Bi-stream Pose Guided Region Ensemble Network for Fingertip Localization from Stereo Images

Guijin Wang, *Senior Member, IEEE,* Cairong Zhang, Xinghao Chen,
Xiangyang Ji, *Member, IEEE,* Jing-Hao Xue, and Hang Wang

***Abstract*—In human-computer interaction, it is important to accurately estimate the hand pose, especially fingertips. However, traditional approaches to fingertip localization mainly rely on depth images and thus suffer considerably from noise and missing values. Instead of depth images, stereo images can also provide 3D information of hands. There are nevertheless limitations on the dataset size, global viewpoints, hand articulations and hand shapes in publicly available stereo-based hand pose datasets. To mitigate these limitations and promote further research on hand pose estimation from stereo images, we build a new large-scale binocular hand pose dataset called THU-Bi-Hand, offering a new perspective for fingertip localization. In the THU-Bi-Hand dataset, there are 447k pairs of stereo images of different hand shapes from 10 subjects with accurate 3D location annotations of the wrist and five fingertips. Captured with minimal restriction on the range of hand motion, the dataset covers large global viewpoint space and hand articulation space. To better present the performance of fingertip localization on THU-Bi-Hand, we propose a novel scheme termed Bi-stream Pose Guided Region Ensemble Network (Bi-Pose-REN). It extracts more representative feature regions around joints in the feature maps under the guidance of the previously estimated pose. The feature regions are integrated hierarchically according to the topology of hand joints to regress a refined hand pose. Bi-Pose-REN and several existing methods are evaluated on THU-Bi-Hand so that benchmarks are provided for further research. Experimental results show that our Bi-Pose-REN has achieved the best performance on THU-Bi-Hand.**

***Index Terms*—Fingertip localization, Hand pose estimation, Region ensemble network, Human-computer interaction, Hand pose dataset.**

I. Introduction

HAND pose estimation is one of the most important techniques in many applications like virtual reality and augmented reality [1]–[4]. Recently hand pose estimation from depth images has drawn considerable research attention [5]–[14], due to the emergence of depth cameras [15]–[18]. Compared with other hand joints, fingertips are much more challenging to localize because of the high fingertip flexibility, large viewpoint variation and poor depth quality [5]. Traditionally stereo-based hand poses are estimated by converting stereo images into depth images with full stereo matching algorithms [19]. However, depth values around fingertips can be inaccurate as large noise is introduced during conversion. As a result, fingertip localization is hindered by accumulated errors in disparity calculation and hand pose estimation.

Recently, several deep network methods have been developed for fingertip localization from binocular images [20]–[22], without converting them into depth images. However, their network architectures could not exploit the features of binocular images effectively. Furthermore, in publicly available datasets of stereo-based hand poses [21], [23], there have been limitations on the dataset size, viewpoints, hand articulations and hand shapes. These limitations substantially limit the generalization ability of trained models.

In this paper, we propose a novel hand pose estimation approach, named as Bi-stream Pose Guided Region Ensemble Network (Bi-Pose-REN), to estimate the locations of the wrist and fingertips from stereo images directly. Feature maps are extracted from left and right images by using DenseNet [24] in a two-stream style. Cropped around the location of joints in an initially estimated hand pose, the two-stream grid feature regions are first fused by concatenation and fully connected (FC) layers, and then integrated hierarchically according to the topology of hand joints. Under an iterative refinement framework, Bi-Pose-REN takes a previously predicted hand pose as input and improves the estimation in each iteration. Benefiting from the ensemble learning of multiple branches and the more representative features of joints, our proposed Bi-Pose-REN achieves the state-of-the-art performance over existing methods on our previous ThuHand17 [21] dataset.

To further promote research on stereo-based hand pose estimation, we build THU-Bi-Hand, a more large-scale binocular hand pose dataset, which contains about 447k pairs of stereo images from 10 different subjects with accurate annotations of six hand joint (five fingertips and the wrist) locations. In the dataset, 16 basic hand poses as well as transforming poses between pairs of basic poses were captured for each subject. The subjects were allowed to move their hands and fingers freely under the restriction that their hands appeared entirely in the valid imaging area. Captured from large diversity of hand shapes and hand poses, the new dataset covers the natural hand pose space commonly used in human-computer interaction (HCI), with little restriction on the range of hand motion including translation and rotation.

