PMD-Net: Privileged Modality Distillation Network for 3D Hand Pose Estimation from a Single RGB Image

Kewen Wang[1,2]
kewen.wang@vipl.ict.ac.cn

Xilin Chen[1,2]
xlchen@ict.ac.cn

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences(CAS), Inst. of Computing Technology, CAS, Beijing, 100190, China

[2] University of Chinese Academy of Sciences, Beijing, 100049, China

Abstract

3D Hand Pose Estimation from a single RGB image is a challenging task due to the significant depth ambiguities and occlusions. In this paper, we propose a Privileged Modality Distillation Network (PMD-Net), which improves the RGB-based hand pose estimation by excavating the privileged information from depth prior during training. Different from existing methods, the PMD-Net is composed of three sub-networks to regress X, Y, and Z coordinates respectively and distills the privileged information from the depth network to the RGB network by transferring constraints between corresponded layers. Furthermore, a random block replacement is adopted and a refine module is added to enhance the robustness of PMD-Net. Experiments on both synthesized and real-world hand pose estimation datasets are conducted, and extensive results demonstrate that the proposed PMD-Net achieves state-of-the-art results and is superior to existing methods.

1 Introduction

Hand pose estimation progresses rapidly in recent years. With the widespread use of depth cameras, several datasets [21, 24, 25, 26, 34] are available. Benefit from deep neural networks and these datasets, 3D hand pose estimation becomes an active topic, and several methods [4, 11, 17, 19, 29, 30] were proposed. These depth-based methods perform well on hand pose estimation. However, as the accuracy of the captured depth decreases quickly at a distance, it is hard to acquire a high-accurate depth map for hand pose estimation in unconstrained cases. There is still a great need for RGB-based 3D hand pose estimation since RGB cameras can easily deal with objects at a distance with a zoom lens.

Some works, such as [20] only utilizes color images for training. The performance is much worse than depth-based methods due to the significant depth ambiguities and occlusions. As an RGB image doesn't contain explicit depth information and even our human beings can inference depth from various hints, it's a big challenge to estimate depth directly

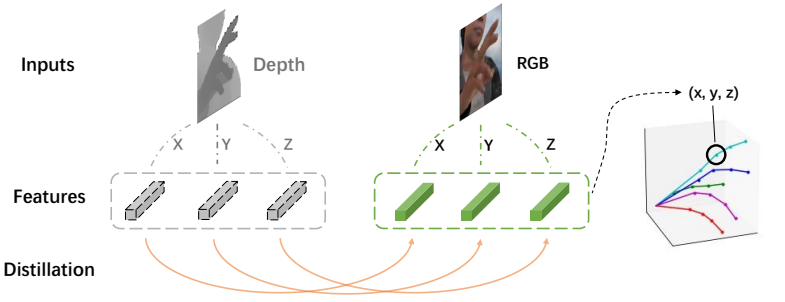


Figure 1: Concept of the proposed method.

from RGB input. Inspired by human beings, several researchers try to apply depth information as prior to improve the accuracy of depth estimation from RGB images. The following works attempt to utilize the depth modality information to enhance training. Gu *et al*. [10] align the hand pose latent space with the depth and RGB modalities by using a cross-modal discriminator with adversarial learning. Yuan *et al*. [35] exploit a pretrained depth-based network and add a middle feature level constraint to supervise the training process of the RGB network.

Meanwhile, privileged modality distillation has been used in challenging situations, in which only limited training data or partially observed modalities are available. For example, privileged modality distillation is used to tackle action detection [5, 23], image classification [14], and vessel border detection [7]. Considering that the RGB-based 3D hand pose estimation task suffers from the significant depth ambiguities in RGB images, we utilize the privileged modality distillation to improve the 3D hand pose predictions in PMD-Net.

Although 2D RGB images provide rich details in texture, they lose the important depth information of objects during the projection from 3D to 2D. However, the lost depth is still implicitly embedded in 2D images in many aspects, such as shadow, reflection, occlusion, motion, and even some prior hints. We human beings learn these cues from experiences. In this work, depth maps are applied as these experiences, and to teach the RGB network during training. More specifically, by utilizing the depth information, the PMD-Net could distinguish the hand from the background and predict hand joints coordinates more precisely, especially for the Z-axis. To this end, we propose to utilize a pretrained depth-based hand pose network and an RGB-based network with privileged modality as side information from depth-based network during training. Only RGB images are used to feed in the RGB-based network to predict 3D hand pose during inference. In short, the PMD-Net is trained in two stages way and makes the predictions in a single end-to-end model.

