Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8432840, 13 pages
https://doi.org/10.1155/2020/8432840

Research Article

Cascaded Hierarchical CNN for RGB-Based 3D Hand Pose Estimation

Shiming Dai,[1] Wei Liu,[2] Wenji Yang ,[1,3] Lili Fan,[4] and Jihao Zhang[5]

[1]*School of Software, Jiangxi Agricultural University, Nanchang 330045, China*
[2]*School of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang 330045, China*
[3]*State Key Lab of CAD & CG, Zhejiang University, Hangzhou 310058, China*
[4]*School of Information Engineering, Nanchang University, Nanchang 330031, China*
[5]*School of Foreign Languages, Huazhong University of Science and Technology, Wuhan 430074, China*

Correspondence should be addressed to Wenji Yang; ywenji614@jxau.edu.cn

Received 4 April 2020; Accepted 13 June 2020; Published 15 July 2020

Academic Editor: George A. Papakostas

3D hand pose estimation can provide basic information about gestures, which has an important significance in the fields of Human-Machine Interaction (HMI) and Virtual Reality (VR). In recent years, 3D hand pose estimation from a single depth image has made great research achievements due to the development of depth cameras. However, 3D hand pose estimation from a single RGB image is still a highly challenging problem. In this work, we propose a novel four-stage cascaded hierarchical CNN (4CHNet), which leverages hierarchical network to decompose hand pose estimation into finger pose estimation and palm pose estimation, extracts separately finger features and palm features, and finally fuses them to estimate 3D hand pose. Compared with direct estimation methods, the hand feature information extracted by the hierarchical network is more representative. Furthermore, concatenating various stages of the network for end-to-end training can make each stage mutually beneficial and progress. The experimental results on two public datasets demonstrate that our 4CHNet can significantly improve the accuracy of 3D hand pose estimation from a single RGB image.

1. Introduction

The hand is the most active organ for humans. Therefore, the gesture is one of the main expressions of human beings, which accounts for the largest proportion of all human posture. With the rapid development of computer vision technology, 3D hand pose estimation is gradually applied to the fields of Human-Machine Interaction (HMI), Virtual Reality (VR), and Augmented Reality (AR) [1–3], which makes vision-based 3D hand pose estimation become an active research area [4], and has achieved great progress after years of research [5–13]. However, this research is still very challenging due to the diversity of gestures, the significant flexibility of finger joints, the high similarity between fingers and severe self-occlusion. In recent years, research on 3D hand pose estimation based on depth images is progressing rapidly with the development of the depth cameras [14–16]. Firstly, the depth information from the depth image is beneficial for 3D hand pose estimation. Secondly, the emergence of cheap depth cameras significantly reduces the difficulty of obtaining depth data, which greatly reduces the production cost of depth data. As a result, 3D hand pose estimation based on depth images has achieved a great many of results [17–21] during this period. Compared with depth images, RGB images lack depth information, which makes it difficult to estimate 3D hand pose directly from 2D RGB images. Therefore, the result of current 3D hand pose estimation based on RGB images is not ideal enough. But 3D hand pose estimation based on RGB images is more realistic because the application based on RGB images is more widespread and the number of users using RGB images is larger. In this paper, we present a four-stage cascaded hierarchical CNN (4CHNet) for RGB-based 3D hand pose estimation. We cascade four stages of the network for end-

to-end training. The four stages include hand mask estimation stage, 2D hand pose estimation stage, hierarchical estimation stage, and 3D hand pose estimation stage. According to the back-propagation mechanism of the neural network, the mutual promotion and common progress can be achieved by each stage. The hierarchical estimation stage processes hand feature extracted hierarchically to extract more effective, deeper, and more representative feature information and finally fuses the feature information of all layers to estimate the 3D hand pose to improve the estimation accuracy of the 3D gesture. Our contributions can be summarized as follows:

(1) We propose a 4CHNet for RGB-based 3D hand pose estimation in which hand pose estimation is divided into two subtasks by using hierarchical thinking, namely, finger pose estimation and palm pose estimation. More representative finger features and pam features are extracted, respectively, and finally fused to estimate the 3D hand pose, which can improve estimation accuracy of 3D gestures.

(2) Four-stage cascaded training, which cascades four stages including hand mask estimation stage, 2D hand pose estimation stage, hierarchical estimation stage, and 3D hand pose estimation stage for end-to-end training, is proposed. According to the back-propagation mechanism, each stage is mutually beneficial and progressive together in the training process to achieve the global optimization and refine the models.

(3) Based on the hierarchical network, 2D finger heatmaps and 2D palm heatmaps are estimated. These two constraints enable the hierarchical network to conduct feature stratification and further estimate 3D finger pose and 3D palm pose. The network can perform better in feature extraction and 3D hand pose estimation by introducing four new constraints.

(4) We conduct experiments on two public datasets, and the results show that our 4CHNet can achieve better 3D hand pose estimation accuracy than previous works.

2. Related Work

Following recent trends in computer vision, methods for 3D hand pose estimation from different input images can be categorized into RGB-based estimation methods [22–30], depth-based estimation methods [17–21], or RGB-D based estimation methods [9, 31, 32]. Because depth information is helpful for 3D estimation, most of previous works are based on the depth image. However, it still has certain practical application limitations. Currently, the research focus is gradually shifting to the RGB-based 3D hand pose estimation.

2.1. Estimation Method Based on RGB Images. Estimating 3D hand pose directly from a single RGB image is far more challenging due to the absence of depth information. Subsequently, researchers have presented different estimation methods. Zimmermann and Brox [23] firstly applied a deep neural network to 3D hand pose estimation based on single RGB images. They used three deep networks to cover important subtasks on the way to the 3D pose. The three networks are hand localization segmentation network, 2D hand pose estimation network, and 3D hand pose estimation network. Spurr et al. [33] extended VAE framework via training several pairs of encoder and decoder to form a joint cross-modal latent space representation and estimated 3D hand pose of the input depth images and RGB images. Since full 3D meshes of hand surface can determine the shape of hands, it is of great help for 3D hand pose estimation. Using 3D meshes to estimate 3D hand pose has been extensively studied recently. Ge et al. [28] added a 3D hand mesh estimation stage in which the Graph CNN [34] uses heatmaps and hand features as input and estimates the full 3D mesh of hand surface which is further used to regress the 3D gesture. Boukhayma et al. [30] leveraged a deep convolution encoder to estimate hand shape parameters and gesture parameters and then fed these parameters to a pretrained hand mesh model to estimate the mesh of hand surface and further estimate 3D hand pose after obtaining hand shapes. Although accurate hand mesh greatly improves the estimation accuracy of 3D gesture, it is hard to generalize estimation methods from hand meshes due to the difficulty of obtaining the hand surface mesh labels. Our early work [35] proposed a three-stage cascaded CNN mask-2d-3d, which cascaded mask estimation stage, 2D hand pose estimation stage, and 3D hand pose estimation stage to estimate 3D hand pose. Here we need to emphasize the difference between our proposed method and the earlier work of mask-2d-3d. Firstly, we add a hierarchical network to form a four-stage cascaded network, which divides 21 key points into 15 key points of finger layer and 6 key points of the palm layer to extract deeper finger features and palm features and then fuses them to estimate more accurate 3D gestures. Secondly, we add 2D palm heatmaps, 2D finger heatmaps, 3D palm poses, and 3D finger poses constraints to train the network effectively. Here, we need to emphasize the differences between us and Zimmermann and Brox [23]; their method was proposed earlier and also has some defects. They trained their networks separately in each estimation stage, which makes estimation effect of each stage reach the local optimum rather than the global optimum. To overcome this shortcoming, we use a 4CHNet, which affects mutually and progresses together to achieve global optimization of 3D hand pose estimation. The second difference is that Zimmermann and Brox [23] only used two simple constraints: 2D hand heatmaps and 3D gestures. However, the two constraints are really difficult to extract deeper features. Differently, we address that the estimation accuracy would be dramatically improved by adding 2D finger heatmaps, 2D palm heatmaps, 3D finger poses, and 3D palm poses constraints via using a hierarchical network, while introducing hand masks and employing hand masks and 2D heatmaps to further guide feature extraction.

