

Article

3D Hand Pose Estimation Based on Five-Layer Ensemble CNN

Lili Fan ¹, Hong Rao ² and Wenji Yang ^{3,4,*}

¹ School of Information Engineering, Nanchang University, Nanchang 330031, China; 411014519016@email.ncu.edu.cn

² Center of Computer, Nanchang University, Nanchang 330031, China; raohong@ncu.edu.cn

³ School of Software, Jiangxi Agricultural University, Nanchang 330045, China

⁴ State Key Lab of CAD & CG of Zhejiang University, Hangzhou 310058, China

* Correspondence: ywenji614@jxau.edu.cn

Abstract: Estimating accurate 3D hand pose from a single RGB image is a highly challenging problem in pose estimation due to self-geometric ambiguities, self-occlusions, and the absence of depth information. To this end, a novel Five-Layer Ensemble CNN (5LENet) is proposed based on hierarchical thinking, which is designed to decompose the hand pose estimation task into five single-finger pose estimation sub-tasks. Then, the sub-task estimation results are fused to estimate full 3D hand pose. The hierarchical method is of great benefit to extract deeper and better finger feature information, which can effectively improve the estimation accuracy of 3D hand pose. In addition, we also build a hand model with the center of the palm (represented as Palm) connected to the middle finger according to the topological structure of hand, which can further boost the performance of 3D hand pose estimation. Additionally, extensive quantitative and qualitative results on two public datasets demonstrate the effectiveness of 5LENet, yielding new state-of-the-art 3D estimation accuracy, which is superior to most advanced estimation methods.

Keywords: hierarchical thinking; 3D hand pose estimation; RGB image; hand topology



Citation: Fan, L.; Rao, H.; Yang, W. 3D Hand Pose Estimation Based on Five-Layer Ensemble CNN. *Sensors* **2021**, *21*, 649. <https://doi.org/10.3390/s21020649>

Received: 28 December 2020

Accepted: 14 January 2021

Published: 19 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The gesture is among the most commonly used expressions by humans, and accurate 3D hand pose estimation has already become a key technology in the fields of Human-Computer Interaction (HCI) and Virtual Reality (VR) [1–5]. It can help humans communicate with machines in a more natural way. Though 3D hand pose estimation has achieved significant progress after years of research [6–11], it is still far from a solved problem due to the challenges of high joint flexibility, local self-similarity, and severe self-occlusion.

At present, mainstream 3D hand pose estimation methods can be classified into two categories: holistic estimation method based on the hand [12–20] and hierarchical estimation method based on hand structure [21–26]. The holistic estimation method based on the hand aims to directly use a complete hand structure for estimation, which has developed into a mainstream method in recent years [12,14]. For instance, Zimmermann et al. [12] took the whole hand as input and used CNNs to estimate 3D coordinates of 21 keypoints. Spurr et al. [14] trained encoder–decoder pairs from the generative perspective, which allows the estimation of full 3D hand pose from different input modalities. However, these methods fail to make good use of hand structure and lose a high quantity of underlying information concerning hand structure. Thus, some researchers have carried out studies based on hand structure stratification [21,22,26], leveraging hierarchical network to explore the structure of the hand and decomposing the task of 3D hand pose estimation into several sub-tasks, then representative features are extracted through the mutual promotion among tasks, which can effectively improve estimation performance. For example, Guo et al. [21] designed an REN network based on hierarchical thinking, which simply partitioned the extracted

feature map into several regions, then estimated the local pose of each region, respectively, and finally merged them into a full hand pose. Zhou et al. [22] divided hand structure into three layers according to the functional importance of different fingers to achieve the estimation of 3D hand pose. Chen et al. [26] proposed a pose guided structured region ensemble network (Pose-REN) to estimate 3D hand pose hierarchically and iteratively. In summary, these methods all exploit the underlying information of hand topology to successfully extract more representative hand features, thereby promoting more accurate hand pose estimation.

Although the above methods [21,22,26] have achieved relatively accurate 3D hand pose estimation, they are all based on depth images. However, depth images are not as universal as RGB images in reality and have low practical applicability. Therefore, it is necessary to develop a hierarchical estimation method based on hand structure for RGB images. Inspired by [22,26], a novel Five-Layer Ensemble CNN (5LENet) for 3D hand pose estimation from a single RGB image is proposed, in which 3D hand pose estimation is decomposed into five 3D single-finger pose estimations, and then the estimates of fingers are fused to estimate a more accurate 3D hand pose. Through effectively utilizing the structural characteristics of the hand to extract deeper and more representative finger feature information, this method can not only promote more accurate 2D finger pose estimation but also can further optimize 3D finger pose estimation, and finally achieve the improvement of 3D hand pose estimation accuracy. The major contributions of this paper can be summarized as follows:

- A 5LENet for 3D hand pose estimation from a single RGB image is proposed, in which hand pose estimation is decomposed into five single-finger pose estimations by using hierarchical thinking. More representative fingers are extracted for estimating a more accurate 3D finger pose, and then the features generated in the process of 3D finger pose estimation are fused to estimate a full 3D hand pose. It can not only extract more effective features but also enhance the association between fingers through the fusion of finger features.
- Five 3D finger pose constraints are newly added, which can not only promote 3D finger pose estimation but also can form soft constraints on 2D finger pose estimation to further indirectly optimize the accuracy of 3D hand pose estimation.
- According to the topology of the hand, a model of the hand with the Palm and middle finger connected is built. This is because the middle finger is located in the middle of five fingers, and its connection with the Palm is tighter. Therefore, we connect the Palm to middle finger, which can successfully solve the accuracy degradation problem caused by connecting the Palm with multiple fingers.
- We conduct experiments on the two public datasets, and results demonstrate that our approach achieves a new state-of-the-art in 3D hand pose prediction, which proves the effectiveness and advancement of 5LENet.

2. Related Works

Vision-based 3D hand pose estimation plays an important role in the field of Augmented/Virtual Reality (AR/VR) and Human–Computer Interaction (HCI). Looking back at the work of previous scholars, we can divide the methods for 3D hand estimation into two categories: (1) holistic estimation method based on the hand; (2) hierarchical estimation method based on hand structure.

