diagram of them is shown in Figure 2. Take the angles of MF as an example. The angle θ_2, θ_3 can be easily calculated by the spatial angle of adjacent bones. MCP joints are more flexible than other joints, thus we define two angles that represent the movements on all sides to better measure the motion constraints of MCP joints. We assume the line between the MCP joint of LF, FF as the reference line L1. The plane formed by the root joint and reference line are regarded as the reference plane P1. l_{12} is a vector that points from a MCP joint to its corresponding PIP joint. θ_1 is defined by the angle between l_{12} and the reference plane, while θ_0 are defined by the angle between l_{12} and the perpendicular line of L1. The perpendicular line of L1 is the one in the plane determined by l_{12} and L1 (or a line parallel to L1). Note that the MCP joints of LF, RF, MF and FF in Figure 2 are considered to be in a straight line, in order to show the definition of θ_0, θ_1 more clearly. Claudia et al. [13] have illustrated the angle constraints of hands. However, this definition is not task specific. For example, when a hand pose dataset only contains the directional gesture, the keypoints of FF are in a straight line. Therefore, we define the angle constraints based on the statistics of the datasets. The max value $\overline{\theta}$ and min value $\underline{\theta}$ of each defined angle in the labeled datasets are calculated as the upper and lower bounds.

3.2.2 Angle Regularizer

Angle regularizer can recognize the invalid hand poses and enforce angle constraints through the loss function. This module first calculates the defined angles in Figure 2 from predictions of the pose estimation network, and then computes the angular loss $Loss_\theta$ for each finger, which is shown in Equation (1). The validity of hand poses can be improved by back propagation of the $Loss_\theta$.

$$Loss_\theta = \sum_{i=0}^{3}[max(\theta_i - \overline{\theta_i}, 0) + max(\underline{\theta_i} - \theta_i, 0)] \tag{1}$$

However, $Loss_\theta$ is not conducive to the back propagation of the networks because of the complex relationship between the angles and keypoint positions. Thus, we use the cosine and sine of the joint angles in the loss function to simplify the optimization process. The new loss is shown in Equation (2). As θ_0 and θ_1 are defined with perpendicular of the reference line and reference plane, the sine function is used to calculate the loss. The θ_2 and θ_3 are defined as the angle between two vectors, thus using the cosine is more effective.

$$Loss_\theta = \sum_{i=0}^{3}[max(f(\theta_i) - f(\overline{\theta_i}), 0) + max(f(\underline{\theta_i}) - f(\theta_i), 0)] \tag{2}$$

where f(·) is sine when i = 0,1, and f(·) is cosine when i = 2,3.

3.2.3 Offset Network

Small pose estimation networks could square up the angle errors in the predictions by the angle regularizer and try to correct these errors by themselves. However, the students may not find a universal way to correct them due to the weak learning ability. Therefore, a further guidance is needed. In view of this consideration, an offset network is proposed to inform the small pose estimation networks the way to correct the mistakes. The samples with invalid angles are sent to the offset network to get the offset vectors of keypoints. The offset vectors represent the difference between the ground truth and predictions. Formally speaking, the input of offset network is the features **f** which contain the image and prediction information, and a weight vector w that represents whether a sample has at least one invalid angle. Note that EKD is a framework, it can be applied to various network architectures. The features sent to the offset network can be selected according to the specific network architecture, as long as the features contain the information of the input image and the student predictions. **f** is usually obtained by the latent feature maps and the predictions $\{C_i^p\}_{i=1}^K$ of the student network. w is calculated by the angle regularizer. The output is the offset vector $\{M_i^p\}_{i=1}^K$ of those physiologically invalid predictions $\{C_i^p\}_{i=1}^K$ to the ground truth $\{C_i^g\}_{i=1}^K$. The offset network has been illustrated in Figure 1. Given the j_{th} sample in the dataset, the loss function of offset network is defined as:

$$Loss_{off} = \sum_{i=1}^{K} |M_i^p - w_j(C_i^g - C_i^p)| \tag{3}$$

where $w_j \in \{0,1\}$ and w_j is set to 1 when the predicted pose of the j_{th} sample has at least one invalid angle. The $Loss_{off}$ makes the teacher give precise corrections of those physiologically invalid predictions of students. The original prediction and corresponding offset are added up to obtain a modification and then a keypoint location loss $Loss_{pos}^{mod}$ is applied to pass this modification to the student network.