2.2. Estimation Method Based on Hierarchical Thinking. Hierarchical network is spurred by the multitask sharing mechanism. In machine learning, multitask sharing has the

advantages of reserving more intrinsic information than single-task learning [36]. The hierarchical network divides hand pose estimation task into several subtasks according to the structure of hand, which extracts more intrinsic information through multiple subtasks and finally shares information to estimate 3D hand pose. Guo et al. [37] proposed a region ensemble network, which simply divided the extracted feature maps into four grid regions of 2×2, and features of each region were fed into FC layers for the ensemble. The method can effectively improve performance without extra heavy computational cost. Madadi et al. [38] firstly divided the hand features into six layers, of which five layers were used to model each finger, and the remaining layer was used to model palm orientation features. Then, the six layers were combined to estimate all joint positions. Zhou et al. [39] divided five fingers into three layers according to the sensitivity and function of fingers, where one layer was correlated with thumb finger, one layer modeled the index finger, and the final layer represented the remaining three fingers. Finally, three layers were combined to estimate the hand pose. Du et al. [40] divided the features of the hand into two layers, that is, finger feature and palm feature, and used a cross-connected network to refine the two-layer features and finally fused them to estimate the hand pose. Our 4CHNet is the closest to Du et al. [40]. Here, we also need to emphasize the difference. Firstly, our method is based on 3D hand pose estimation of RGB images. However, the method proposed by Du et al. [40] is based on depth images. Secondly, we use a 4CHNet, exploiting the hand mask estimation, 2D hand pose estimation, hierarchical estimation and 3D hand pose estimation to estimate 3D gesture jointly, which is essentially different from the network architecture of Du et al. [40].

3. Four-Stage Cascaded Hierarchical CNN

3.1. Overview. We propose a 4CHNet for estimating 3D hand pose from a single RGB image, as illustrated in Figure 1. Firstly, we use a localization segmentation network to localize and crop the hand of the RGB image for preprocessing RGB images. The cropped RGB image is used as the input of 4CHNet to estimate hand masks, 2D hand heatmaps, 2D finger heatmaps, 2D palm heatmaps, 3D finger poses, and 3D palm poses and then to estimate the full 3D hand poses through fusing 3D poses of fingers and palms.

3.2. Localization and Segmentation Network. The localization segmentation network is used to determine the location of hand, and then the low-resolution hand is obtained and enlarged, which is the basis for subsequent gesture estimation. If there is no appropriate localization segmentation network, the accurate 3D hand pose estimation will also lack practical significance. We use a simplified version of Convolutional Pose Machines [41] as the localization segmentation network and extract the spatial features of hand by estimating two-channel hand masks. Furthermore, the loss is calculated by hand mask labels to feedback the network to achieve the goal of training a localization segmentation network. Through the estimated hand mask, we can locate the hand in RGB image and then crop and resize the hand to 256×256 size.

3.3. 4CHNet. We intend to use the principle of the cascade into our overall network, cascading four stages for end-to-end training. The four stages include hand mask estimation stage, 2D hand pose estimation stage, hierarchical estimation stage, and pose estimation stage, respectively. Furthermore, four stages can benefit mutually and progress together, thereby achieving global optimization and the goal of improving the accuracy of 3D hand pose estimation.

3.3.1. Hand Mask Estimation Stage. In the hand mask estimation stage, we use a simplified version of VGG-19 network [42]. Both 128-channel image feature $F1$ and 2-channel spatial feature, namely, hand mask M, are extracted by convolution, and mask labels of dataset are used to train the network. Hands can be better tracked through the spatial feature, which is helpful for subsequent hand pose estimation.

3.3.2. 2D Hand Pose Estimation Stage. 2D hand pose estimation stage consists of five substages. In the first substage, it takes 130-channel features S as input, which consisted of 128-channel image features $F1$ and 2-channel spatial features M extracted from mask estimation stage and then outputs 21-channel heatmaps. In the last four substages, 21-channel hand heatmaps estimated from the previous stage and 130-channel image feature S are connected to form 151-channel feature which is taken as the input to estimate five substages 2D hand heatmaps. We use the final substage hand heatmaps as the final output and then use 2D labels of datasets to train the network.

3.3.3. Hierarchical Estimation Stage. The hierarchical estimation stage is similar to the 2D hand pose estimation stage, both of which estimate 2D heatmaps, but the hierarchical estimation stage divides features of hands into two layers: finger features and palm features. The 21 key points of hands are shown in Figure 2(a). We divide 6 key points into palm key points and the remaining 15 key points into finger key points. The key points division demonstration of the real dataset STB is shown in Figure 2(b). And the key points division demonstration of the synthetic dataset RHD is shown in Figure 2(c). The left side of the demonstration is an example of finger key points, and the right is an example of palm key points.

The hierarchical network estimates 2D finger heatmaps and 2D palm heatmaps independently and helps to further estimate 3D finger pose and 3D palm pose (see Figure 3). There are three substages in each layer of this stage. Taking the finger layer as an example, firstly, the first substage connects 130-channel feature S and 21-channel hand heatmaps outputted from the previous stage to form 151-channel full hand feature F, which is as the input to estimate 15-channel finger heatmaps F_{f1}. Then, the last two substages connect the 15-channel finger heatmaps obtained from the

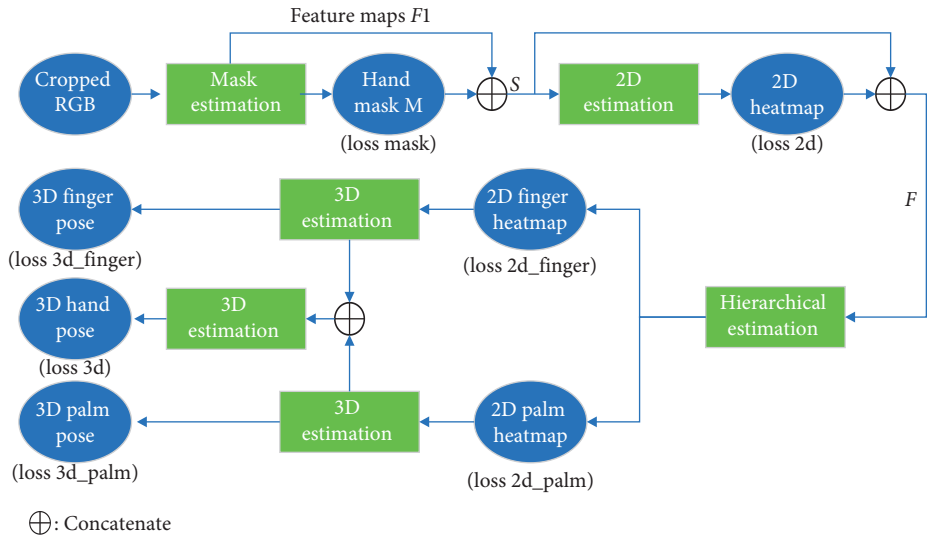

FIGURE 1: Basic framework diagram of 4CHNet. The cropped color images are used as the input of 4CHNet to, respectively, estimate masks of hands and heatmaps of hand through the mask estimation stage and the 2D hand pose estimation stage, then to estimate the 2D heatmaps of fingers and palms through the hierarchical estimation stage, and finally to estimate 3D hand poses, 3D finger poses, and 3D palm poses through the 3D hand pose estimation stage, respectively.