The concept of the proposed method is shown in Figure 1. During the training stage, a depth network is trained from depth images to form the capability of hand pose estimation. Then, paired RGB and depth images are used by the two networks as inputs simultaneously. The distillation is operated between the depth network whose weights are fixed and the RGB network by constraints from both the middle-layer features and the features before final predictions. In this way, the RGB network learns privileged information from the depth modality to achieve more accurate predictions. Moreover, we adopt a random block replacement strategy to estimate invisible hand joints. To verify the effectiveness of PMD-Net, we perform experiments on two public 3D hand pose estimation datasets: RHD [38] and STB [37]. Extensive results on two datasets demonstrate that PMD-Net achieves state-of-the-art performances.

The main contributions of the proposed method can be summarized as follows:

- A two-stage network with privileged modality distillation. The PMD-Net utilizes privileged depth modality to teach RGB based 3D hand pose estimation. The distillation loss constraints are employed to both the middle-layer features and the features before the final regression layers.
- Two steps for joints' coordinates prediction. In the first step, joints coordinates are predicted with three sub-networks for X, Y, and Z separately. This makes the privileged modality distillation more clear and specific. In the second step, a refine module is added to utilize the correlations among X, Y, and Z.
- Augmentation in training for occlusion. A random block replacement strategy is adopted to the RGB network during training, which enhances the generalization ability of the model to predict invisible hand joints.

2 Related Work

3D hand pose estimation has been studied with great passion in recent years. Benefit from the widely used depth sensor, depth-based 3D hand pose estimation has progressed rapidly. Voxel-based methods [6, 8, 17] convert depth image into voxels and use 3D convolution to capture the spatial representation. Instead of voxels, Ge *et al*. [9] utilize hand pointnet to process the 3D point cloud that models the visible surface of the hand for pose regression. Xiong *et al*. [30] propose an anchor-based approach with 2D convolutions achieving better performance.

Meanwhile, as RGB camera is still the majority, many applications related to hand pose estimation still depends on RGB input. To enhance pose estimation from RGB, depth images can be used as input during training as an implicit label and only RGB images are available in inference. Researchers propose a few approaches trying to tackle this problem and their approaches can be categorized into generative methods and discriminative methods. Generative methods aim at modeling hand and parameters are learned to represent the hand model. Spurr *et al*. [22] learn a statistical hand model represented by a co-trained latent space. A VAE is learned to model the 3D hand representation and the KL-divergence across multiple modalities is jointly optimized. In addition, Yang and Yao [32] propose a disentangled VAE to learn disentangled representations of hand poses and hand images.

As for discriminative methods, researchers directly map the features or heatmaps of the input modality to 3D hand pose predictions. Panteleris *et al*. [20] perform real-time hand pose estimation from single RGB images. Zimmermann and Brox [38] propose a deep network with three steps to estimate 3D hand poses. Firstly a segmentation network is used to predict the hand mask. Next, a sub-network predicts the 2D hand pose heatmap. Finally, a transformation matrix and PosePrior for 3D pose estimation is predicted. Similar to Zimmermann and Brox, Cai *et al*. [2] propose a network to learn 2D heatmaps to regress 3D hand joints and utilize depth maps as a regularizer. They adopt a weakly-supervised training strategy during training and test on single RGB images. Abdi *et al*. [1] use a dual network model to unveil the latent variable of depth for 3D hand pose estimation. They use unsupervised training method with both synthetic and partially-labeled real data.