The holistic estimation method based on the hand is currently mainly based on CNNs, such as Zimmermann et al. [12] who firstly applied CNN for 3D hand pose based on CNN for a single RGB image. The method is composed of three independent networks: the first network performed image segmentation and hand location, the second network performed 2D hand keypoint detection, and the third network performed 2D–3D lifting to derive 3D keypoints. However, their 3D hand pose estimation only relies on 2D keypoints, ignoring rich texture features and latent space information of the RGB image, which would affect the estimation performance to a certain extent. In addition, each network is trained

separately, which is prone to falling into local optimization. Wang et al. [13] estimated 2D keypoint features with RGB image features and introduced a channel domain attention mechanism as a weighting scheme to reweight the concatenated features, which can exploit features more efficiently to estimate 3D hand pose. In order to solve the problem of local optimization in literature [12], Liu et al. [27] presented a three-stage cascaded CNN, which realized the mutual promotion between estimation stages, so as to achieve global optimization. Some researchers proposed different solutions to the problem of difficulty in obtaining ground truth 3D keypoints. Cai et al. [15] proposed an estimation method based on weakly-supervised deep learning, in which the estimated 3D hand pose from the RGB image is converted into a depth map through introducing a depth regularizer. Then, the estimated depth map is used to perform weak supervision on the 3D hand pose regression; it can effectively overcome the difficulty of obtaining ground truth 3D annotations, but the method severely relies on the paired RGB and depth images. Based on [15], Chen et al. [19] added a depth-map reconstruction module, which employed a conditional generative adversarial network (GAN) model for generating pseudo-real depth images of color images, then the paired color and pseudo-real depth images were used as input to the hand pose estimation module. Though it eliminates the need for real depth images, it still has a certain domain gap between pseudo-real and real depth images, which leads to low estimation accuracy. Since full 3D mesh information of hand surface is beneficial for the estimation of 3D hand pose, some scholars also integrate mesh information into 3D hand pose estimation. For example, Ge et al. [20] introduced a Graph CNN to estimate 3D hand meshes, in which the extracted hand features and keypoint heatmaps were used as its input, then the estimated mesh information was used to regress 3D hand pose. Though this work achieves a high estimation precision and provides a new direction for the research of 3D hand pose, the ground truth 3D hand mesh is rather rare in the existing datasets.

The abovementioned methods have improved the estimation performance of 3D hand pose to a certain extent, but they all have their problems. Inspired by the multi-task information-sharing mechanism [28], the hierarchical estimation method based on hand structure has become popular in 3D hand pose estimation. It decomposes the task of 3D hand pose estimation into several sub-tasks through employing a hierarchical network to analyze and explore the structure of the hand, then each sub-task learns parallelly based on the shared representation, and finally information from sub-tasks is shared to improve estimation performance. For example, based on the finger functional importance, Zhou et al. [22] designed a three-layer network including the thumb layer, the index finger layer, and the remaining fingers layer, then the local pose of the corresponding layer was estimated separately, and finally the local poses were merged to estimate a full 3D hand pose. Madadi et al. [23] firstly proposed a hierarchical method dividing hand features into six layers based on hand structure, where each layer was associated with a local pose containing a specific single finger, and the remaining layer was used to regress Palm viewpoint. Then, they fused all the features of the last convolutional layers to extract higher-level feature information to estimate 3D global hand pose. Du et al. [24] leveraged the characteristics of related sub-tasks that can facilitate the learning of the main task; they employed a hierarchical model to decompose the hand pose estimation task into two related sub-tasks, palm pose estimation task and finger pose estimation task, and shared useful information between two sub-tasks to guide deeper feature extraction via a cross-connection, and finally fused the deeper features to estimate 3D hand pose. Chen et al. [26] firstly exploited an initially estimated hand pose as guided information to extract effective joint features. Secondly, joint features belonging to the same finger were integrated to form finger features, and then finger features were further integrated to regress 3D hand pose. The entire estimation process was performed iteratively. Dai et al. [25] further extended the hierarchical idea of 3D hand pose estimation from color images, which improved the performance of 3D hand pose estimation from color images.

To further improve the performance of 3D hand pose estimation based on RGB images, a novel hierarchical estimation method based on hand structure is proposed after analyzing the hand structure, which is most similar to the method proposed in [26]. Here, it is necessary for us to emphasize the difference between them. First, Ref. [26] is based on depth images, while our work is based on RGB images, which is essentially different from Chen et al. [26]. Second, it first extracts joint features and then integrates them into finger features, while our work directly extracts finger features through a hierarchical method, which reduces the depth of the network while ensuring accuracy. Again, ours works without estimated 3D hand pose to guide the feature extraction, which eliminates the need for iteration and reduces the complexity of the network. Finally, unlike the model proposed by Chen et al. [26], where the Palm is connected to five fingers, our method connects only the Palm and the middle finger. Subsequent experimental analysis shows that the Palm connected with multiple fingers simultaneously affects the estimation accuracy.

3. Methodology

3.1. Overview

Our target is to infer a 3D hand pose from a single RGB image. To achieve this goal, a novel hierarchical network called Five-Layer Ensemble CNN (5LENet) is proposed, as illustrated in Figure 1. First, a localization segmentation network is used to locate the salient hand and crop hand image as the input of our network, which will be briefly described in Section 3.2. Next, 3D hand pose is estimated, and the whole estimating process can be divided into three stages: hand mask estimation stage, 2D hand heatmap estimation stage, and 3D hand pose estimation stage. The first stage obtains a hand mask by processing the cropped image. The second stage connects the hand mask and the intermediate image features to estimate the 2D hand heatmaps. The last stage connects hand heatmaps with the intermediate image features to estimate single finger heatmaps hierarchically and then estimates their corresponding 3D finger pose, finally fusing the extracted 3D finger feature information to estimate 3D hand pose. We will introduce the first and second stages in Section 3.3. The method of estimating a single finger pose based on the hierarchical network and fusing 3D feature information of five fingers to estimate 3D hand pose will be focused on in Section 3.4.

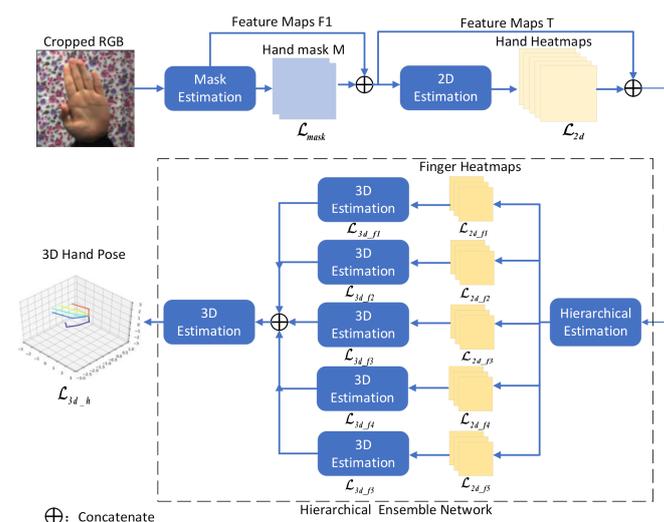


Figure 1. Schematic diagram of proposed 5LENet framework. Our 5LENet receives a cropped RGB image as the input to estimate hand mask. Then, the 2D hand heatmaps are estimated according to the features of RGB image and hand mask. Finally, 3D hand pose is estimated through the hierarchical ensemble network in the dotted box.