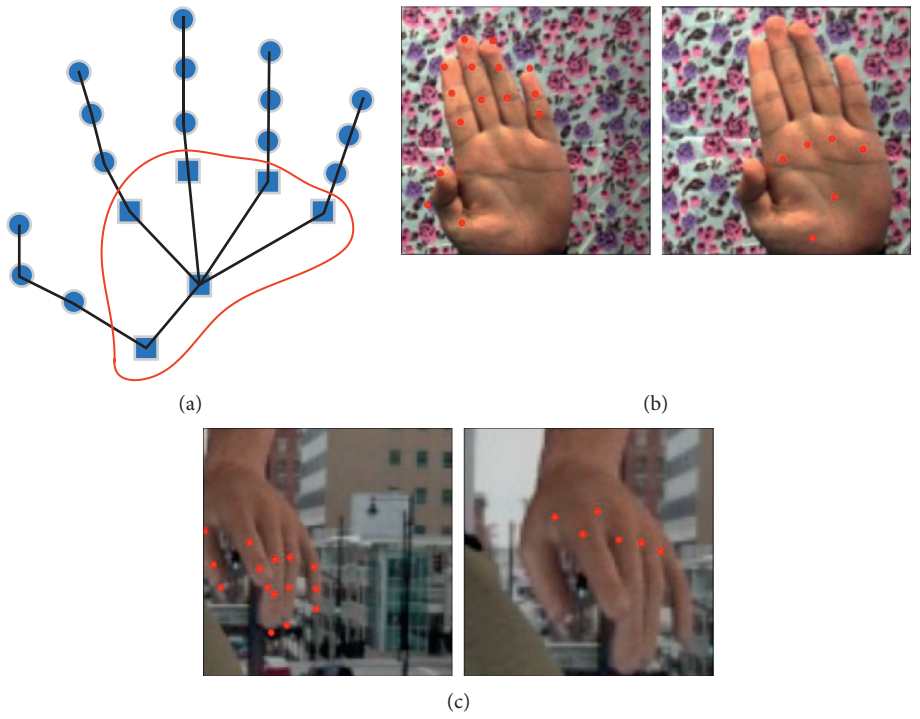

FIGURE 2: The diagram and examples of hand key points division. (a) Skeleton graph of 21 key points of the hand, in which squares represent key points of the palm and circles represent key points of fingers. (b) An example diagram of finger key points and palm key points from the real dataset STB. (c) An example diagram of finger key points and palm key points from the synthetic dataset RHD.

previous stage with 151-channel full hand feature F as input. Finally, a total of three substages finger heatmaps are estimated, and the final substage estimated finger heatmaps F_{f3} are as the output. The principles employed for the finger layer is the same as the palm layer. Here, we use 2D finger and 2D palm labels of datasets to train the hierarchical

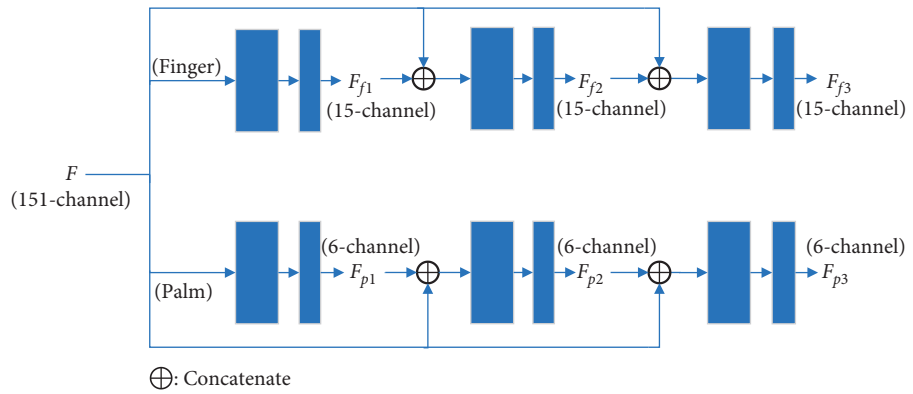

Figure 3: Hierarchical network structure diagram. Finger features and palm features are extracted by the hierarchical network.

network. F represents full hand features, F_f represents finger features, F_p represents palm features, $C15$ and $C6$ represent the convolutional neural network which is employed to extract features of fingers and palms, respectively:

$$\begin{aligned} F_f &= C15(F), \\ F_p &= C6(F). \end{aligned} \tag{1}$$

3.3.4. 3D Hand Pose Estimation Stage. The 3D hand pose estimation stage takes 2D finger heatmaps and 2D palm heatmaps outputs of the hierarchical network as inputs to estimate 3D finger poses and 3D palm poses and fuses them to estimate the 3D hand pose. We employ the method proposed by Zimmermann and Brox [23] to represent the 3D pose. In order to estimate the relative normalized coordinates x^{rel} of key points, the first bone's length of index finger $\|x_{k+1} - x_k\|$ is selected as the standard length. x_{k+1} and x_k represent the two endpoints of the first bone of the index finger and palm point x_r as origin:

$$x_i^{\text{rel}} = \frac{x_i - x_r}{\|x_{k+1} - x_k\|}. \tag{2}$$

In order to facilitate the estimation of hands with different poses, the relative normalized coordinates x^{rel} are rotated by using a 3D rotation matrix R to obtain the canonical coordinates x^c. The gesture directions of these canonical coordinates are consistent, which is convenient for 3D hand pose estimation. We estimate the canonical coordinates x^c and 3D rotation matrix R to indirectly estimate the relative normalized 3D coordinates x^{rel} of the 21 key points:

$$x^{\text{rel}} = x^c \cdot R^T. \tag{3}$$

3.4. Design of Loss Function

3.4.1. Estimation Loss of Mask. The mask estimation loss loss mask uses *standard softmax cross-entropy* loss, where y is its label, s_u is output score of the uth label in the mask estimation stage, and the mask is a binary map, $u \in \{0, 1\}$:

$$\text{loss of mask} = -\sum \log\left(\frac{e^{s_y}}{\sum_u e^{S_u}}\right). \tag{4}$$

3.4.2. Estimation Loss of Mask. A squared *L2* loss is imposed on the 2D heatmaps loss of 21 key points to calculate the estimation loss of 2D hand pose loss 2d, where pre_j is estimated 2D hand heatmaps and gt_j is its corresponding label, and j represents the index of key point:

$$\text{loss 2d} = \left\|\text{pre}_j - \text{gt}_j\right\|_2. \tag{5}$$

3.4.3. Estimation Loss of Hierarchical. The estimation loss of hierarchical loss L is sum of the loss of 2D finger heatmaps loss 2d_finger and the loss of 2D palm heatmaps loss 2d_palm, which is calculated by using L2 loss, where pre_{j_f} and pre_{j_p} are estimated 2D finger heatmaps and 2D palm heatmaps respectively, and gt_{j_f} and gt_{j_p} are their corresponding 2D key points label of finger and palm separately, j_f represents finger key points, and j_p represents palm key points:

$$\begin{aligned} \text{loss 2d\_finger} &= \left\|\text{pre}_{j\_f} - \text{gt}_{j\_f}\right\|_2, \\ \text{loss 2d\_palm} &= \left\|\text{pre}_{j\_p} - \text{gt}_{j\_p}\right\|_2, \\ \text{loss } L &= \text{loss 2d\_finger} + \text{loss 2d\_palm}. \end{aligned} \tag{6}$$

3.4.4. Estimation Loss of 3D Hand Pose. The estimation loss of 3D hand pose loss 3d includes estimation loss of 3D finger pose loss 3d_f, 3D palm pose loss 3d_p, and full hand pose loss 3d_h, which is computed by using the squared *L2* loss for canonical coordinate x^c and 3D rotation matrix R, respectively. The estimation loss of 3D finger pose is

$$\text{loss 3d\_}f = \left\|x^c_{\text{pre\_}f} - x^c_{\text{gt\_}f}\right\|_2^2 + \left\|R_{\text{pre\_}f} - R_{\text{gt\_}f}\right\|_2^2. \tag{7}$$