Our main contributions can be summarized as follows.

(1) We proposed a new approach, Bi-Pose-REN, to estimate fingertip locations directly from stereo images. Taking a previously estimated pose as input, Bi-Pose-REN extracts

G. Wang and C. Zhang are with the Department of Electronic Engineering, Tsinghua University, China. (Corresponding author: Guijin Wang)

X. Chen is with Huawei Noah's Ark Lab, China. Work was done when X. Chen was with the Department of Electronic Engineering, Tsinghua University, China.

X. Ji is with the Department of Automation, Tsinghua University, China.

J.-H. Xue is with the Department of Statistical Science, University College London, UK.

H. Wang is with Beijing Huajie IMI Technology Co., Ltd., China.

more representative feature regions of joints and provides more accurate localization.

(2) We built a new large-scale dataset THU-Bi-Hand[1], which consists of about 447k stereo image pairs from 10 subjects with large variant hand poses and movements as well as accurate 3D location annotations of fingertips and wrist.

(3) We provided several benchmarks on the THU-Bi-Hand dataset, offering a new perspective for fingertip localization. We evaluated several methods including [20], TSBNet [21], Bi-REN [22] and the proposed Bi-Pose-REN. Ablation studies of Bi-Pose-REN were also introduced to analyze the module effects.

The rest of this paper is organized as follows. We review related work in Section II, describe details of the new Bi-Pose-REN in Section III, and introduce the construction procedure and detailed information of the THU-Bi-Hand dataset in Section IV. Comparative studies with extensive experiments are presented in Section V, and Section VI concludes the whole paper and discusses some future work.

II. Related Work

In this section, we first review popular hand pose datasets, then discuss existing methods of hand pose estimation from stereo images, and finally review different feature extraction methods used in hand pose estimation from depth images.

A. Hand Pose Datasets

Publicly available hand pose datasets can be classified into two kinds: depth image based datasets [5], [25]–[28] and stereo image based datasets [21], [23].

1) Depth image based datasets: The NYU dataset [5] contains over 72k RGB-D images from one subject in the training set and 8k images from two different subjects (one of them is the subject in the training set) for testing with 36 annotated joints. The depth maps were collected from Microsoft Kinect camera [15] with missing values along occluded boundaries and noisy outlines [29].

The ICVL dataset [25] has 300k images with different rotations from 10 subjects with 26 gestures for training and 1.6k images for testing. The depth images were captured by Intel RealSense [18] with locations of 16 joints annotated. The depth maps have a high quality with few missing values and sharp outlines with little noise, but lots of samples were annotated incorrectly in both training and test sets (about 36% of the poses from the test set were annotated with errors of at least 10mm) [29].

The MSRA dataset [26] contains 76.5k depth images collected with Intel Creative Interactive Camera. Totally 21 joints were annotated. There are 9 subjects with 17 gestures for each subject. The variation of hand poses is limited in this dataset.

The BigHand2.2M dataset [27] contains 2.2 million depth maps with 21 accurately annotated joint locations. It has large diversity of global viewpoint, hand articulation and orientation. There are 10 hand shapes in the training set and an additional shape for testing.

[1]Dataset available at https://sites.google.com/view/thubihand or http://image.ee.tsinghua.edu.cn/data/thubihand.

The HandNet dataset [28] was created from 10 participants, containing more than 210k depth images captured by Intel RealSense camera, with annotations of the hand center and five fingertips.

2) Stereo image based datasets: In [23], a stereo hand pose dataset was established, containing 18k stereo image pairs with annotations of palm and finger joints. The images were captured by a Point Grey Bumblebee2 stereo camera, divided into 12 different sequences. There is only one subject in this dataset. It is too small and contains only one hand shape, which limits the generalization ability of trained models.

In our previous ThuHand17 dataset [21], there are 117k binocular samples in the training set captured by Leap Motion from eight subjects. The dataset covers 16 basic hand poses and extra transitional poses. One subject mainly performed some basic poses while the other subjects performed all the basic and transitional poses. The test set contains another 10k binocular samples of two subjects. However, ThuHand17 is still not large enough, and the two subjects of its test set are included in the eight subjects of its training set.