Vapnik and Izmailov [27] first introduce the concept of training with privileged information (PI), which is similar to the teacher-student relationship in human society, the teacher network with PI teaches the student network during training since the teacher network has

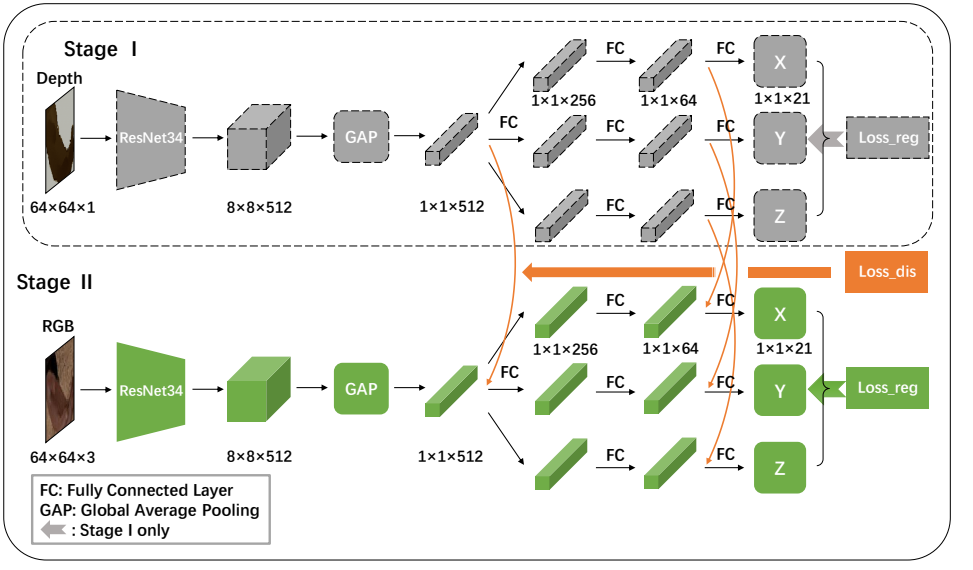


Figure 2: Structure of PMD-Net. (1) In the first stage, a depth-based network is trained with depth input. (2) In the second stage, paired RGB and depth images are used by the two networks as inputs. The weights of the depth network are fixed and the privileged modality distillation is operated between the depth network and the RGB network by constraints from both the middle-layer features and the features before final predictions. (3) During inference, the RGB network estimates the 3D pose from single RGB images.

additional information to teach. During test time, however, the student network should predict on its own as PI is not available. Then, several works support this idea in person re-identification [33], classification [28] and action recognition [36]. Yuan *et al*. [35] utilize a pretrained depth-based network as privileged information and add a middle feature level constraint to supervise the finetuning of the pretrained RGB network. Hinton *et al*. propose the idea of knowledge distillation (KD)[13] to distill the knowledge in an ensemble of models into a student model by adding a loss constraint between the outputs of the student network and the high-temperature outputs of the teacher network. Lopez *et al*. [15] propose a generalized version of distillation which combines knowledge distillation and PI. Luo *et al*. [16] introduce privileged modality distillation to action detection. Different modalities such as RGB, depth, and optical flow are utilized by a distillation graph layer that can dynamically learn to distill knowledge across multiple privileged modalities. Inspired by [16] and [35], we propose a two-stage privileged modality distillation network to distill privileged information in the depth modality. We argue that the middle feature level loss in [35] is ambiguous and nothing related directly to the network output is constrained. Different to [35], we perform privileged modality distillation by constraining the loss between both the middle layer features and the features before the final regression layers.

3 Methodology

The proposed PMD-Net is illustrated in Figure 2. The training of PMD-Net is comprised of two stages. In stage I, we train a depth-based network with depth images from the given

dataset. In stage II, an RGB-based network is trained under the supervision of the depth network and hand joints coordinates. The paired depth and RGB images are used as inputs to the networks and both regression and distillation losses are utilized to constrain the training of the RGB-based network. During inference, the RGB network predicts 3D hand poses from single RGB images.

3.1 Network Architecture

The architecture of PMD-Net is shown in Figure 2. The RGB-based network and the depth-based network share the same structure except for the channels of the input image.

Feature Extractor. Without bells and whistles, we simply exploit a ResNet backbone with a Global Average Pooling (GAP) layer to obtain the feature vector. The input size of the network is 64×64 and a backbone network ResNet-34 [12] is used as a feature extractor. Then we get a feature tensor with the shape of $8 \times 8 \times 512$ and a GAP is operated. After the GAP operation, we extract a middle-layer feature vector of 512 dimensions.