The estimation loss of 3D palm pose is

$$\text{loss 3d\_}p = \left\|x^c_{pre\_p} - x^c_{gt\_p}\right\|_2^2 + \left\|R_{pre\_p} - R_{gt\_p}\right\|_2^2. \tag{8}$$

The estimation loss of full hand pose is

$$\text{loss 3d\_}h = \left\| x^c_{\text{pre\_}h} - x^c_{\text{gt\_}h} \right\|_2^2 + \left\| R_{\text{pre\_}h} - R_{\text{gt\_}h} \right\|_2^2. \quad (9)$$

The sum of 3D estimated loss is

$$\text{loss 3d} = \text{loss 3d\_}f + \text{loss 3d\_}p + \text{loss 3d\_}h. \quad (10)$$

The total loss of 3D hand pose estimation is

$$\text{loss} = v^* \text{loss mask} + \text{loss 2d} + \text{Loss } L + \text{Loss 3d}. \quad (11)$$

Because the loss value of loss mask is large, we add a weight ratio v to this item to reduce its loss value. It is found that $v = 0.05$ can achieve a best result via a large number of experiments.

4. Experiments

4.1. Datasets

4.1.1. OneHand10 K. *OneHand10 K* dataset [27] is one single-handed RGB-based dataset, hereinafter, referred to as OHK. Images in OHK are real images, including 10000 images for training, and the remaining 1703 images are used for testing, which are captured under different backgrounds and lighting conditions. Each RGB image has a corresponding mask label and 2D labels for 21 key points. In this work, we use hand mask labels of real dataset OHK to train localization segmentation network for the purpose of enhancing adaptability of the network in a real world and then employ localization segmentation network to localize the hand of RGB image and crop and enlarge hand size to get cropped RGB image for facilitating subsequent accurate 3D hand pose estimation. Because image resolution of this dataset is not uniform, we have adjusted and filled the OHK data. The size of unified OHK image is 320×320, and the adjustment ratio is m, where w and h are original width and height of the image. After the ratio is adjusted, we fill the lower right corner of the RGB image with gray value (128,128,128), zero-fill the lower right corner of the mask, and finally output the RGB image with a resolution of 320×320 and its corresponding mask:

$$m = \min\left(\frac{320}{w}, \frac{320}{h}\right). \quad (12)$$

4.1.2. RHD. Rendered Hand Pose Dataset (RHD) [23] is a synthetic RGB image based hand dataset, which is composed of 41258 images for training and 2728 images for testing with a resolution of 320×320, and it is obtained by requiring 20 different human models randomly to perform 39 different actions and randomly generate arbitrary backgrounds. The dataset is considerably challenging due to large variations in viewpoints and hand proportion, as well as large visual diversity induced by random noise and ambiguity of the images. For each RGB image, it provides corresponding depth image, mask label, 2D label, and 3D label of 21 key points of the hand. We use the mask labels, 2D labels, and 3D labels to train the entire network. However, due to a certain gap between the synthetic data and real data, it is difficult for a network trained by synthetic data to adapt directly to the real world, so it is necessary to use real data for adaptive adjustment later.

4.1.3. STB. Stereo Hand Pose Tracking Benchmark (STB) [43] is a real RGB image hand dataset containing two subsets: the stereo subset STB-BB captured from the stereo vision camera and the color-depth subset STB-SK captured from the Intel active depth camera. Since no deep data is used in our method, we only use the subset STB-BB. STB-BB has a total of 36000 images which is divided into 12 pairs. Following the same condition used in [23], we use 10 parts of 30000 images as training set and the remaining 2 parts of 6000 images as testing set. Each RGB image of this dataset has 2D and 3D labels of 21 key points of the hand and corresponding depth map, but we only use its 2D and 3D labels. On the basis of RHD training using synthetic dataset, we use real dataset STB to refine model and make the model adapt to the real world.

4.2. Evaluation Metric. We evaluate our proposed 4CHNet on two public datasets, RHD and STB, by using two evaluation metrics:

(1) Endpoint error (EPE), which includes the average endpoint error (EPE mean) and median endpoint error (EPE median)

(2) The area under the curve (AUC) on the percent of correct key points (PCK). Our evaluation fully adopts the same metrics as [23]

4.3. Experimental Details. Our 4CHNet is implemented by Tensorflow [44] on a single server with single GPU of Nvidia RTX2080Ti for training and testing.

4.3.1. Localization Segmentation Network Training Details. We use real dataset OHK with mask label to train the localization segmentation network. A batch size of 8 and an initial learning rate of 1×10^{-5} are employed for training 40 K iterations. To prevent overfitting, we have set decay ratio as 0.1. Learning rate is 1×10^{-6} for the first 20 K iterations and then decays every 10 K iterations.

4.3.2. Training Details of 4CHNet

(1). Pretraining on Synthetic Dataset RHD. We adopt synthetic dataset RHD to pretrain the 4CHNet and use mask labels and 2D and 3D labels of dataset to supervise the training. The training batch size is 8 and an initial learning rate is 5×10^{-5} for training 300 K iterations, while the decay ratio of learning rate is 0.3, which decays every 50 K iterations.

(2). Refinement on the Real Dataset. Based on the RHD pretrained network, in order to adapt the model to the real world, we use a real dataset STB to refine the model by using its 2D and 3D label to train the network for training 250K

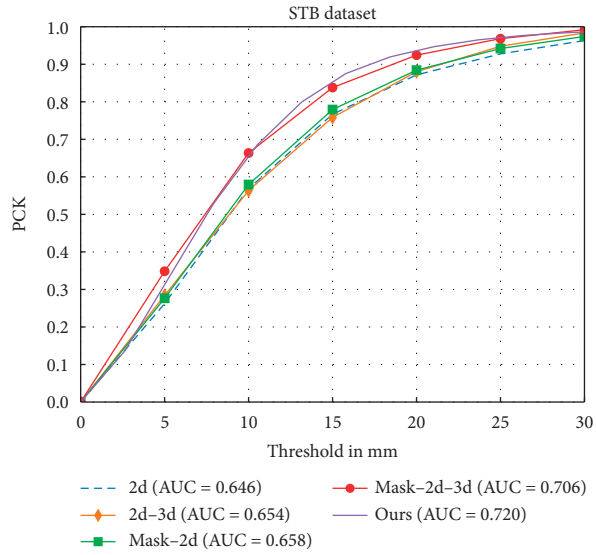

Figure 4: Self-comparison experiment.

iterations. The remaining training parameters are consistent with that of the pretraining stage.

4.4. Self-Comparison Experiment. Our early work [35] has experimented on a three-stage cascaded network and compared ablation experiments with other methods, which has demonstrated the effectiveness of newly added mask estimation stage and cascaded network. On this basis, we propose a four-stage cascaded network and compare it with the three-stage cascaded network to demonstrate the effectiveness of the newly added hierarchical network. In this experiment, we also designed the other four network training methods, where 2d means that 2D and 3D networks are trained separately without a mask estimation stage, and *mask-2d* means mask estimation and 2D hand pose estimation are trained jointly, while 3D estimation stage is trained alone; *2d-3d* represents the cascaded training of 2D and 3D estimation without mask estimation stage, *mask-2d-3d* represents a three-stage cascaded network, and *Ours* is 4CHNet we have proposed. Previous work [35] has verified the superiority of OHK for training segmentation networks, so our experiment uses localization segmentation network trained by OHK, fuses RHD and STB to train networks, and keeps the parameters consistent. Figure 4 and Table 1 show the experimental results. The experimental results show that the AUC of four-stage cascaded network denoted by *Ours* reaches 0.720 and 0.822 within the error threshold of 0–30 mm and 0–50 mm, which is higher than 0.706 and 0.811 of three-stage cascaded network mask-2d-3d and far higher than that of other network structures. The average endpoint error of our four-stage cascaded network is reduced to 8.878 mm, which is reduced by 5.53% compared with 9.398 mm of three-stage cascaded network and the median endpoint error of the two networks is similar. This self-comparison experiment verifies the superiority of proposed 4CHNet over the three-stage cascaded network. Because of newly added hierarchical network, 2D finger heatmaps constraint, 2D palm heatmaps constraint, 3D finger poses constraint, 3D palm poses constraint and four-stage cascaded, and estimation accuracy of 3D key points have greatly been improved.