Currently, publicly available datasets of stereo-based hand poses are insufficient for fingertip localization. In order to promote research of fingertip localization from stereo images, we built a new large-scale binocular hand pose dataset called THU-Bi-Hand. Totally about 447k stereo images from 10 different subjects were collected. The training set contains all samples of seven subjects and half of the samples of another two subjects, while the test set contains the rest samples. There are about 357k and 90k samples in the training and test sets, respectively. THU-Bi-Hand is the largest binocular hand pose dataset with a large variety of hand poses, hand movements and hand shapes.

B. Stereo-based Hand Pose Estimation

Recent approaches to hand pose estimation from stereo images can be categorized into two categories: indirect methods [23], [30] and direct methods [20]–[22], [31], [32]. Indirect methods first compute depth maps from stereo images, and then estimate hand poses from depth images. Direct methods estimate hand poses directly from stereo images.

The indirect method of [23] incorporates on-line training based skin color detector and constrained stereo matching to compute depth maps from stereo images and conduct hand segmentation. Then, depth-based hand pose tracking algorithms [33], [34] are used to estimate hand poses in stereo image sequences. However, it still cannot get rid of poor depth quality around fingertips, which will cause difficulties in fingertip localization. Furthermore, the error in depth map calculation from stereo images hinders the performance of depth-based hand pose estimation. In [30], depth proposals and hand poses are jointly optimized by using Markov-chain Monte Carlo (MCMC) sampling and two CNNs. The first CNN evaluates the consistency between the proposed depth images and the observed stereo images, while the second CNN estimates hand poses from the proposed depth images. However, it also suffers from poor depth quality around fingertips, as with [23]. Besides, it consumes much time with

a lot of depth proposals. A frame of stereo images under 200 MCMC proposals takes about 360 seconds during prediction.

Direct methods can avoid the influence of noise introduced during depth map conversion. In [32], a generative hand model based framework is proposed to optimize the hand pose that maximizes the color consistency of the two views of the hand, avoiding the explicit computation of disparity maps of relatively uniformly colored hands. However, an explicit definition of the hand model is required for model-driven methods. It is also sensitive to the initialization of hand poses and suffers from tracking failures. In [20], hand mask images extracted from binocular images are exploited to localize the palm center and fingertips by using a deep CNN without explicitly computing depth maps. However, it is unable to use the binocular image features sufficiently and ignores informative details of hands. In [21], original and hand mask images are concatenated along the channel dimension as input. Two-stream convolutional layers with the same structure but different parameters are used to extract high level features of left and right images separately. The features from two streams are fused by FC layers to estimate the pixel coordinates of the joints. However, the whole feature maps are used to regress joint positions. Localized features can be used for better estimation. In [22], inspired by the region ensemble strategy (REN) [35], [36], multi-view feature regions are extracted from the feature maps and fused for hand pose estimation from stereo images. But the positions of regions are fixed and same for all samples, which is not optimal for each joint in different samples. Different from [22], we exploit pose guided structured region ensemble network (Pose-REN) [37] to improve the performance of localizing hand joints. Moreover, we build a new large-scale dataset THU-Bi-Hand to promote research on fingertip localization from stereo images. Besides, we also provide several benchmarks on the THU-Bi-Hand dataset, offering a new perspective for fingertip localization.

C. Feature Extraction and Region Ensemble

CNN-based architectures are proved to be very powerful in many computer vision tasks due to their strong ability of image feature extraction [24], [38]–[41]. In [41], residual representations and shortcut connections are incorporated into CNNs to address the problem of accuracy degradation when networks go deeper. By inserting identity shortcut connections between convolutional layers, the network is forced to learn residual mapping, which is beneficial for improving performance in deeper networks. Residual connections also ease the optimization by providing faster convergence for relatively shallow networks.

In [24], the benefits of connections between layers are further exploited to formulate the DenseNets. DenseNets have several dense blocks connected by transition layers. Inside each dense block, every layer is connected to all other layers. For each layer in a dense block, the feature maps of all preceding layers are concatenated as inputs, and its own feature maps are used as inputs of all subsequent layers after being concatenated with feature maps of other layers. The dense connections and feature map concatenations can alleviate the vanishing-gradient problem, strengthen feature propagation and encourage feature reuse. However, DenseNets are highly memory consuming because of fast feature maps growing.