Regression Module. Previous works always predict 3D coordinates together in one single stream network. However, we find that in this situation the prediction of Z coordinates tends to have larger deviations than the X and Y coordinates. In order to make more precise predictions and make the privileged modality distillation more specific, we separate the hand joints coordinates prediction into three sub-networks for X, Y, and Z coordinates regression. Each sub-network is composed of three fully connected layers with 256, 64, and 21 units respectively.

3.2 Privileged Modality Distillation

Given network f and input modality m, we define the regression loss $l_{reg}(f,\, m)$ as:

$$l_{reg}(f,\, m) = (\, ||f^X_{-1}(\boldsymbol{x}^m) - \boldsymbol{y}^X||_2^2 + ||f^Y_{-1}(\boldsymbol{x}^m) - \boldsymbol{y}^Y||_2^2 + ||f^Z_{-1}(\boldsymbol{x}^m) - \boldsymbol{y}^Z||_2^2\,)^{\frac{1}{2}}, \tag{1}$$

where f^i_j denotes the activation value in the j-th layer of the sub-network which predicts i ($i \in \{X,\, Y,\, Z\}$) coordinates and the subscript number -1 denotes the last layer; $\boldsymbol{x}^m$ indicates input sample in m modality and $\boldsymbol{y}^k$ indicates 3D hand pose label in k-axis.

In stage I, we train a depth-based network with the input of the depth modality. Let the depth-based network be ϕ and depth modality be D, the training loss of stage I is simply the regression loss:

$$\begin{aligned} Loss_I &= l_{reg}(\phi,\, D) \\ &= (\, ||\phi^X_{-1}(\boldsymbol{x}^D) - \boldsymbol{y}^X||_2^2 + ||\phi^Y_{-1}(\boldsymbol{x}^D) - \boldsymbol{y}^Y||_2^2 + ||\phi^Z_{-1}(\boldsymbol{x}^D) - \boldsymbol{y}^Z||_2^2\,)^{\frac{1}{2}}. \end{aligned} \tag{2}$$

In stage II, we freeze the parameters in the depth network. To distill the privileged information in the depth modality, the loss constraints are employed to both the middle-layer features and the features before the final regression layers between the network of the RGB and depth modalities. To be more specific, given depth and RGB network ϕ and ψ, the imitation loss of distillation is defined as:

$$\begin{aligned} \mathcal{L}_{dis} = \lambda_1 ||\psi_{feat}(\boldsymbol{x}^R) - \phi_{feat}(\boldsymbol{x}^D)||_1 + \lambda_2 ||\psi^X_{-2}(\boldsymbol{x}^R) - \phi^X_{-2}(\boldsymbol{x}^D)||_1 \\ + \lambda_3 ||\psi^Y_{-2}(\boldsymbol{x}^R) - \phi^Y_{-2}(\boldsymbol{x}^D)||_1 + \lambda_4 ||\psi^Z_{-2}(\boldsymbol{x}^R) - \phi^Z_{-2}(\boldsymbol{x}^D)||_1, \end{aligned} \tag{3}$$

where λ_1, λ_2, λ_3, and λ_4 are hyperparameters to balance the four parts. D and R represent depth and RGB modalities. The subscript $feat$ and -2 denote the 512-dimension feature layer in the middle of the network and the second last layer.

Similar to Eq. 2, the regression loss in stage II is:

$$\begin{aligned}\mathcal{L}_{reg} &= l_{reg}(\psi,\ R) \\ &= (\ ||\psi^X_{-1}(\boldsymbol{x}^R) - \boldsymbol{y}^X||_2^2 + ||\psi^Y_{-1}(\boldsymbol{x}^R) - \boldsymbol{y}^Y||_2^2 + ||\psi^Z_{-1}(\boldsymbol{x}^R) - \boldsymbol{y}^Z||_2^2\ )^{\frac{1}{2}}.\end{aligned} \tag{4}$$