4.5. Comparison with Other Methods. We compare our 4CHNet on two public datasets with most of state-of-the-art methods [23, 35] on RHD and state-of-the-art methods [23, 25, 33, 35, 43, 45] on STB and the comparison adopts the same evaluated metrics in [23]. Particularly, we use a localization segmentation network to locate the hand in the image instead of directly processing the original image; therefore, in addition to the pose estimation error, a part of our total error also comes from hand positioning. The methods involved in the comparison also need to add localizing errors if they also have a localization segmentation network. The comparison experiment results on the synthetic dataset RHD are shown in Figure 5. The results show that the 4CHNet achieves an AUC of 0.770 within the error threshold 20–50 mm, which is significantly better than that of the state-of-the-art method.

Figure 6 shows a comparison test on the STB dataset. *Ours* and *Ours (without OHK)* both represent 4CHNet, and both fuse the synthetic dataset RHD and the real dataset STB for training, of which *Ours* uses OHK to train localization segmentation network to achieve a more accurate hand localization in a real world, while *Ours (without OHK)* uses the localization segmentation network model of [23], which only uses synthetic dataset RHD for training the localization segmentation network. The mask-2d-3d and mask-2d-3d *(without OHK)* represent three-stage cascaded network; the latter one uses a localization segmentation network model of [23]. The experimental results show that the AUC of *Ours* reaches 0.988, which is a significant improvement over 0.948 in Zimmermann and Brox [23] and 0.977 in the three-stage cascaded network. At the same time, it is also better than the state-of-the-art result on STB dataset, which verifies the

Table 1: The error analysis of self-comparison experiment.

Network	AUC (0–30 mm)	AUC (0–50 mm)	EPE mean (mm)	EPE median (mm)
2d	0.646	0.768	11.653	9.619
2d-3d	0.654	0.777	11.134	9.773
mask-2d	0.658	0.777	11.189	9.298
mask-2d-3d	0.706	0.811	9.398	7.821
Ours	**0.720**	**0.822**	**8.878**	**7.830**

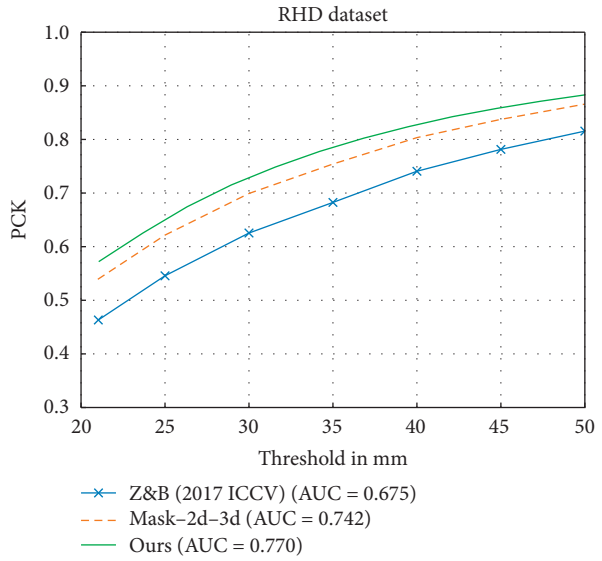

Figure 5: Comparative experiment on synthetic dataset RHD.

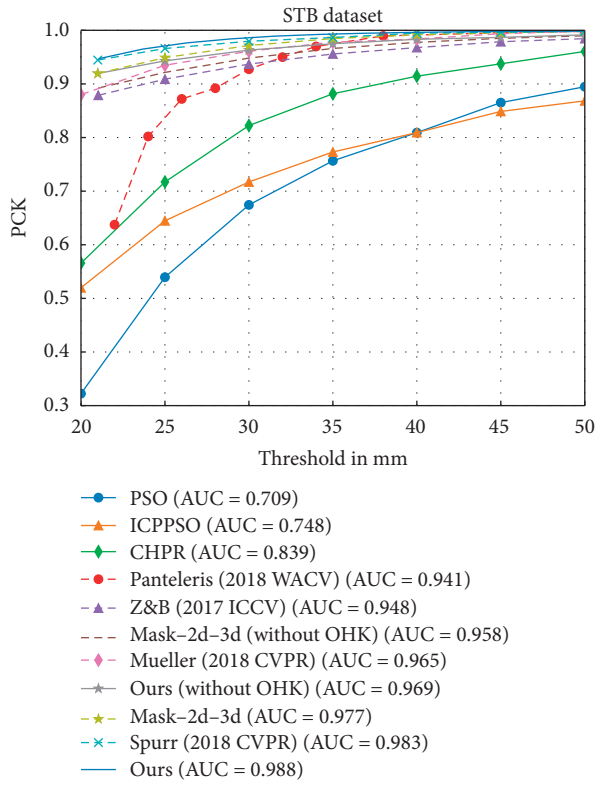

Figure 6: Comparative experiment on real dataset STB.

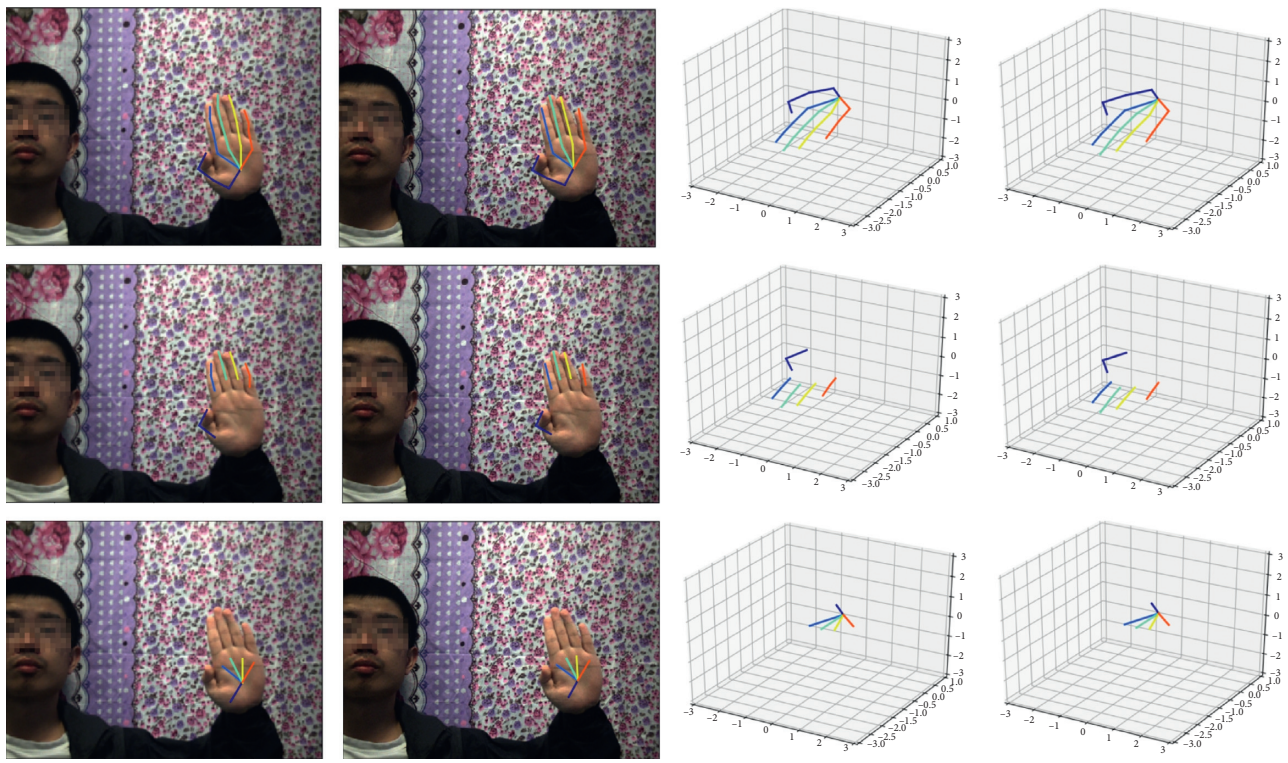

Figure 7: The qualitative results on STB.