Therefore, with the regression loss of the student network $Loss_{pos}^{stu}$, the overall loss function of proposed method is

$$Loss_{total} = Loss_{pos}^{stu} + \lambda_1 Loss_{\theta} + \lambda_2 Loss_{off} + \lambda_3 Loss_{pos}^{mod} \tag{4}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

Three benchmark datasets are used in evaluation: Large-scale Multiview 3D Hand Pose Dataset (LM) [5], Stereo Hand Pose Tracking Benchmark (STB) [25] and Rendered Hand Pose Dataset (RHP) [31].

LM is a real-world hand pose dataset, which is structured in 21 sequences. The 2D and 3D annotations of 21 keypoints are provided. We use the first 20% samples for testing and the rest 80% samples for training.

STB is another real-world dataset. It contains 12 sequences with six difference backgrounds. Stereo and depth images were captured from a Point Grey Bumblebee2 stereo camera and an Intel Real Sense F200 active depth camera simultaneously. 2D and 3D annotations of 21 keypoints are provided. Following the same setting in [3, 10], 10 sequences are adopted for training and the other two for testing.

RHP is a synthetic hand pose dataset. It contains 41,258 training images and 2,728 test images. Precise 2D and 3D annotations of 21 keypoints are provided, as well as the mask of hands and depth images.

The Area Under the Curve (AUC) on Percentage of Correct Keypoints (PCK) and angle violation frequency are used in evaluation. The definitions of these metrics can be found in the supplementary document. The global hand scale and the root depth are assumed to be known in order to report the PCK curve. This is the same condition used in [3].

4.2 Network Architecture

Five different sized networks are constructed as students with different learning capabilities. The first network is a two-stack network proposed by [10], whose multiply-accumulate operations (MAC) are about 120G. Then a stack and the intermediate layers of the first network are removed to obtain the second network, which has around 20G MAC. The next network is obtained by lessening the number of channels per layer of the second network and the MAC is reduced to 500M. Another two networks are constructed with resnet50 [6] and squeezenet [9], and the final fully connected layer is replaced with 2.5D representations [10], which are used to restore the 3D positions of 21 hand joints. For easier representation, the five networks are recorded as TS120G, TS20G, TS500M, Resnet50 and Squeezenet. The detail of these networks and corresponding training setup can be found in supplementary document.

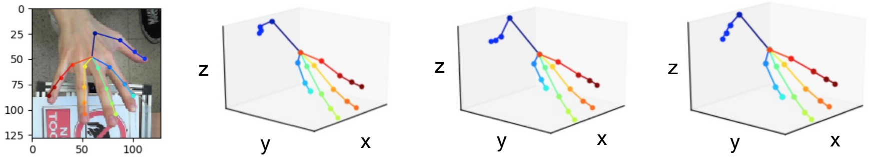
Methods	baseline	w/$Loss_\theta$	w/$Loss_{op}$	w/$Loss_\theta$ +w/$Loss_{op}$
3D AUC	79.3%	82.0%	83.1%	**85.1%**
V.F.	4.83%	2.83%	3.25%	**1.74%**

Table 1: The ablation study of different loss functions of the proposed method. The $Loss_{op}$ represents $Loss_{off} + Loss_{pos}^{mod}$ in the offset network. V.F. is the shorthand of Violation Frequency. The dataset used in this experiment is LM. The backbone is TS500M. 3DAUC is the area under 3D PCK curve between 20-50mm.

4.3 Ablation Study

4.3.1 Different Loss Functions

Several ablation studies with the TS500M model on the LM dataset are conducted to evaluate the role of $Loss_\theta$, $Loss_{op} = Loss_{off} + Loss_{pos}^{mod}$. Results in Table 1 show that adding $Loss_\theta$ provides a 2.7% improvement over the baseline. The proportion of invalid poses has also been reduced. The loss of the offset module $Loss_{op}$ enhances the performance by 3.8% while the angle violation frequency is higher than that of the model with $Loss_\theta$. The advantage of our teacher network is that it not only introduces angle constraints, but also informs the student network the way to correct the invalid poses. Introducing only $Loss_\theta$ is like a teacher only tells the students that the results are wrong, but does not provide the solutions to correct them. An excellent student is capable of recognizing how to fix it, while a student with insufficient ability finds it difficult to correct it by itself. This phenomenon is shown in Figure 3. After adding both $Loss_\theta$ and $Loss_{op}$ during the training, the 3D AUC is increased by **5.80%** over the baseline and the violation frequency is **3.09%** less than the baseline.