The final loss function of stage II is a combination of regression loss and imitation loss of distillation with hyperparameter α:

$$Loss_{II} = \mathcal{L}_{dis} + \alpha \cdot \mathcal{L}_{reg}. \tag{5}$$

3.3 PMD-Net with Refine Module

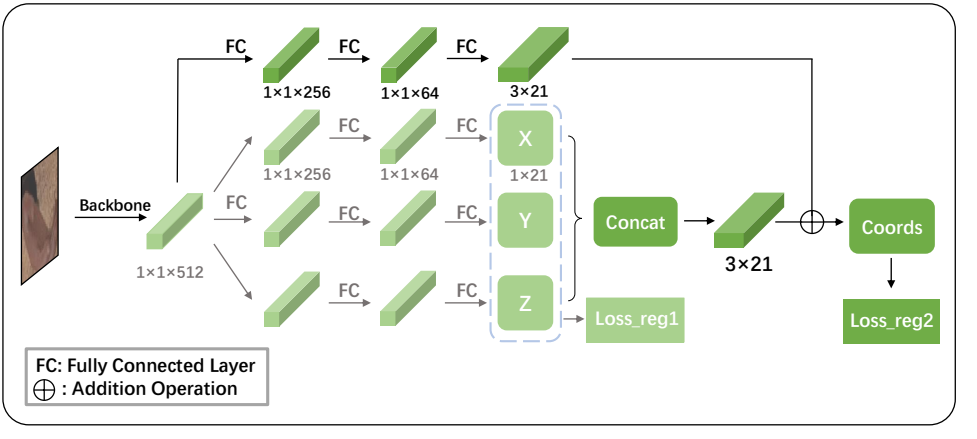


Figure 3: Architecture of PMD-Net with refine module. A refine branch is added on the stage II of PMD-Net to estimate the residual.

As illustrated in Figure 3, in order to further utilize features of RGB input and improve the relevance between X, Y, and Z coordinates, we add a refine module in stage II of PMD-Net.

Refine Module. A refine branch is added on the middle layer of PMD-Net, which has three fully connected layers to predict the residual between XYZ coordinates estimated by PMD-Net and the ground truth. By estimating the residual we can achieve more precise predictions.

Loss Functions. The loss function in stage I is same as PMD-Net which is shown in Eq. 2. In stage II, as shown in Figure 3, the regression loss is composed of two parts: $\mathcal{L}_{reg1}$ and $\mathcal{L}_{reg2}$. We use a hyperparameter β to balance the two regression losses:

$$\mathcal{L}_{reg} = \beta \cdot \mathcal{L}_{reg1} + \mathcal{L}_{reg2}. \tag{6}$$

The distillation loss function and final loss function are the same as Eq. 3 and Eq. 5.

Figure 4: Random block replacement. The pixel values in solid red frame is replaced by the values in dotted red frame.

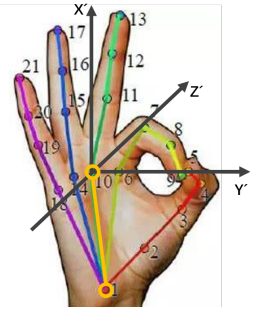


Figure 5: Hand pose normalization. The middle finger MCP joint is shifted to origin and the yellow line is set to 1.

3.4 Random Block Replacement

3D hand pose estimation from RGB images is challenging because of self-occlusion and input ambiguity. Prior work [31] demonstrates that data-distortion operation used in knowledge distillation can improve the performance of the model. By transferring knowledge between different distorted versions, the potential capacity of the network is excavated.

To improve the capability of the RGB network to estimate invisible hand joints and the generalization ability, we propose a random block replacement strategy which is illustrated in Figure 4. We randomly choose a block of $K \times K$ size(the solid red frame in the figure) in the input RGB image and replace its pixel values with the adjacent block on the right(the solid dotted frame in the figure). In this way, the network is guided to learn from distorted images to handle invisible hand joints issues.

4 Experiments

4.1 Dataset and Evaluation Metrics

We evaluate our proposed models on two public datasets: the Rendered Hand Pose Dataset (RHD) [38] and the Stereo Hand Pose Tracking Benchmark (STB) [37].

RHD is a synthesized dataset of rendered hand images from 20 different characters performing 39 actions with various hand sizes. It contains 41238 samples for training and 2728 samples for testing, with a resolution of 320×320. For each pair of RGB and depth images, 3D annotations for 21 hand joints and intrinsic camera parameters are provided.