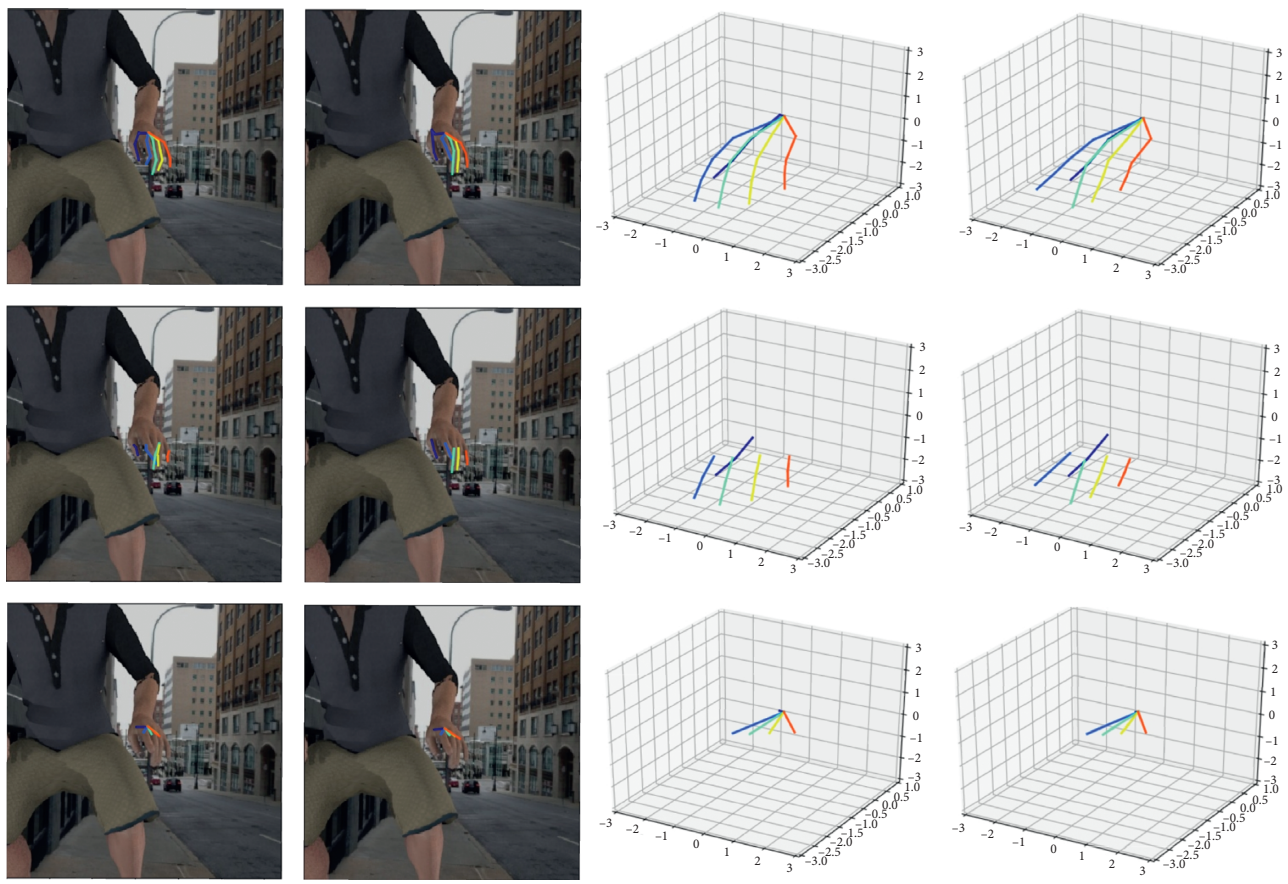

Figure 8: The qualitative results on RHD.

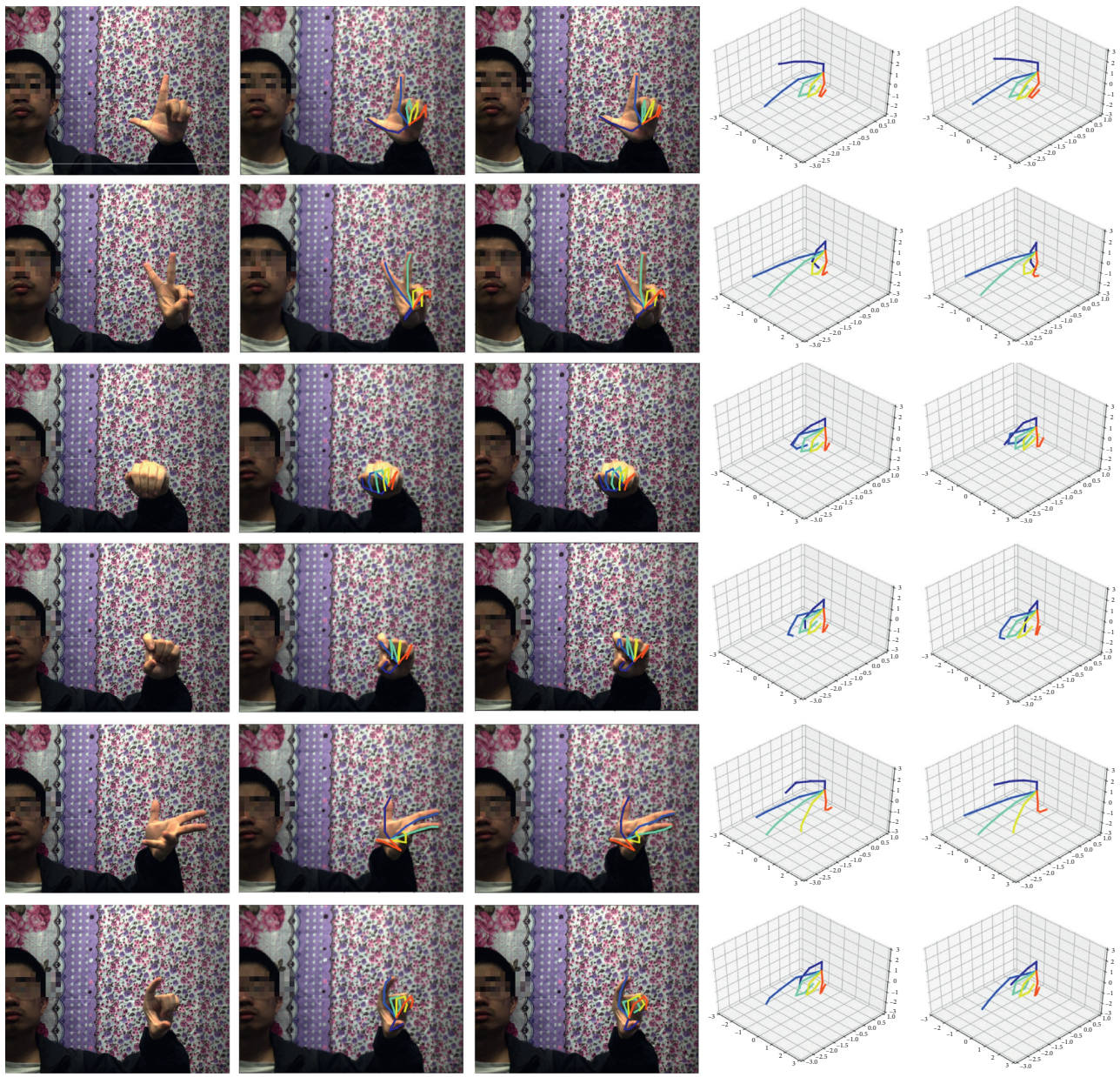

Figure 9: Full-hand pose estimation results on STB.

superiority of 4CHNet. Furthermore, the AUC of 4CHNet *Ours (without OHK)* also reaches 0.969, which is superior to most existing methods; there is no doubt that it further validates the superiority of four-stage cascaded network.