In [35]–[37], CNNs with residual connections are used for feature extraction in hand pose estimation and achieve promising results. In [35], [36], the REN is proposed to improve performance for hand pose estimation from depth images. With feature maps of the last convolutional layer, REN divides them into several grid regions. A region ensemble strategy is used to concatenate the FC layer outputs of different regions, which can represent multiple views of input images. Benefiting from the multi-view strategy in both training and testing as well as the ensemble learning of multiple branches, REN achieves a great improvement in depth-based hand pose estimation.

In [37], the region ensemble method is further exploited to generate the Pose-REN to boost the performance of hand pose estimation from depth images. Pose-REN is an iterative refinement procedure estimating more accurate hand poses in each iteration, taking previously estimated poses as input. It crops spatial regions around each joint of the previously predicted hand pose from the feature maps. The cropped feature regions are integrated hierarchically by FC layers following the topology of hand joints and produce a refined hand pose, which is used as a guidance for feature cropping in the next iteration.

III. Bi-Pose-REN

In this section, we describe the details of Bi-Pose-REN, the proposed approach to fingertip and wrist localization directly from stereo images. The framework is shown in Fig. 1.

Both cropped stereo images and masks are used as input to enhance the precision and the robustness of Bi-Pose-REN considering that original images have informative details of hands, while masks are robust to variation of hand appearances [21].

First, left image is concatenated with left mask along the channel dimension, while right image is concatenated with right mask, producing two-stream $96 \times 96 \times 2$ inputs. Then, two DenseNet branches with the same structure and parameters are used to extract feature maps from left and right inputs respectively. In contrast to [24], to ease the GPU memory consumption in Bi-Pose-REN, the input images are first forwarded to convolutional layers and one average pooling layer before being passed to dense blocks. Furthermore, batch normalization (BN) does not help in our regression task in practice. A similar phenomenon has been observed in other regression tasks such as image super-resolution [43]. As suggested in [43], unlike image classification tasks where scale-invariant softmax is used to make predictions, the different formulations of training and testing in the BN layers may deteriorate the accuracy for regression tasks. Therefore, we remove the BN layers of standard DenseNets in order to improve the performance, accelerate training and inference procedures, as well as reduce memory consumption.

Later on, inspired by Pose-REN [37], we use pose-guided region ensemble to estimate joint positions from stereo images.