3.2. Localization and Segmentation Network

The goal of the localization and segmentation network is to locate the salient hand position of the RGB image and obtain the hand bounding box to crop the hand image. A lightweight Convolutional Pose Machine (CPM) [29] is used as the localization and segmentation network, which exploits a series of convolutional operations to encode the spatial structure of the RGB image, namely, a 2-channel hand mask, and calculates the center of mass of the hand mask. Then, the hand area is cropped around the center of mass and resized to 256×256 .

3.3. 2D Pose Estimation Network

Two-dimensional pose estimation is the key to 3D pose estimation. For the sake of accurately estimating 2D hand pose, we first use a lightweight VGG-19 [30] mask estimation network to encode the input cropped image. Both 128-channel image features F1 and 2-channel hand mask M are extracted by convolution. The Boolean hand mask distinguishes the hand and background areas, which is beneficial to the subsequent 2D hand pose estimation.

Based on the discussion in [31], the promotion of regressing 2D keypoint heatmaps is better than direct regressing 2D keypoint coordinates. Therefore, we employ the method of regressing 2D keypoint heatmaps to estimate 2D pose, taking 130-channel features T as the input of the 2D pose estimation network, which consists of a 2-channel hand mask M and 128-channel image features F1, to output 2D heatmaps of 21 keypoints. The 2D pose estimation network consists of five sub-networks, in which 21-channel hand heatmaps output by the previous sub-network and 130-channel image features T are connected to form 151-channel features as the input of the latter sub-network, then the location of 21 keypoints are refined iteratively. We take the 21-channel 2D heatmaps estimated by the final sub-network as the final output.

3.4. Hierarchical Ensemble Network

The hierarchical ensemble network is the main part of our proposed 5LENet. It first adopts a hierarchical estimation network to divide hand features into five layers: the thumb layer, index finger layer, middle finger layer, ring finger layer, and pinky finger layer, respectively, and each finger heatmap is estimated separately, of which only the middle finger layer adds the Palm joint. Based on the 2D finger heatmap features, deeper-level 3D finger features are extracted to estimate 3D finger pose, and finally, five layers of 3D finger features are assembled to estimate 3D hand pose. This hierarchical ensemble method is conducive to extracting deeper-level and more effective hand feature information, thereby improving the performance of 3D hand pose estimation.

3.4.1. Hierarchical Estimation Network

The target of our hierarchical estimation network is to estimate 2D heatmaps of five fingers separately, and the division of hand keypoints is shown in Figure 2a. Among the five fingers, only the middle finger is connected to the Palm, which is distinguished by different colors (blue and red) in Figure 2a. Figure 2b shows the examples of the keypoint partition of each finger from the samples of the real dataset Stereo Hand Pose Tracking Benchmark (STB).

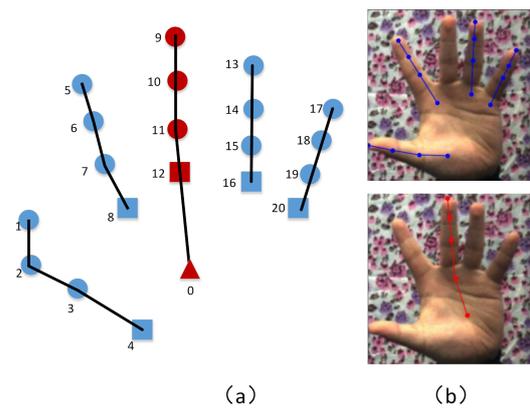


Figure 2. The diagram and examples of hand keypoint division. (a) Skeleton graph of 21 hand keypoints, in which triangle, square, and circle represent Palm, the metacarpophalangeal joint, and the phalangeal joint, respectively. The middle finger containing the Palm is marked in red, and others are marked in blue. (b) An example diagram of different finger keypoints from the real dataset Stereo Hand Pose Tracking Benchmark (STB).

The hierarchical estimation network decomposes hand features into five layers according to five different fingers to estimate the 2D heatmaps of each corresponding finger. As illustrated in Figure 3, there are five layers in the whole network, and each layer is composed of three sub-networks. Taking the first layer (thumb layer) as an example, the 4-channel thumb heatmaps output by the previous sub-network are connected with 151-channel features F to form 155-channel features as the input of the next sub-network, which optimizes the location of thumb keypoints iteratively. Then, we take 4-channel 2D thumb heatmaps F_{f13} estimated by the final sub-network as the output for the thumb layer. The 2D heatmap estimation process of other fingers is the same as that of the thumb layer, the only difference being that the number of output channels of middle finger is 5, while others are 4.

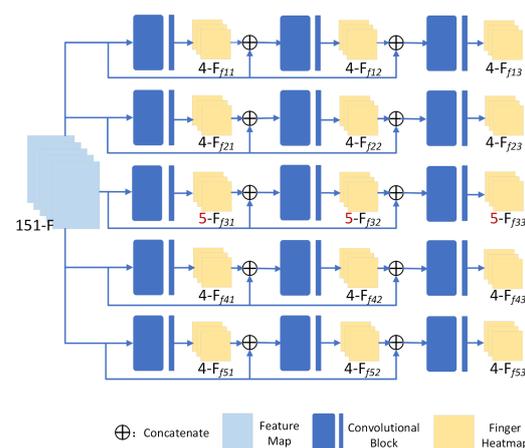


Figure 3. The architecture of the hierarchical estimation network. Five 2D single finger heatmaps are estimated by the network, of which n - F_{fij} represents the n -channel 2D heatmaps of the i th finger estimated by j th layer.

3.4.2. 3D Pose Estimation Network

Based on the five 2D finger heatmaps estimated by the hierarchical network, we further estimate the 3D pose of each finger, and then 3D hand pose is regressed from the concatenated 3D finger features of the last FC layer of all 3D finger pose estimation. For regression, we first feed the integrated 3D features into two 512-dimensional FC layers with a dropout rate of 0.2 and then regress the 3D coordinates of 21 keypoints.

This paper follows the representation in [12], using relative normalized coordinates. $W^{rel} = \{w_j^{rel}\}_{j=1}^J \in \Lambda_{3D}$ represents 3D keypoint coordinates, where Λ_{3D} and w_j^{rel} indicate the $J \times 3$ -dimensional hand joints space and the relative normalized coordinates of the j th keypoint, respectively. The 3D canonical coordinates W^c and 3D rotation matrix R are estimated parallelly within the canonical frame, and then W^{rel} is obtained indirectly, of which R is equivalent to predicting the viewpoint of a given sample concerning the canonical frame.

$$W^{rel} = W^c \cdot R^T \quad (1)$$

3.5. Loss Functions

3.5.1. Estimation Loss of 2D Pose

Before estimating 2D hand pose, the cropped image is fed to the mask estimation network to estimate a 2-channel hand mask M . The standard softmax cross-entropy loss is adopted to calculate the hand mask loss \mathcal{L}_{mask} , and it is defined as follows:

$$\mathcal{L}_{mask} = - \sum \log \left(\frac{e^{s_y}}{\sum_k e^{s_k}} \right) \quad (2)$$

where y is the corresponding label of hand mask, and s_k is the output score of the k th label, where the mask is a binary map, $k \in \{0, 1\}$.