4.6. Display and Comparison of Estimated Results. In this section, we make a qualitative analysis of the proposed 4CHNet by visualizing the hand pose estimation results and comparing them with their corresponding labels. Figures 7 and 8 are the estimation results of 4CHNet on STB and RHD, respectively. And their first, second, and third rows represent full hand pose, finger pose, and palm pose estimation, respectively. The first and third columns represent the estimation results, while the second and fourth columns represent their corresponding labels. As shown in Figures 7 and 8, the full hand pose, finger pose, and palm pose estimated by 4CHNet have obtained good results, which reflects the effectiveness of the hierarchical estimation. Furthermore, we present more results of the full hand pose estimation, as shown in Figures 9 and 10, respectively, representing the qualitative results on STB and RHD. The first column represents original RGB images, and the second and fourth columns represent the full hand pose estimation of 2D and 3D, respectively; the third and fifth columns are their corresponding labels. As can be seen from Figure 9, our 2D and 3D estimated results of 4CHNet are basically consistent with the labels on the real dataset STB. Only in a few gestures with complex motions and severe occlusions, the estimation results are slightly biased, which indicates that 4CHNet can

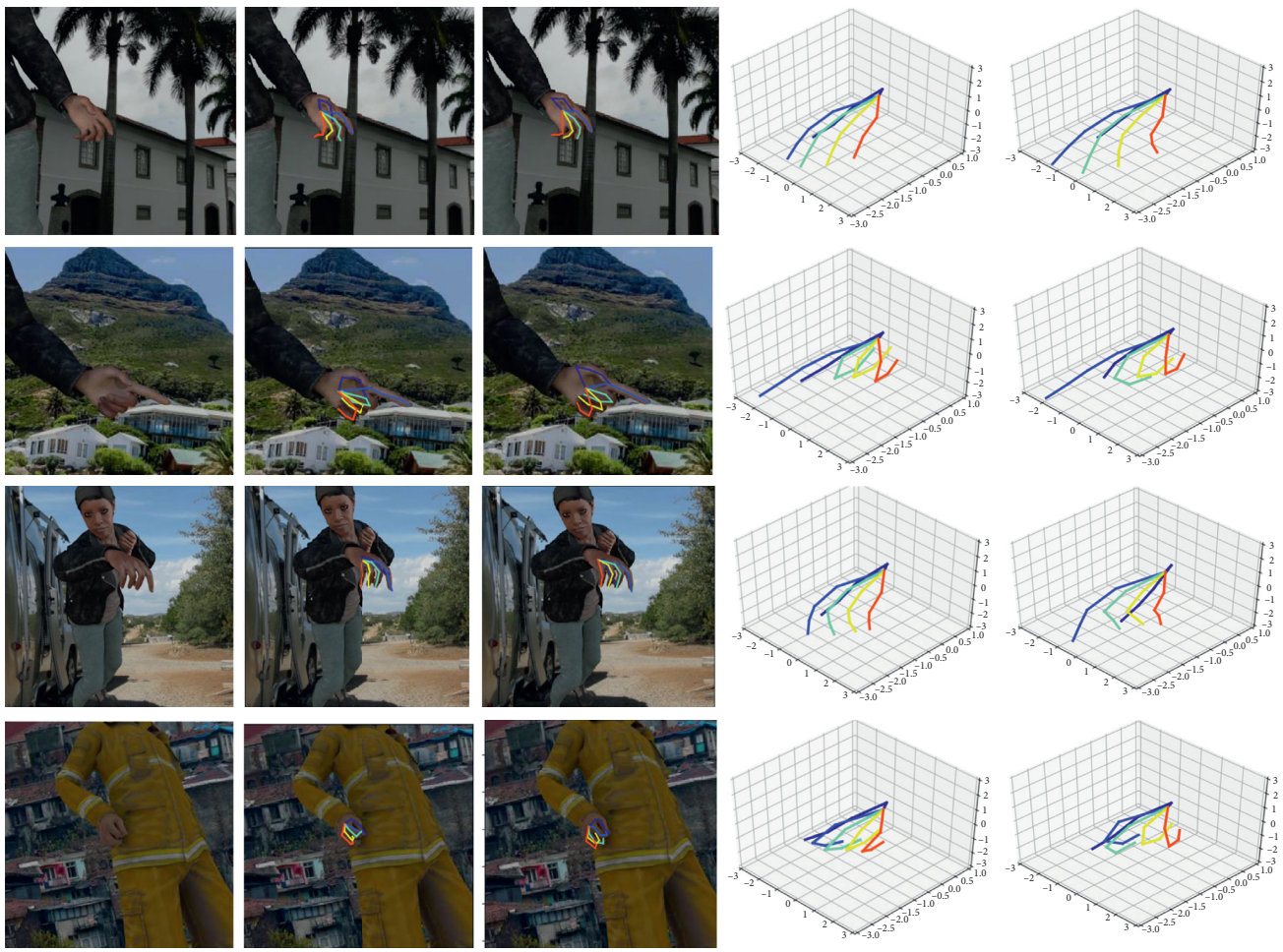

Figure 10: Full-hand pose estimation results on RHD.

be well promoted in the real world. From Figure 10, we can find that, on the synthetic dataset RHD, the estimated results are close to the labels but still have a gap. This is because synthetic dataset RHD has a lot of noise and ambiguity, and the proportion of hands is small, which results in highly difficult estimation.

5. Conclusions

Based on the cascaded CNN and hierarchical CNN, we have proposed a novel four-stage cascaded hierarchical CNN (4CHNet) for estimating 3D hand pose of a single RGB image. Four stages include mask estimation stage, 2D hand pose estimation stage, hierarchical estimation stage, and 3D hand pose estimation stage. The four stages are cascaded for end-to-end training to achieve mutually beneficial progress. At the same time, the extracted hand features are divided into the finger layer and palm layer in hierarchical estimation stage to estimate corresponding finger pose and palm pose respectively. Finally, we concatenate them to estimate full 3D hand pose. This hierarchical network leverages finger and palm constraints to extract deeper and more representative feature information to improve accuracy of 3D hand pose estimation. In this work, we have experimented on two public datasets and compared 4CHNet with the state-of-art methods on two datasets. The experimental results verify the significant promotion and conspicuous advantages of our proposed method.

Data Availability

Previously reported data were used to support this study and are available at 10.1109/TCSVT.2018.2879980 and 10.1109/iccv.2017.525 (https://arxiv.org/abs/1610.07214). These prior studies and datasets are cited at relevant places within the text as references.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant nos. 61462038, 61562039, and 61502213, in part by the Science and Technology Planning Project of Jiangxi Provincial Department of Education under Grant GJJ190217, and in part by the Open

Project Program of the State Key Lab of CAD & CG of Zhejiang University under Grant A2029.

References

[1] W. Hürst and C. Van Wezel, "Gesture-based interaction via finger tracking for mobile augmented reality," *Multimedia Tools and Applications*, vol. 62, no. 1, pp. 233–258, 2013.