(a) Input image (b) Ground truth (c) Wo $Loss_\theta$ (d) With $Loss_\theta$

Figure 3: The schematic diagram of the role of $Loss_\theta$. (a) the input image and its 2D annotations. (b) the 3D annotations of input image. (c) the predictions of the network trained without $Loss_\theta$, where the thumb is in an abnormal state. (d) the predictions after adding the $Loss_\theta$, which are physiologically valid. There is still a gap between (d) and (b).

4.3.2 Different Architectures

For the convenience of presentation, the pose estimation network without EKD are defined as the baselines: a) TS120G + wo/EKD; b) TS20G + wo/EKD; c) TS500M + wo/EKD; d) Resnet50 + wo/EKD; e) Squeezenet + wo/EKD. Those with EKD are recorded as: a) TS120G + w/EKD; b) TS20G + w/EKD; c) TS500M + w/EKD; d) Resnet50 + w/EKD; e) Squeezenet + w/EKD. The algorithm is comprehensively evaluated on the LM [5], STB [25] and RHP [31] datasets. The results are shown in Table 2. Through the experimental results, the 3DAUC of all the models has been improved after adding the proposed teacher network.

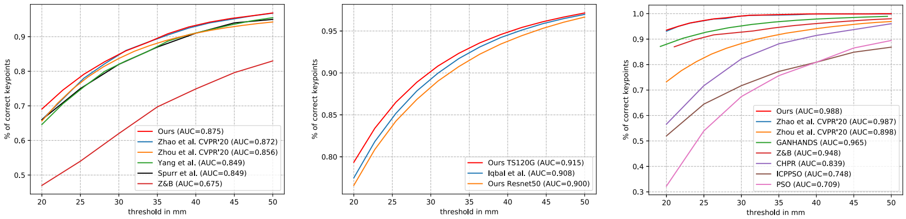
3D AUC/V.F.	LM	STB	RHP
TS120G + wo/EKD	90.8%/1.45%	98.3%/**13.5%**	86.9%/1.74%
TS120G + w/EKD	**91.5%/1.34%**	**99.1%**/14.6%	**87.0%/1.48%**
TS20G + wo/EKD	85.2%/3.06%	84.9%/17.7%	82.5%/2.33%
TS20G + w/EKD	**87.5%/2.40%**	**87.1%/16.7%**	**83.7%/1.82%**
TS500M + wo/EKD	79.3%/4.83%	82.8%/17.2%	61.6%/6.22%
TS500M + w/EKD	**85.1%/1.74%**	**87.8%/13.6%**	**64.3%/4.74%**
Resnet50 + wo/EKD	86.6%/1.48%	95.1%/2.35%	86.4%/1.08%
Resnet50 + w/EKD	**90.0%/0.85%**	**98.8%/1.02%**	**87.5%/0.98%**
Squeezenet + wo/EKD	82.6%/6.94%	82.3%/6.47%	59.6%/2.71%
Squeezenet + w/EKD	**88.5%/2.88%**	**85.9%/3.93%**	**62.1%/1.17%**

Table 2: Comparisons of different baselines with our method on LM, STB and RHP.

And we achieved comparable results to TS120G [10] with a far smaller resnet50 architecture (resnet50+w/EKD). At the same time, the joint angle violation frequency is calculated to verify whether the angle constraints knowledge is passed to the student network. The angle violation frequencies of all the models except the TS120G model trained on STB dataset have been reduced to varying degrees. The STB dataset has more violated samples than others, since the labels of STB dataset are relatively noisy due to the self-occlusion problem and uncontrollable human factors. Therefore, we argue that our algorithm can further improve the performance of STB dataset when the labeling becomes more accurate. Note that our method is a model-agnostic algorithm, in the sense that it is compatible with almost any model for 3D hand pose estimation. Although different networks presents different performance, e.g. TS120G model achieves best 3DAUC and resnet50 model gives lowest violation frequency, the proposed method can always enhance their performance.