During the process of full 2D hand pose estimation, the L_2 loss is imposed on the 2D heatmaps of 21 keypoints to calculate the estimation loss of 2D hand pose \mathcal{L}_{2d} , and the loss function is defined as follows:

$$\mathcal{L}_{2d} = \sum_{j=1}^{J=21} \|h_j^{pre} - h_j^{gt}\|_2 \quad (3)$$

where h_j^{pre} and h_j^{gt} represent the estimated and ground truth heatmaps of the j th keypoint, respectively. The ground truth heatmaps for the j th keypoint are generated by applying 2D Gaussian centered at its ground truth 2D keypoint location.

3.5.2. Loss of Hierarchical Estimation

The loss of hierarchical estimation \mathcal{L}_L is the sum of the loss of five 2D finger heatmaps; fi ($i = 1, 2, 3, 4, 5$) represents the corresponding finger, where $f1$ represents the first finger (thumb), and the others are the index finger, middle finger, ring finger, and pinky. The loss of hierarchical estimation \mathcal{L}_L is summarized as follows:

$$\mathcal{L}_L = \sum_{i=1}^5 \|h_{fi}^{pre} - h_{fi}^{gt}\|_2 \quad (4)$$

where h_{fi}^{pre} and h_{fi}^{gt} denote the estimated and ground truth heatmaps of the i th finger.

3.5.3. Estimation Loss of 3D Pose

The estimation loss of 3D pose \mathcal{L}_{3d} is composed of six parts, which are the 3D pose estimation loss of five fingers \mathcal{L}_{3d_fi} ($i = 1, 2, 3, 4, 5$) and the full 3D hand pose loss \mathcal{L}_{3d_h} , where fi represents the thumb, index finger, middle finger, ring finger, and pinky, respectively, of which only the middle finger contains the Palm joint. \mathcal{L}_{3d_fi} is used as supervision to guide the network to extract specific 3D features of each finger and promote a more accurate 3D estimation of the whole hand, and \mathcal{L}_{3d_h} is used as supervision to guide the predictions of the whole hand. We employ L_2 loss to calculate the loss of the corresponding part of 3D canonical coordinates W^c and 3D rotation matrix R . They are defined as follows:

$$\mathcal{L}_{3d_fi} = \|W_{fi}^{pre-c} - W_{fi}^{gt-c}\|_2 + \|R_{fi}^{pre} - R_{fi}^{gt}\|_2, \quad i = (1, 2, 3, 4, 5) \quad (5)$$

$$\mathcal{L}_{3d_fi} = \|W_{fi}^{pre_c} - W_{fi}^{gt_c}\|_2 + \|R_{fi}^{pre} - R_{fi}^{gt}\|_2, \quad i = (1, 2, 3, 4, 5) \quad (6)$$

$$\mathcal{L}_{3d} = \sum_{i=1}^5 \mathcal{L}_{3d_fi} + \mathcal{L}_{3d_h} \quad (7)$$

where $W_{fi}^{pre_c}$ and $W_h^{pre_c}$ denote the estimated 3D canonical coordinates of the i th finger and the full hand, and $W_{fi}^{gt_c}$ and $W_h^{gt_c}$ are their corresponding 3D labels. R_{fi}^{pre} and R_h^{pre} denote the estimated 3D rotation matrix of the i th finger and the full hand, and R_{fi}^{gt} and R_h^{gt} are their ground truth.

3.5.4. Total Loss of Network

Because the loss value of hand mask is too large, we add a weight ratio λ to this item to reduce its loss value. Experimental results show that the model achieves the best performance when $\lambda = 0.05$, and the total loss of our entire network \mathcal{L}_{total} is defined as follows:

$$\mathcal{L}_{total} = \lambda * \mathcal{L}_{mask} + \mathcal{L}_{2d} + \mathcal{L}_L + \mathcal{L}_{3d} \quad (8)$$

4. Experiments

4.1. Datasets and Metrics

4.1.1. OneHand10K

OneHand10K [32] is a dataset based on single-handed RGB images with different resolutions, which is composed of 10,000 images for training and 1703 images for testing. All samples are annotated with 2D keypoint locations and hand masks. The early work of Liu et al. [27] proved the superiority of using OneHand10K to train the localization segmentation network, so we follow their method to exploit ground truth hand masks provided by OneHand10K to train our localization segmentation network and promote the model's adaptability to the real environment.

Because the image resolution of this dataset is not uniform, it is necessary to adjust and fill the original image $I \in \mathbb{R}^{w \times h \times 3}$, where w and h denote the width and height of the image, respectively. We first adjust the resolution of all color images and corresponding masks with a ratio \square , then fill the $m \times m$ image with the resized color image from the left-top corner, where $m = 320$, and fill the remaining empty region with the gray value (128,128,128). Similar procedures are performed for the hand mask, but filled with zeros.

$$\square = \min\left(\frac{m}{w}, \frac{m}{h}\right) \quad (9)$$

4.1.2. RHD

Rendered Hand Pose Dataset (RHD) [12] is a fully synthetic dataset with rendered hand images, which is composed of 41,258 training images and 2728 testing images with the resolution of 320×320 . All samples come with full annotation of 2D and 3D labels of 21 keypoints and additionally corresponding mask labels and depth images. The dataset is highly challenging due to the diverse visual scenery, illumination, and noise. We use all labels except the depth images to train the entire network.

4.1.3. STB

Stereo Hand Pose Tracking Benchmark (STB) [33] is a real dataset, which contains two different subsets of image resolution of 640×480 : STB-BB and STB-SK. The images in STB-BB and STB-SK are captured by the Point Grey Bumblebee2 stereo camera and the Intel F200 depth camera, respectively. Both subsets provide 2D and 3D labels of 21 keypoints. In our experiments, we only use the STB-BB subset for 5LENet training and testing. STB-BB has 12 sequences totaling 36,000 images, which includes six different backgrounds. Following the training and evaluation scheme in [12], 10 of these sequences are used as the training set, and the remaining are the testing set. Due to the domain gap

between the synthetic and real images, the model trained by RHD cannot adapt well to the real environment. Thus, we use 2D and 3D labels of the STB-BB subset to fine-tune the model pre-trained by RHD to promote greater adaptability of the model.