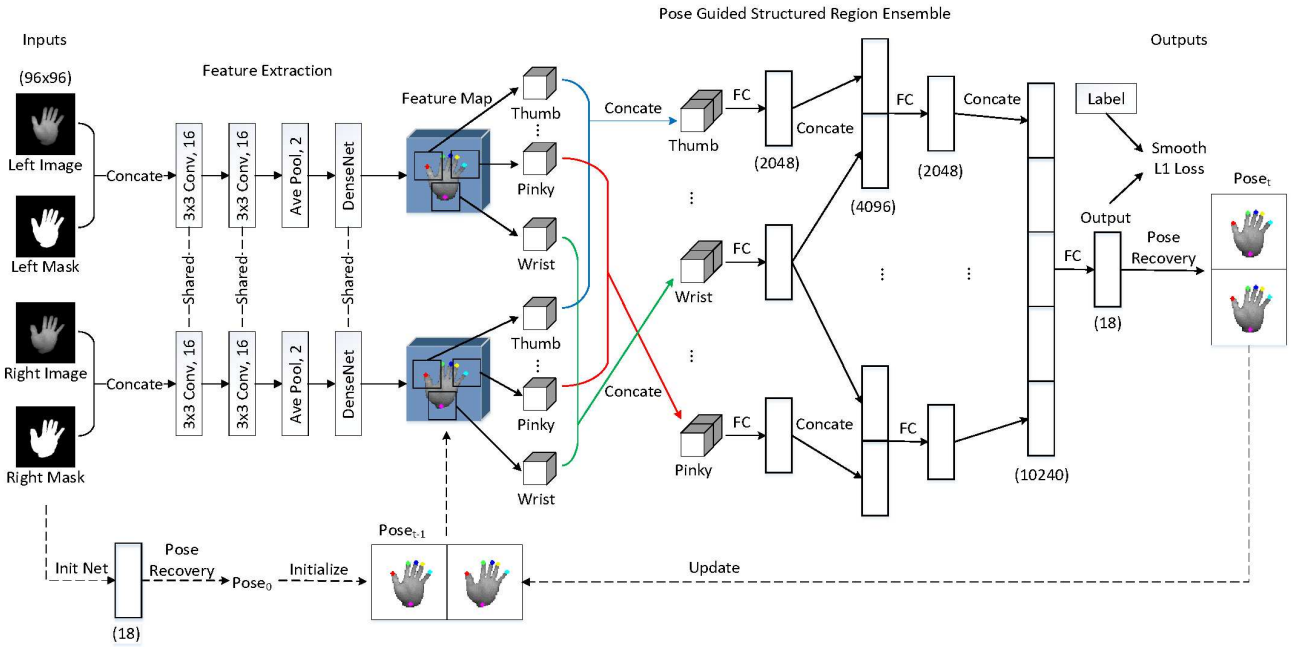


Fig. 1. The framework of Bi-Pose-REN. "$k \times k$ $Conv, c$" denotes a $k \times k$ convolutional layer with c output channels ($stride = 1, same\ padding$). "$Ave\ Pool, 2$" represents a 2×2 average pooling layer ($stride = 2, no\ padding$). "FC, n" indicates an n-dimensional fully connected layer. The sizes of images and vectors are in parentheses. Rectified linear unit (ReLU) [42] is used as the activation function.

From the feature maps in each stream, grid regions are cropped around previously estimated joints. Every two corresponding regions are concatenated to fuse the information from different streams. Then FC layers are used to integrate the fused regions hierarchically following the topology of hand joints. Finally, the 18-dimensional outputs are transformed into hand poses so Bi-Pose-REN can predict hand poses end-to-end.

Bi-Pose-REN refines joint locations iteratively from an initial hand pose predicted by an initialization network (Init Net). To explore the robustness of Bi-Pose-REN, different Init Nets are studied during inference.

Bi-Pose-REN differs from our previous Pose-REN [37] in three aspects. First, instead of using convolutional layers with residual connections, Bi-Pose-REN employs DenseNet [24] to extract feature maps. Second, Bi-Pose-REN processes feature extraction in two streams, which suits stereo inputs. Furthermore, the information from different streams is fused by concatenating corresponding feature regions after they are extracted from the feature maps.

A. Preprocessing and Inputs

As in [22], in order to preserve the completeness of hand regions as well as remove the background noise, two thresholds th_1 and th_2 are used to extract hand regions. First, rough binary images are obtained by thresholding the infrared images at th_1. Only the largest connected components are preserved for the calculation of hand region centroids, around which the stereo images are cropped. Then delicate binary mask images are acquired by thresholding the cropped stereo images at th_2 (which is smaller than th_1). The pixels are set to zero if they are outside the hand masks in the cropped stereo images.

After resized into $w_p \times h_p$, the left and right images (both cropped stereo and mask images concatenated along the channel dimension in each stream) are used as inputs of Bi-Pose-REN. Multi-scale training as well as random translation and scaling are used to make the model more robust to different hand shapes. Three sizes (240×240, 220×220, and 200×200) are used to crop image patches from the full-size images. For testing, the images are cropped into the size of 200×200.

B. Feature Extraction by Bi-stream DenseNets

Like [22], Bi-Pose-REN extracts feature maps from left and right inputs by using two-stream DenseNet layers (with the same structure and parameters) with 3 dense blocks. But unlike [24], the BN layers are removed from Bi-Pose-REN. Each dense block contains two 3×3 convolutional layers with the growth rate of 24 (fewer layers and smaller growth rate than [24]). Before the first dense block with an initial channel number of 16, there are two convolutional layers and one average pooling layer. As for the transition layers between contiguous dense blocks, a 2×2, $stride$ 2 average pooling layer following a 1×1 convolution is exploited. In each stream, there are 10 convolutional layers in total.