STB is a real hands dataset. It is composed of a single person's left hand in front of six real-world indoor backgrounds. The dataset has 18000 pairs of RGB and depth images with 3D annotations in 12 sequences. Each of the 12 sequences contains 1500 frames with a resolution of 640×480. In our experiments, we follow [2, 3, 35, 38] and use 10 sequences (15000 samples) for training and the other 2 sequences (3000 samples) for testing.

Evaluation Metrics. To evaluate the accuracy of the estimated 3D hand poses, we use the common metrics: 1) mean end-point-error (EPE) 2) Area Under the Curve (AUC) on the Percentage of Correct Keypoints (PCK) curve. EPE measures the average Euclidean distance between the predicted hand joints and the ground truth hand joints in millimeters. PCK is

the joint success rate that presents the percentage of predicted keypoints that fall within a given threshold range of the Euclidean distance.

4.2 Implementation Details

Hand Pose Normalization. We follow the assumption that the hand root position and the global hand scale are already known, which is used in [2, 22, 32, 35]. As is shown in Figure 5, we perform hand pose normalization that the middle finger metacarpophalangeal (MCP) joint is shifted to the origin and the distance between the wrist joint and the middle MCP joint is set to 1 as in [18].

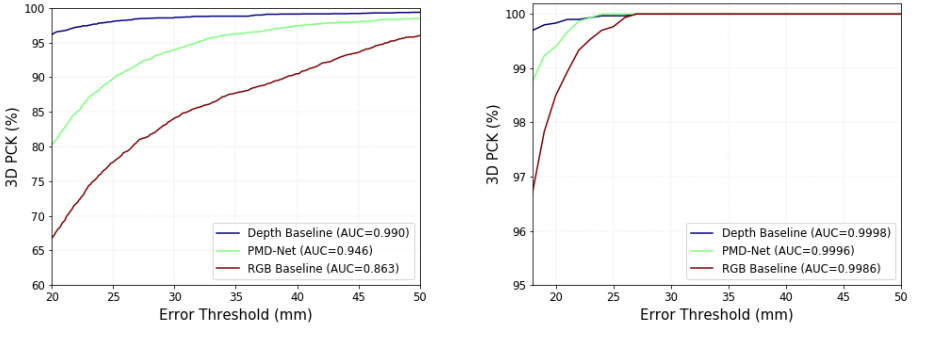
Data Augmentation. In our experiments, we only use image flip augmentation. The image is flipped horizontally and vertically in the RHD dataset and not in the STB dataset.

Random Block Replacement. In our experiments, this module reaches its best performance when the block length k=14 for STB and k=5 for RHD, and the replacement probability is set to 0.5.

We use Pytorch to implement the PMD-Net and train it using an NVIDIA TITAN Xp GPU. We set the hyperparameters λ_i (i from 1 to 4) to (1/16, 5/16, 5/16, 5/16) and α to 100 for all our experiments. In our experiments, we use ADAM as the optimizer with a weight decay of 0.0005 and a learning rate decay strategy is adopted. For the RHD dataset, we set the learning rate as 1e-4 and we decay the learning rate by a factor of 10 after 60K iterations. For the STB dataset, we set the learning rate as 7e-5 and we decay the learning rate by a factor of 10 after 10K iterations.

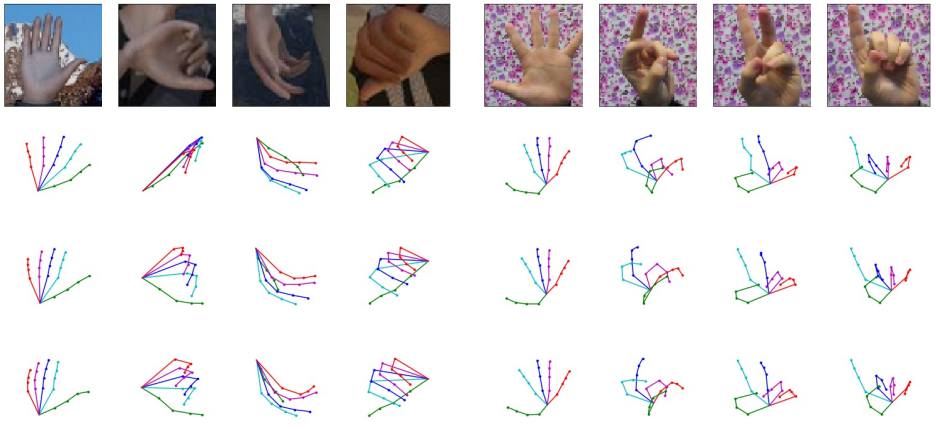
4.3 3D Hand Pose Estimation from RGB images