4.1.4. Evaluation Metrics

As proposed in [12], we quantitatively evaluate the 3D hand pose estimation performance with two metrics. The first metric is end-point-error (EPE), the Euclidean distance between the predicted 3D keypoint coordinates and the ground truth, and the second metric is the area under the curve (AUC) on the percentage of correct keypoints (PCK) under different error thresholds σ in the 3D space. They are defined as follows:

$$PCK_{\sigma} = \frac{1}{J} \sum_{j=1}^J \delta(\|w_j^{pre} - w_j^{gt}\| < \sigma) \quad (10)$$

$$AUC_J = \int PCK_{\sigma}^J \quad (11)$$

where w_j^{pre} represents the estimated 3D coordinates of the j th keypoint, w_j^{gt} is its corresponding 3D labels, and δ is an indicate function.

4.2. Experimental Details

Our method is implemented with TensorFlow [34], and an Adam optimizer [35] is used to train the entire network. All experiments are conducted on a single GPU server with CUDA 8.0 NVIDIA RTX2080Ti.

The input of our network is a cropped RGB image preprocessed by localization and segmentation network, which is trained by the OneHand10K in Section 4.1.1. Its training batch is 8, and the initial learning rate is set to 5×10^{-5} ; the learning rate decreases at a rate of 0.1 every 10 K iterations, with a total of 40 K iterations.

The entire 5LENet is trained in an end-to-end manner. We first use RHD to pre-train the network; the training batch is 8, and the initial learning rate is 5×10^{-5} . The learning rate decreases by 0.3 for every 50 K iterations, with a total of 300 K iterations. Then, the STB is used to fine-tune the pre-training model to improve its adaptability to the real environment. Except for the different learning rate settings, the remaining training parameters are the same as the pre-training process. The learning rate is 1.5×10^{-5} for the first 50 K iterations and then decays every 50 K iterations with the decay ratio 0.3; after 250 K iterations, the network is trained at 3.645×10^{-8} until the end.

4.3. Ablation Study

In order to verify the effectiveness of different components of our 5LENet, we conduct extensive ablation studies on the number of network layers, the newly added 3D finger pose constraints, and Palm-to-finger connection problems. For clarity of discussion, we mark the thumb, index finger, middle finger, ring finger, and pinky as T, I, M, R, and P, respectively, and the center of the palm as Palm. Note that the setting of all ablation experimental parameters in this paper is consistent, and the model is pre-trained by RHD and then fine-tuned by STB using the same OneHand10K-trained localization segmentation network to preprocess the input image.

4.3.1. Effectiveness of Five-Layer Network

To verify the superiority of our five-layer network, in this experiment, we uniformly use a hand model with the Palm connected to the middle finger and add 3D finger constraints. Only the hand hierarchical structure is changed, which is divided into five layers, four layers, and three layers, represented by Ours, 4L-M+Palm, and 3L-M+Palm, respectively. The hand division structure of the four-layer and three-layer layers is T, I, M, R/P, and T/I, M, R/P. The experimental results of using the evaluation metrics in Section 4.1.4

to evaluate Ours, 4L-M+Palm, and 3L-M+Palm on the STB dataset are shown in Figure 4 and Table 1.

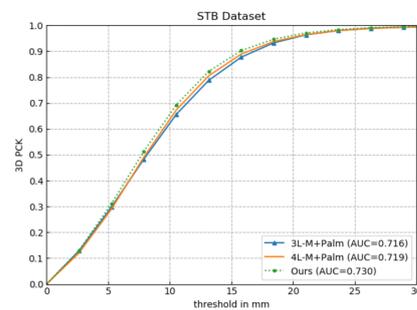


Figure 4. Comparative experiment of the effectiveness of the five-layer network.

Table 1. Error analysis of five-layer network effectiveness.

Network	EPE Median (mm)	EPE Mean (mm)	AUC
3L-M+Palm	8.408	8.957	0.716
4L-M+Palm	8.253	8.842	0.719
Ours	7.937	8.492	0.730

As illustrated in Figure 4, Ours achieves an AUC of 0.730 within the error threshold 0–30 mm, which improves by 1.96% and 1.53%, respectively, compared with 0.716 of 3L-M+Palm and 0.719 of 4L-M+Palm. From Table 1, we can observe that as the number of network layers increases, the estimation error gradually decreases, which successfully verifies the effectiveness of the five-layer network to extract deeper and more representative features, thereby further improving the accuracy of 3D hand pose estimation.

4.3.2. Effectiveness of Newly Added 3D Finger Pose Constraints

In this section, the effectiveness of another component of our proposed method will be demonstrated: the newly added 3D finger pose constraints. The basic network architecture is kept the same, among which Ours adds five 3D finger pose constraints, while 2d-Ensemble directly fuses the 2D finger feature information to estimate full 3D hand pose. Ours and 2d-Ensemble are evaluated on STB with the metrics in Section 4.1.4; the experimental results are shown in Figure 5 and Table 2.

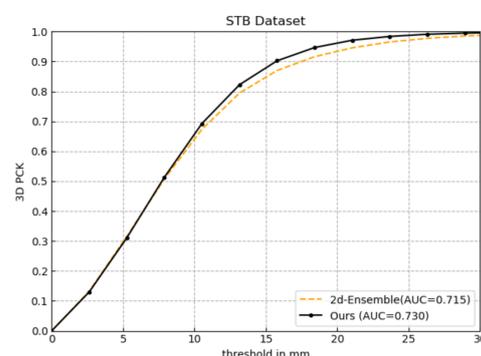


Figure 5. Comparative experiment of the effectiveness of newly added five 3D finger pose constraints.

Table 2. Error analysis of newly added five 3D finger pose constraints effectiveness.

Network	EPE Median (mm)	EPE Mean (mm)	AUC
2d-Ensemble	8.047	9.003	0.715
Ours	7.937	8.492	0.730

As shown in Figure 5, compared with 2d-Ensemble 0.715, the AUC of ours has improved 0.015 within the error threshold 0–30mm. Table 2 displays the 3D hand pose estimation errors of 2d-Ensemble and Ours. Compared with 2d-Ensemble, the average endpoint error and median endpoint error of Ours are reduced by 0.511 (mm) and 0.110 (mm), respectively, which demonstrates that the newly added constraints can effectively improve the accuracy of 3D hand pose estimation.

4.3.3. Effectiveness of Connecting Palm with a Single Finger

In order to validate the effectiveness of the Palm connected with a single finger for 3D hand pose estimation, we keep the five-layer network and newly added five 3D finger pose constraints the same and change the number of fingers connected to the Palm, of which Ours represents the Palm connected to the middle finger, and TIMRP+Palm and IMR+Palm represent the Palm connected with all fingers, and connected with the index finger, middle finger, and ring finger, respectively. The metrics mentioned in Section 4.1.4 are used to evaluate them on STB. Figure 6 and Table 3 show their experimental results.