4.4 Comparisons with State-of-the-arts

In this section, the proposed method is compared with state-of-the-art (SOTA) hand pose estimation methods, Zhao et al. [28], Zhou et al. [30], Yang et al. [23], Spurr et al. [19], GANHANDS [11], Z&B [31], CHPR [20], ICPPSO [18] and PSO [14]. Following the same criterion of [28], those methods [4, 26] that aim to predict hand shapes, which are with different research targets compared with ours, are not included here. For sake of fairness, we adopt resnet50 network as the backbone, where [11, 28, 30] all used resnet50 network. In addition, LM dataset does not provide depth images, while many SOTA methods rely on depth images during training. Thus, few papers report the results of this dataset and we only compare our method with the method of Iqbal et al. [10] on LM dataset. The results are shown in Figure 4. Our method achieves best results among these SOTA methods. In Figure 4(b), the proposed method achieves comparable performance to the model of Iqbal et al. with a far smaller resnet50 network. As for comparisons with Zhao et al. [28] in Figure 4(a) and Figure 4(c), they used depth images when training on RHP and a additional synthetic dataset when training on STB. Our method does not utilize any extra data during training and achieves comparable performance to their method. We also compare with SOTA knowledge distillation methods and the results are shown in the supplementary document.



(a) RHP (b) LM (c) STB

Figure 4: Comparisons with SOTA methods on three datasets.

5 Conclusion

In this paper, we propose an explicit knowledge distillation method to simulate the teaching process of students in real world. The proposed teacher network helps students generate physiologically valid gestures as well as boost the performance. The effectiveness of the proposed EKD method is validated with different networks on three datasets. We will focus on designing a pose-related network architecture in future studies to further enhance the performance of small networks.

6 Acknowledgements

This research was supported by Tsinghua-Kuaishou Institute of Future Media Data. Li Chen is the corresponding author. She was partially supported by the National Natural Science Foundation of China (Grant Nos. 61972221, 61572274). Junhai Yong was partially supported by the National Natural Science Foundation of China (61672307), National Key R&D Program of China(2019YFB1405703) and TC190A4DA/3.

References

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

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019.

[3] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *the European Conference on Computer Vision*, pages 666–682, 2018.

[4] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019.

[5] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. Large-scale multiview 3d hand pose dataset. *Image and Vision Computing*, 81:25–33, 2019.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

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

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

[9] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

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

[11] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018.

[12] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *the European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[13] Claudia Nolker and Helge Ritter. Visual recognition of continuous hand postures. *IEEE Transactions on Neural Networks*, 13(4):983–994, 2002.

[14] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BmVC*, volume 1, page 3, 2011.

[15] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision*, pages 436–445. IEEE, 2018.

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

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

[18] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1106–1113, 2014.

[19] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2018.

[20] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. Cascaded hand pose regression. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 824–832, 2015.

[21] Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. Robust articulated-icp for real-time hand tracking. In *Computer Graphics Forum*, volume 34, pages 101–114. Wiley Online Library, 2015.

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

[23] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9877–9886, 2019.

[24] Feng Zhang, Xiatian Zhu, and Mao Ye. Fast human pose estimation. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2019.

[25] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing, ICIP 2017, Beijing, China, September 17-20, 2017*, pages 982–986, 2017.

[26] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *the IEEE International Conference on Computer Vision*, pages 2354–2364, 2019.

[27] Yumeng Zhang, Li Chen, Yufeng Liu, Junhai Yong, and Wen Zheng. Adaptive wasserstein hourglass for weakly supervised hand pose estimation from monocular rgb. *arXiv preprint arXiv:1909.05666*, 2019.

[28] Long Zhao, Xi Peng, Yuxiao Chen, Mubbasir Kapadia, and Dimitris N Metaxas. Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6528–6537, 2020.

[29] Xingyi Zhou, Qingfu Wan, Wei Zhang, Xiangyang Xue, and Yichen Wei. Model-based deep hand pose estimation. *arXiv preprint arXiv:1606.06854*, 2016.

[30] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020.

[31] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *the IEEE International Conference on Computer Vision*, pages 4903–4911, 2017.