Effectiveness of Privileged Modality Distillation. We compare our proposed PMD-Net with the RGB baseline network and the depth-based network. The network architectures of the RGB baseline and the Depth baseline are the same as the RGB and the Depth part of the proposed model. Figure 6 shows the PCK curves and AUC values on the 3D hand pose estimation task. The visualization results are shown in Figure 7. Privileged modality distillation improves the performance on the basis of the RGB network and closing the gap to the depth network. This experiment demonstrates convincingly the effectiveness of the privileged modality distillation.



(a) RHD (b) STB

Figure 6: Comparison to baseline networks. The AUC between 20mm and 50mm is showed on the figure legend.



(a) RHD (b) STB

Figure 7: 3D hand pose estimation results on RHD and STB datasets. From top to bottom: RGB images, predictions of RGB baseline network, predictions of PMD-Net, ground truth poses. Note that the center point in RHD dataset is the wrist joint while in STB is the hand's centroid point.

Ablation Study. Firstly, we evaluate the PMD-Net with refine module in different hyperparameters settings. The results are summarized in Table 1 and we compare them with vanilla PMD-Net. In our experiment, we set the hyperparameter β in Eq. 6 to 0.1, 0.3, 0.5, 1, 2 and the network with $\beta = 0.5$ has the best performance.

Method	PMD-Net	PMD-Net with refine module				
		$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 1$	$\beta = 2$
EPE mean (mm)	7.46	7.26	7.12	**6.92**	6.94	7.16

Table 1: EPE mean comparison with different settings.

Then, we conduct an ablation study for different settings on RHD and STB datasets. The results are summarized in Table 2. Compared with the RGB baseline network, the EPE mean is improved by 4.57mm and 2.16mm on RHD and STB, respectively. PMD-Net performs better than the network without 3 sub-networks. The refine module improves the performance on the STB dataset for 0.54mm while decreasing the performance slightly by 0.19mm on the RHD dataset. We think this inconsistent performance is probably because the RHD is a synthesized dataset which has very precise depth information while the STB is a real-world dataset where the captured depth information is not so accurate. As a result, the refine module which relies on RGB features may reduce the effect of distillation from depth on the RHD dataset. Moreover, our PMD-Net has achieved further accurate predictions with the refine module and random replacement strategy.

The test time comparison with different experiment settings is shown in Table 3. The result shows that the three sub-networks and refinement designs only take an extra 0.11ms and 0.08ms per image, respectively, while really contribute to performance a lot. These structures are not very time-consuming because they only use the FC layers with several units.

Method	EPE mean (mm)	
	RHD	STB
RGB Baseline Network	20.83	9.62
PMD-Net without 3 sub-networks	16.52	7.60
PMD-Net	16.26	7.46
PMD-Net + Refine	16.45	6.92
PMD-Net + Refine + Random Replacement	**16.14**	**6.80**

Table 2: EPE mean comparison with different settings. PMD-Net without 3 sub-networks denotes the network which predicts XYZ coordinates together.

Method	Test time (ms/img)
PMD-Net without 3 sub-networks	5.90
PMD-Net	6.01
PMD-Net + Refine	6.09

Table 3: Test time comparison with different settings.

Comparison to state-of-the-art. We compare our method with the state-of-the-art methods in recent years on both the RHD and STB datasets. The comparison results are reported in Table 4. On both datasets, our method achieves better results than all other current methods.

Method	EPE mean (mm)	
	RHD	STB
Zimmermann and Brox [38]	30.42	8.68
Yang and Yao [32]	19.95	8.66
Gu *et al.* [10]	17.11	7.27
PMD-Net	**16.14**	**6.80**

Table 4: Comparisons to state-of-the-art on the RHD and STB with EPE mean.