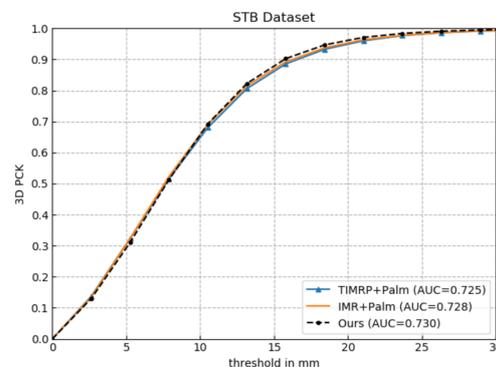


Figure 6. Comparative experiment of the effectiveness of the Palm connecting with a finger.

Table 3. Error analysis of connecting Palm with a single finger.

Network	EPE Median (mm)	EPE Mean (mm)	AUC
TIMRP+Plam	8.032	8.672	0.725
IMR+Palm	7.949	8.575	0.728
Ours	7.937	8.492	0.730

As shown in Figure 6, Ours achieves the best performance within the given error threshold, and its AUC value reaches 0.730. From Table 3, we can find that as the number of fingers connected to the Palm increases, the estimation accuracy of the network decreases instead. The estimated result shows that the structure with the Palm connected with multiple fingers cannot promote estimation accuracy; on the contrary, it will affect the accuracy. Thus, this experiment demonstrates that the Palm connected with a single finger helps to boost the performance of estimation.

4.3.4. Effectiveness of Connecting Palm with Middle Finger

Since the validity of the Palm connected with a single finger has been verified in Section 4.3.3, this section will further demonstrate the effectiveness of the Palm connected with the middle finger. The five-layer network and the newly added five 3D finger pose constraint are unchanged, and the Palm is changed to connect with the thumb, index finger, middle finger, and pinky, respectively, which are represented by T+Palm, I+Palm, Ours (M+Palm), R+Palm, and P+Palm. Similarly, using the same metrics in Section 4.1.4 to evaluate them on the STB dataset, the experimental results are shown in Figure 7 and Table 4.

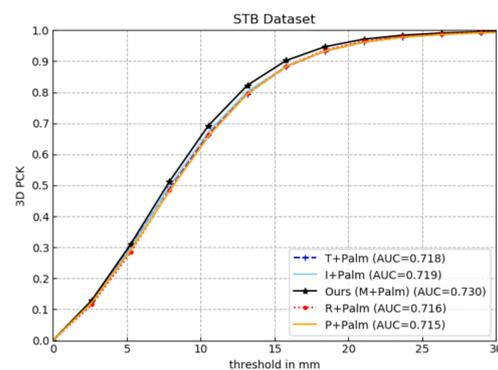


Figure 7. Comparative experiment of the effectiveness of the Palm connecting with middle finger.

Table 4. Error analysis of connecting Palm with middle finger.

Network	EPE Median (mm)	EPE Mean (mm)	AUC
t+palm	8.258	8.858	0.718
I+Palm	8.195	8.852	0.719
R+Palm	8.351	8.940	0.716
P+Palm	8.375	8.989	0.715
Ours (M+Palm)	7.937	8.492	0.730

It can be seen from Figure 7 that the AUC value of Ours (M+Palm) within the error threshold is higher than that of other networks, while the estimation performance of others is considerably close. In Table 4, Ours (M+Palm) achieves the minimum error both in the average endpoint error and the median endpoint error. The reason why Ours (M+Palm) achieves the highest estimation accuracy is that the Palm and the middle finger are more closely related with each other than other fingers, according to the topological features of the hand.

4.4. Comparison with the State-of-the-Art Methods

As shown in Figures 8 and 9, under the same evaluation conditions and criteria, we perform a quantitative comparison of our proposed 5LENet with other previous competitive methods on RHD and STB datasets. Since the 3D hand pose estimation method based on depth image is inconsistent with the evaluation metric based on color image method, here, we do not make a quantitative comparison with the most similar work [26] to ours.

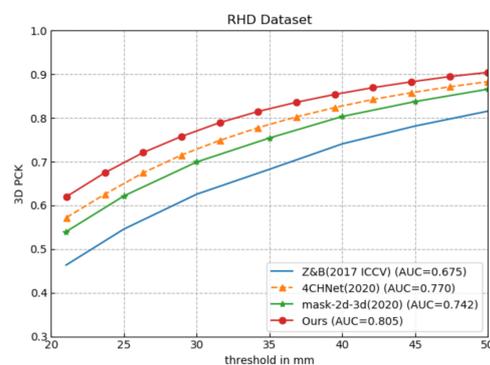


Figure 8. Comparison with the state-of-the-art methods on Rendered Hand Pose (RHD) dataset [12].

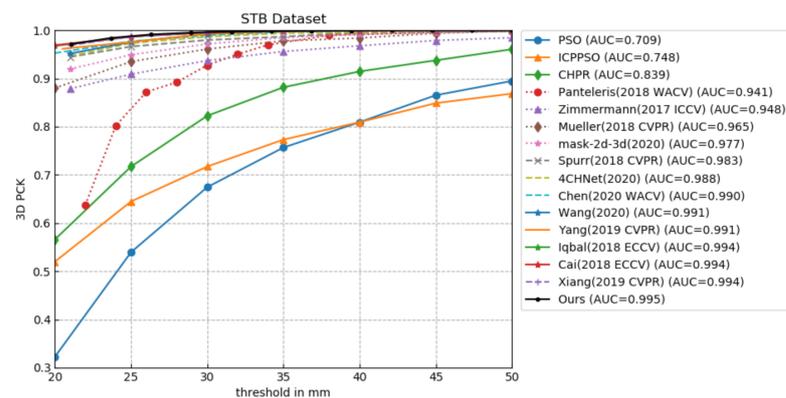


Figure 9. Comparison with the state-of-the-art methods on STB dataset [33].

As illustrated in Figure 8, our 5LENet outperforms state-of-the-art methods [12,25,27] by a large margin on the RHD dataset. The AUC value of Ours reaches 0.805 within the error threshold 20–50 mm, which is improved by 19.26%, 8.49%, and 4.55%, respectively, compared to Zimmermann et al. [12], mask-2d-3d [27], and 4CHNet [25]. All experimental results of the competitive methods are derived from the corresponding literature.