C. Pose Guided Structured Region Ensemble Network

Inspired by Pose-REN [37], which can accurately estimate hand poses from depth images, we further extend the pose guided region ensemble idea to stereo-based fingertip localization. We use the initial network (Init Net) to estimate an initial hand pose, which is then used as the guidance in the

first iteration in Bi-Pose-REN. The feature maps extracted in Section III-B are cropped into several regions around each joint location according to the initial hand pose in both streams separately. To fuse the information from left and right streams, we concatenate the feature regions for the same cropping joint location along the channel dimension. Then the concatenated regions of the six joints are integrated hierarchically according to their topology, with the wrist and each fingertip concatenated first and the five finger parts concatenated later (See Fig. 2 (right)). FC layers are used to fuse the information of concatenated feature regions and refine the hand pose, which is then to guide the feature maps cropping procedure in the next iteration. In this manner, the estimated hand pose can be refined iteratively under the pose guided framework.

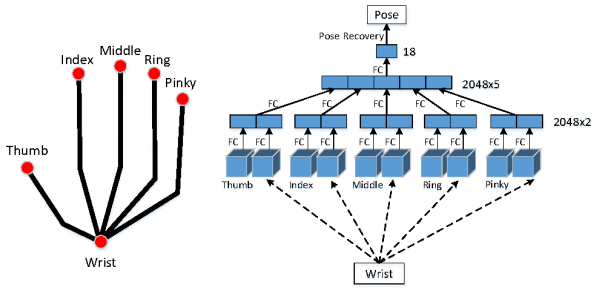


Fig. 2. The hand topology (left) and the architecture of the proposed region ensemble method (right). Each fingertip is fused with the wrist first. Afterwards, FC (fully connected layer) outputs of different fingers are fused to regress the final hand pose.

Denote the pixel coordinates of hand joints in the left and right images as (u_l, v_l) and (u_r, v_r), respectively, where v_l equals to v_r. These original coordinates are inconsistent with the Bi-Pose-REN inputs. Therefore, in order to reduce the difficulties of mapping joint positions from cropped stereo inputs, we use an easy mapping form:

$$label = \begin{pmatrix} \frac{(u_l - c_{xl}) + (u_r - c_{xr})}{w} \\ \frac{2((u_l - c_{xl}) - (u_r - c_{xr}))}{w} \\ \frac{(v_l - c_{yl}) + (v_r - c_{yr})}{h} \end{pmatrix}, \quad (1)$$

where (c_{xl}, c_{yl}) and (c_{xr}, c_{yr}) are the centroids of the segmented hand region in the left and right images, respectively (c_{yr} and c_{yl} are set to equal), and w and h denote the size of the cropped images (width and height before resized). $(u_l + u_r)$ and $(u_l - u_r)$ are the elements corresponding to the horizontal coordinates and disparities of joint positions; $(v_l + v_r)$ corresponds to the vertical coordinates. The coordinates of each joint in cropped images (with centroids subtracted) are normalized by w and h.

Bi-Pose-REN optimizes the smooth L1 loss [44] between the last FC layer output and the transformed label (3-dimensional target in Eq. (1) of six joints concatenated), with the threshold $th = 0.01$:

$$smooth_{L1}(x) = \begin{cases} 0.5|x|^2/th, & \text{if } |x| < th \\ |x| - 0.5th, & \text{otherwise} \end{cases}. \quad (2)$$

Afterwards, hand poses $((u_l, v_l)$ and (u_r, v_r), with v_l equals to $v_r)$ are obtained from the last FC layer outputs by a pose recovery layer.

IV. Dataset: THU-Bi-Hand

In this section, we introduce our method of building the THU-Bi-Hand dataset, using Leap Motion [45] to capture binocular images of hands with a resolution of 640×480, and the TrakSTAR tracking system with 6D magnetic sensors [46] to obtain accurate annotations of the locations of the wrist and five fingertips. We also present detailed information about the THU-Bi-Hand dataset, including the size of the dataset, etc.

A. Dataset Construction

Our hand model has 6 joints: the wrist and five fingertips (See Fig. 2 (left)). To avoid an ambiguous definition of the wrist, we consistently define the location of the wrist as the intersection of two lines: the first line is the bone connecting the root of the middle finger and the center of the palm; the second line is perpendicular to the first line and passes through the root of the thumb.