5 Conclusion

In this paper, we propose an end-to-end network, PMD-Net, which utilizes privileged modality distillation to enhance 3D hand pose estimation from single RGB images. To overcome the problem of the curse of dimensionality, we separate the hand joints predictions into three sub-networks. Moreover, We adopt a random block replacement and a refine module in order to improve robustness to occlusion. Our experimental results demonstrate the effectiveness of the privileged modality distillation as well as refine and replace modules. Compared to existing methods, PMD-Net has achieved better performance and outperforms the previous state-of-the-art ones.

References

[1] Masoud Abdi, Ehsan Abbasnejad, Chee Peng Lim, and Saeid Nahavandi. 3d hand pose estimation using simulation and partial-supervision with a shared latent space. *arXiv preprint arXiv:1807.05380*, 2018.

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

[3] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, Wei Fan, and Xiaohui Xie. Dggan: Depth-image guided generative adversarial networks fordisentangling rgb and depth images in 3d hand pose estimation. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 411–419, 2020.

[4] Xinghao Chen, Guijin Wang, Hengkai Guo, and Cairong Zhang. Pose guided structured region ensemble network for cascaded hand pose estimation. *Neurocomputing*, 2019.

[5] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7882–7891, 2019.

[6] Xiaoming Deng, Shuo Yang, Yinda Zhang, Ping Tan, Liang Chang, and Hongan Wang. Hand3d: Hand pose estimation using 3d neural network. *arXiv preprint arXiv:1704.02224*, 2017.

[7] Zhifan Gao, Jonathan Chung, Mohamed Abdelrazek, Stephanie Leung, William Kongto Hau, Zhanchao Xian, Heye Zhang, and Shuo Li. Privileged modality distillation for vessel border detection in intracoronary imaging. *IEEE transactions on medical imaging*, 2019.

[8] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1991–2000, 2017.

[9] Liuhao Ge, Yujun Cai, Junwu Weng, and Junsong Yuan. Hand pointnet: 3d hand pose estimation using point sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8417–8426, 2018.

[10] Jiajun Gu, Zhiyong Wang, Wanli Ouyang, Jiafeng Li, and Li Zhuo. 3d hand pose estimation with disentangled cross-modal latent space. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 391–400, 2020.

[11] Hengkai Guo, Guijin Wang, Xinghao Chen, Cairong Zhang, Fei Qiao, and Huazhong Yang. Region ensemble network: Improving convolutional network for hand pose estimation. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4512–4516. IEEE, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

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

[14] Saurabh Kumar, Biplab Banerjee, and Subhasis Chaudhuri. Online sensor hallucination via knowledge distillation for multimodal image classification. *arXiv preprint arXiv:1908.10559*, 2019.

[15] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[16] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 166–183, 2018.

[17] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR)*, pages 5079–5088, 2018.

[18] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 49–59, 2018.

[19] Markus Oberweger and Vincent Lepetit. Deepprior++: Improving fast and accurate 3d hand pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 585–594, 2017.

[20] 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 (WACV)*, pages 436–445. IEEE, 2018.

[21] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1106–1113, 2014.

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

[23] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 625–634, 2020.

[24] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. Cascaded hand pose regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 824–832, 2015.

[25] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3786–3793, 2014.

[26] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):1–10, 2014.

[27] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *Journal of machine learning research*, 16(2023-2049): 2, 2015.

[28] Vladimir Vapnik and Rauf Izmailov. Knowledge transfer in svm and neural networks. *Annals of Mathematics and Artificial Intelligence*, 81(1-2):3–19, 2017.

[29] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5147–5156, 2018.

[30] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 793–802, 2019.

[31] Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5565–5572, 2019.

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

[33] Xun Yang, Meng Wang, and Dacheng Tao. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27(2):791–805, 2017.

[34] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4866–4874, 2017.

[35] Shanxin Yuan, Bjorn Stenger, and Tae-Kyun Kim. Rgb-based 3d hand pose estimation via privileged learning with depth images. *arXiv preprint arXiv:1811.07376*, 2018.

[36] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with deeply transferred motion vector cnns. *IEEE Transactions on Image Processing*, 27(5):2326–2339, 2018.

[37] 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)*, pages 982–986. IEEE, 2017.

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