Figure 9 shows the comparisons with state-of-the-art methods [8,12–15,19,25,36–42] on the STB dataset; Ours represents the proposed 5LENet. It is worth noting that in addition to the pose estimation error, the process of hand positioning will also produce the error, so the methods involved in the comparison also need to add the corresponding localization error if they do not have it. As presented in Figure 9, our method achieves the highest AUC value 0.995 within the error threshold of 20–50mm; compared with 0.948 of Zimmermann et al. [12] and 0.988 of 4CHNet [25], the performance improves by 0.047 and 0.007, respectively, where 4CHNet [25] employs a hierarchical network to divide hand features into two layers of palm and finger features. The high precision estimation result of ours again verifies the effectiveness and advancement of the five-layer network and five newly added 3D finger constraints. There is a substantial improvement compared with the 0.991 of Wang et al. [13] and Yang et al. [39], while Wang et al. [13] use an RHD-trained localization segmentation network to crop the hand area, which cannot adapt well to the real environment compared with the OneHand10K-trained network. Yang et al. [39] achieve image synthesis and pose estimation by learning disentangled representations of hand poses and hand images, but to a certain extent, the disentangling process will lead to the missing information that helps to generate useful data. Additionally, although the estimation results of Iqbal [42], Xiang [41], and Cai et al. [15] are significantly close to ours, it is necessary to emphasize the difference between theirs and ours here. Among them, in [41,42], the hand location error is not considered, and in [15], the paired depth and RGB image are required, which are rather difficult to collect, so our network is superior to theirs to a certain extent. The results from competing methods are derived from the corresponding literature.

4.5. Qualitative Results

Some qualitative results of 5LENet on STB and RHD can be seen in Figures 10 and 11. For each dataset, the first row represents the cropped input RGB image, the second row shows the estimated 3D hand pose results of the proposed 5LENet, and the third row is the ground truth. It is observed that the estimation of 3D hand poses is visually very similar to the ground truth ones, and for some self-occluded hand poses, as shown in Figure 10, our 5LENet also has comparable estimated performance. Due to the complexity and challenges of the synthetic dataset RHD, as shown in Figure 11, though the 3D hand poses predicted by 5LENet are similar to the ground truth, there are still some gaps.

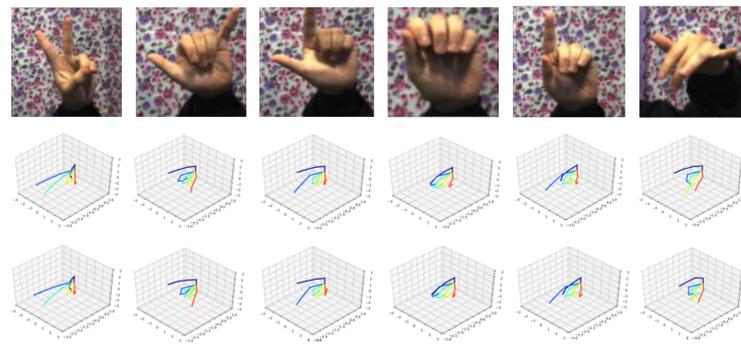


Figure 10. Qualitative results for STB dataset [33].

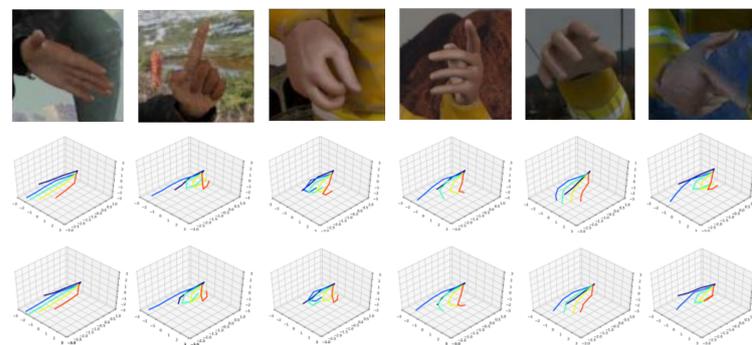


Figure 11. Qualitative results for RHD dataset [12].

5. Conclusions

In this paper, a novel five-layer ensemble CNN (5LENet) is proposed for accurate 3D hand pose estimation from a single RGB image. Based on hierarchical thinking, we designed a hierarchical network to decompose the full hand pose estimation into five single-finger pose tasks. It can not only promote a more accurate 2D finger pose estimation but also can further optimize the estimation of 3D finger hand pose. Then, the 3D finger features optimized in the estimation process are fused to estimate the full 3D hand pose. At the same time, 5LENet leverages five newly added 3D finger pose constraints to extract more representative features to improve estimation accuracy. Furthermore, we also build a hand model with the Palm connected to the middle finger, which can further help to improve the estimation performance. We make quantitative comparisons with other state-of-the-art methods on RHD and STB datasets. The high-precision 3D estimation results demonstrate the advancement and superiority of our proposed network.