[2] J. Song, G. Sörös, F. Pece et al., "In-air gestures around unmodified mobile devices," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, pp. 319–329, Honolulu, HI, USA, October 2014.

[3] Y. Jang, S.-T. Noh, H. J. Chang, T.-K. Kim, and W. Woo, "3d finger cape: clicking action and position estimation under self-occlusions in egocentric viewpoint," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 4, pp. 501–510, Apr. 2015.

[4] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: a review," *Computer Vision and. Image Understanding*, vol. 108, no. 1–2, pp. 52–73, 2007.

[5] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical Bayesian filter," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1372–1384, 2006.

[6] D. Tang, H. J. Chang, A. Tejani, and T. Kim, "Latent regression forest: structured estimation of 3d articulated hand posture," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3786–3793, Columbus, OH, USA, June 2014.

[7] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics*, vol. 33, no. 5, pp. 1–10, 2014.

[8] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *Proceddings of the IEEE Conference on Computer Vision and Pattern Recognit (CVPR)*, pp. 824–832, Boston, MA, USA, June 2015.

[9] S. Sridhar, F. Mueller, M. Zollhöfer, D. Casas, A. Oulasvirta, and C. Theobalt, "Real-time joint tracking of a hand manipulating an object from RGB-D input," in *Proceedings of the Computer Vision-ECCV 2016*, pp. 294–310, Amsterdam, The Netherlands, October 2016.

[10] C. Wan, A. Yao, and L. Van Gool, "Hand pose estimation from local surface normals," in *Proceedings of the Computer Vision-ECCV 2016*, pp. 554–569, Amsterdam, The Netherlands, October 2016.

[11] L. Ge, H. Liang, J. Yuan, and D. 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 Recognitition (CVPR)*, pp. 1991–2000, Honolulu, HI, USA, July 2017.

[12] C. Wan, T. Probst, L. V. Gool, and A. Yao, "Combining gans and vaes with a shared latent space for hand pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and. Pattern Recognitition (CVPR)*, pp. 1196–1205, Honolulu, HI, USA, July 2017.

[13] H. Liang, J. Yuan, J. Lee, L. Ge, and D. Thalmann, "Hough forest with optimized leaves for global hand pose estimation with arbitrary postures," *IEEE Transactions on Cybernetics*, vol. 49, no. 2, pp. 527–541, 2017.

[14] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[15] G. Wang, X. Yin, X. Pei, and C. Shi, "Depth estimation for speckle projection system using progressive reliable points growing matching," *Applied Optics*, vol. 52, no. 3, pp. 516–524, 2013.

[16] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel realsense stereoscopic depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and. Pattern Recognitition Workshops (CVPRW)*, pp. 1–10, Honolulu, HI, USA, July 2017.

[17] M. Oberweger and V. Lepetit, "Deepprior++: Improving fast and accurate 3d hand pose estimation," in *in Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 585–594, Venice, Italy, October 2017.

[18] X. Chen, G. Wang, C. Zhang, T.-K. Kim, and X. Ji, "Shpr-net: deep semantic hand pose regression from point clouds," *IEEE Access*, vol. 6, pp. 43425–43439, 2018.

[19] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand pointnet: 3d hand pose estimation using point sets," in *Proceedings of the IEEE Conference on Computer Vision and. Pattern Recognitition (CVPR)*, pp. 8417–8426, Salt Lake City, UT, USA, June2018.

[20] L. Ge, Z. Ren, and J. Yuan, "Point-to-point regression pointnet for 3d hand pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 475–491, Munich, Germany, September. 2018.

[21] G. Moon, J. Y. Chang, and K. M. 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 Recognitition (CVPR)*, pp. 5079–5088, Salt Lake City, UT, USA, June 2018.

[22] H. Liang, J. Yuan, and D. Thalman, "Egocentric hand pose estimation and distance recovery in a single RGB image," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, Turin, Italy, July 2015.

[23] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4903–4911, Venice, Italy, October 2017.

[24] U. Iqbal, P. Molchanov, T. B. J. Gall, and J. Kautz, "Hand pose estimation via latent 2.5 d heatmap regression," in *Proceedings of the. European. Conference on Computer Vision (ECCV)*, pp. 118–134, Munich Germany, September 2018.

[25] F. Mueller, F. Bernard, O. Sotnychenko et al., "Ganerated hands for real-time 3d hand tracking from monocular rgb," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 49–59, Salt Lake City, UT, USA, June 2018.

[26] Y. Cai, L. Ge, J. Cai, and J. Yuan, "Weakly-supervised 3d hand pose estimation from monocular rgb images," in *Proceedings of the European Conference Computer Vision. (ECCV)*, pp. 666–682, Munich, Germany, September 2018.

[27] Y. Wang, C. Peng, and Y. Liu, "Mask-pose cascaded CNN for 2D hand pose estimation from single color image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3258–3268, 2019.

[28] L. Ge, Z. Ren, Y. Li et al., "3D hand shape and pose estimation from a single RGB image," in *Proceedings of the. IIEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10833–10842, Long Beach, CA, USA, June 2019.

[29] Y. Zhang, L. Chen, Y. Liu, J. Yong, and W. Zheng, "Adaptive wasserstein hourglass for weakly supervised hand pose estimation from monocular RGB," 2019, https://arxiv.org/abs/1909.05666.

[30] A. Boukhayma, R. D. Bem, and P. H. Torr, "3d hand shape and pose from images in the wild," in *Proceedings of the IIEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10843–10852, Long Beach, CA, USA, June 2019.

[31] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, "Real-time hand tracking under occlusion from an egocentric rgb-d sensor," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1284–1293, Venice, Italy, October 2017.

[32] E. Kazakos, C. Nikou, and I. A. Kakadiaris, "On the fusion of RGB and depth information for hand pose estimation," in *Proceedings of the. IEEE International Conference on Image Processing (ICIP)*, pp. 868–872, Athens, Greece, October 2018.

[33] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 89–98, Salt Lake City, UT, USA, June 2018.

[34] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proceedings of the Advances in Neural Information. Processing. System. (NIPS)*, pp. 3844–3852, Barcelona, Spain, December 2016.

[35] W. Liu, S. Dai, W. Yang, H. Yang, and W. Qian, "Color image 3d gesture estimation based on cascade convolution neural network," *Journal of Chinese Computer Systems*, vol. 41, no. 3, pp. 558–563, 2020.

[36] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, https://arxiv.org/abs/1706.05098.

[37] H. Guo, G. Wang, X. Chen, C. Zhang, F. Qiao, and H. Yang, "Region ensemble network: improving convolutional network for hand pose estimation," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 4512–4516, Beijing, China, September 2017.

[38] M. Madadi, S. Escalera, X. Baró, and J. Gonzalez, "End-to-end global to local cnn learning for hand pose recovery in depth data," 2017, https://arxiv.org/abs/1705.09606.

[39] Y. Zhou, J. Lu, K. Du, X. Lin, Y. Sun, and X. Ma, "Hbe: hand branch ensemble network for real-time 3d hand pose estimation,," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 501–516, Munich Germany, September 2018.

[40] K. Du, X. Lin, Y. Sun, and X. Ma, "CrossInfoNet: multi-task information sharing based hand pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9896–9905, Long Beach, CA, USA, June 2019.

[41] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 4724–4732, Las Vegas, NV, USA, July 2016.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https:// arxiv.org/abs/1409.1556.

[43] J. Zhang, "3d Hand Pose Tracking and Estimation Using Stereo Matching," 2016, https://arxiv.org/abs/1610.07214.

[44] M. Abadi, "Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," 2016, https://arxiv.org/ abs/1603.04467.

[45] P. Panteleris, I. Oikonomidis, and A. Argyros, "Using a single RGB frame for real time 3D hand pose estimation in the wild," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 436–445, Lake Tahoe, NV, USA, March 2018.