Author Contributions: Conceptualization, L.F. and H.R.; funding acquisition, W.Y.; methodology, L.F. and W.Y.; resources, W.Y.; software, L.F.; supervision, H.R. and W.Y.; validation, L.F.; writing—original draft, L.F.; writing—review and editing, H.R. and W.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under grants 61462038, 61562039, 61502213, and 62062041; 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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, X.Y.; Tsai, M.S.; Huang, C.C. 3D Virtual-Reality Interaction System. In Proceedings of the 2019 IEEE International Conference on Consumer Electronics-Taiwan, ICCE-TW 2019, Ilan, Taiwan, 20–22 May 2019; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2019.
2. Hsieh, M.C.; Lee, J.J. Preliminary Study of VR and AR Applications in Medical and Healthcare Education. *J. Nurs. Heal. Stud.* **2018**, *3*. [[CrossRef](#)]
3. Mahesh, M.M.; Jerry, M.S.; Wadala, D.S.B. FPGA based hand sign recognition system. In Proceedings of the RTEICT 2017–2nd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, Bangalore, India, 19–20 May 2017; pp. 294–299. [[CrossRef](#)]
4. Antoshchuk, S.; Kovalenko, M.; Sieck, J. Gesture recognition-based human-computer interaction interface for multimedia applications. In *Digitisation of Culture: Namibian and International Perspectives*; Springer: Singapore, 2018; pp. 269–286. ISBN 9789811076978.
5. Yang, J.; Wang, C.; Jiang, B.; Song, H.; Meng, Q. Visual Perception Enabled Industry Intelligence: State of the Art, Challenges and Prospects. *IEEE Trans. Ind. Inf.* **2020**, *17*, 2204–2219. [[CrossRef](#)]
6. Stenger, B.; Thayananthan, A.; Torr, P.H.S.; Cipolla, R. Model-based hand tracking using a hierarchical bayesian filter. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*. [[CrossRef](#)] [[PubMed](#)]
7. Tompson, J.; Stein, M.; Lecun, Y.; Perlin, K. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Trans. Graph.* **2014**, *33*. [[CrossRef](#)]
8. Sun, X.; Wei, Y.; Liang, S.; Tang, X.; Sun, J. Cascaded hand pose regression. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015. [[CrossRef](#)]
9. Ge, L.; Liang, H.; Yuan, J.; Thalmann, D. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 5679–5688. [[CrossRef](#)]
10. Wan, C.; Probst, T.; Van Gool, L.; Yao, A. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 1196–1205. [[CrossRef](#)]
11. Cai, Y.; Ge, L.; Cai, J.; Magnenat-Thalmann, N.; Yuan, J. 3D Hand Pose Estimation Using Synthetic Data and Weakly Labeled RGB Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [[CrossRef](#)] [[PubMed](#)]
12. Zimmermann, C.; Brox, T. Learning to Estimate 3D Hand Pose from Single RGB Images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2017; pp. 4913–4921. [[CrossRef](#)]
13. Wang, X.; Jiang, J.; Guo, Y.; Kang, L.; Wei, Y.; Li, D. CFAM: Estimating 3D hand poses from a single RGB image with attention. *Appl. Sci.* **2020**, *10*, 618. [[CrossRef](#)]
14. Spurr, A.; Song, J.; Park, S.; Hilliges, O. Cross-Modal Deep Variational Hand Pose Estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [[CrossRef](#)]
15. Cai, Y.; Ge, L.; Cai, J.; Yuan, J. Weakly-supervised 3D hand pose estimation from monocular RGB images. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Suzhou, China, 16–19 September 2018. [[CrossRef](#)]
16. Baek, S.; Kim, K.I.; Kim, T.K. Pushing the envelope for rgb-based dense 3D hand pose estimation via neural rendering. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 1067–1076. [[CrossRef](#)]
17. Guo, S.; Rigall, E.; Qi, L.; Dong, X.; Li, H.; Dong, J. Graph-based CNNs with self-supervised module for 3D hand pose estimation from Monocular RGB. *IEEE Trans. Circuits Syst. Video Technol.* **2020**. [[CrossRef](#)]
18. Ge, L.; Liang, H.; Yuan, J.; Thalmann, D. Robust 3D Hand Pose Estimation from Single Depth Images Using Multi-View CNNs. *Proc. IEEE Trans. Image Process.* **2018**, *27*. [[CrossRef](#)] [[PubMed](#)]
19. Chen, L.; Lin, S.Y.; Xie, Y.; Lin, Y.Y.; Fan, W.; Xie, X. DGGAN: Depth-image guided generative adversarial networks for disentangling RGB and depth images in 3D hand pose estimation. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, 1–5 March 2020. [[CrossRef](#)]
20. Ge, L.; Ren, Z.; Li, Y.; Xue, Z.; Wang, Y.; Cai, J.; Yuan, J. 3D hand shape and pose estimation from a single RGB image. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019. [[CrossRef](#)]
21. Guo, H.; Wang, G.; Chen, X.; Zhang, C.; Qiao, F.; Yang, H. Region ensemble network: Improving convolutional network for hand pose estimation. In Proceedings of the International Conference on Image Processing, ICIP, Beijing, China, 17–20 September 2017; pp. 4512–4516. [[CrossRef](#)]
22. Zhou, Y.; Lu, J.; Du, K.; Lin, X.; Sun, Y.; Ma, X. HBE: Hand branch ensemble network for real-time 3d hand pose estimation. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Hong Kong, China, 8–10 December 2011; Springer: Berlin/Heidelberg, Germany, 2018; pp. 521–536. [[CrossRef](#)]

23. Madadi, M.; Escalera, S.; Baro, X.; Gonzalez, J. End-to-end Global to Local CNN Learning for Hand Pose Recovery in Depth Data. *arXiv* **2017**, arXiv:1705.09606.
24. Du, K.; Lin, X.; Sun, Y.; Ma, X. Crossinfonet: Multi-task information sharing based hand pose estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9888–9897. [[CrossRef](#)]
25. Dai, S.; Liu, W.; Yang, W.; Fan, L.; Zhang, J. Cascaded Hierarchical CNN for RGB-Based 3D Hand Pose Estimation. *Math. Probl. Eng.* **2020**, *2020*, 1–13. [[CrossRef](#)]
26. Chen, X.; Wang, G.; Guo, H.; Zhang, C. Pose guided structured region ensemble network for cascaded hand pose estimation. *Neurocomputing* **2019**, *395*, 138–149. [[CrossRef](#)]
27. Liu, W.; Dai, S.; Yang, W.; Yang, H.; Qian, W. Color image 3d gesture estimation based on cascade convolution neural network. *J. Chin. Comput. Syst.* **2020**, *41*, 558–563.
28. Zhang, Y.; Yang, Q. An overview of multi-task learning. *Natl. Sci. Rev.* **2018**, *5*, 30–43. [[CrossRef](#)]
29. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; IEEE Computer Society, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4724–4732. [[CrossRef](#)]
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.
31. Pfister, T.; Charles, J.; Zisserman, A. Flowing ConvNets for Human Pose Estimation in Videos. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: Piscataway Township, NJ, USA, 2015; pp. 1913–1921. [[CrossRef](#)]
32. Wang, Y.; Peng, C.; Liu, Y. Mask-Pose Cascaded CNN for 2D Hand Pose Estimation from Single Color Image. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 3258–3268. [[CrossRef](#)]
33. Zhang, J.; Jiao, J.; Chen, M.; Qu, L.; Xu, X.; Yang, Q. 3D Hand Pose Tracking and Estimation Using Stereo Matching. *arXiv* **2016**, arXiv:1610.07214.
34. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, 2–4 November 2016.
35. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings; International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May 2015.
36. Qian, C.; Sun, X.; Wei, Y.; Tang, X.; Sun, J. Realtime and robust hand tracking from depth. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014. [[CrossRef](#)]
37. Panteleris, P.; Oikonomidis, I.; Argyros, A. Using a Single RGB Frame for Real Time 3D Hand Pose Estimation in the Wild. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 436–445. [[CrossRef](#)]
38. Oikonomidis, I.; Kyriazis, N.; Argyros, A. Efficient Model-Based 3D Tracking of Hand Articulations Using Kinect. In Proceedings of the British Machine Vision Conference, Dundee, UK, 29 August–2 September 2011; pp. 101.1–101.11. [[CrossRef](#)]
39. Yang, L.; Yao, A. Disentangling latent hands for image synthesis and pose estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9869–9878. [[CrossRef](#)]
40. Mueller, F.; Bernard, F.; Sotnychenko, O.; Mehta, D.; Sridhar, S.; Casas, D.; Theobalt, C. GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [[CrossRef](#)]
41. Xiang, D.; Joo, H.; Sheikh, Y. Monocular total capture: Posing face, body, and hands in the wild. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10957–10966. [[CrossRef](#)]
42. Iqbal, U.; Molchanov, P.; Breuel, T.; Gall, J.; Kautz, J. Hand Pose Estimation via Latent 2.5D Heatmap Regression. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Hong Kong, China, 14–16 March 2018. [[CrossRef](#)]