To build and annotate the THU-Bi-Hand dataset, we used Leap Motion and the TrakSTAR system with six 6D magnetic sensors. Leap Motion is a kind of infrared imaging binocular camera, so we can perform segmentation easily, using thresholds of grayscale. The TrakSTAR system can track the attached magnetic sensors by capturing their orientations and locations under the precision of 1.4mm. As shown in Fig. 3, we attached the six sensors to the back of the hand and on the defined wrist and five fingertips. Note that the locations of the magnetic sensors are a bit different from the real locations of the six joints since the joints are located inside the hand or fingers, not on the surface of the back. The coordinates of the six sensors in the coordinate system of TrakSTAR were acquired, while the coordinates that we really needed were the ones in the coordinate system of the Leap Motion (Fig. 4). To simplify the problem, the Leap Motion and the Transmitter of the TrakSTAR system were placed parallel to each other. Then we measured the position of the coordinate origin of the Leap Motion using one magnetic sensor; afterwards we can do calibration easily since there is no explicit rotation:

$$\begin{aligned} x &= y' - y'_0 \\ y &= z' - z'_0 \\ z &= -x' + x'_0 \end{aligned}, \quad (3)$$

where (x', y', z') and (x, y, z) are coordinates in the TrakSTAR coordinate system and the Leap Motion system respectively, and (x'_0, y'_0, z'_0) is the coordinates of the origin of the Leap Motion in the TrakSTAR coordinate system.

For better segmentation, given that Leap Motion is based on infrared imaging, volunteers were asked to wear black wrist straps. The TrakSTAR system consists of two electronic magnetic units, which are synchronized using Multi-Unit Sync Connections [46]. Each magnetic unit can attach at most four magnetic sensors, which are 2mm wide, with a 1.2mm wide and 3.3m long cable. The cables were attached to the hand using small and short tubes so they can move freely through

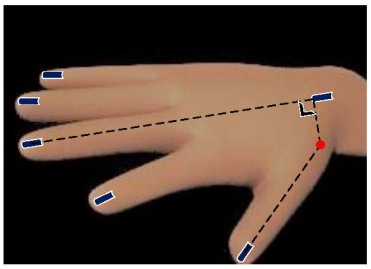


Fig. 3. Annotation setting. The sensors were attached to the back of the hand, on the locations of five fingertips and the wrist.

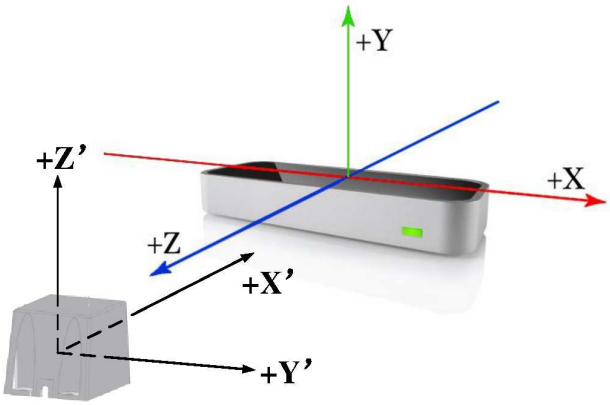


Fig. 4. The coordinate systems of TrakSTAR [46] and Leap Motion [45]. The devices were placed parallel to each other.

the tubes and the hand movements were less affected. The images and annotations were synchronized as in [27].

There are 16 basic hand poses in the THU-Bi-Hand dataset (see Fig. 5). Each subject was asked to perform the basic hand poses one by one, transforming from preceding pose to its next pose in the order shown in Fig. 5. The transformation was carried out several times before the subject performed the next basic pose. Both basic poses and transforming poses were captured. A reverse procedure was conducted after the 16th basic pose was finished. That is, the subjects transformed their hands from succeeding basic pose to its previous basic pose. Afterwards, the subjects chose several pairs of nonadjacent basic poses at random, and changed between each pair of chosen basic poses, for instance, transformed from the third basic pose (with the index finger and the middle finger stretched out) to the fifth basic pose (with only the thumb not stretched out), and vice versa.

The subjects were asked to move their hands freely during the whole sampling process. Finger movements like clicking and swinging were appreciated. As long as the valid imaging area contained the entire hand, the hands were allowed to translate in all directions. Besides, provided that the initial palm was on a plane parallel to the Leap Motion's imaging plane with the hand direction extended perpendicularly to the Leap Motion's baseline (the virtual line connecting the left camera and the right camera), the subjects could rotate their hands around all the three axis, each within 90 degrees.

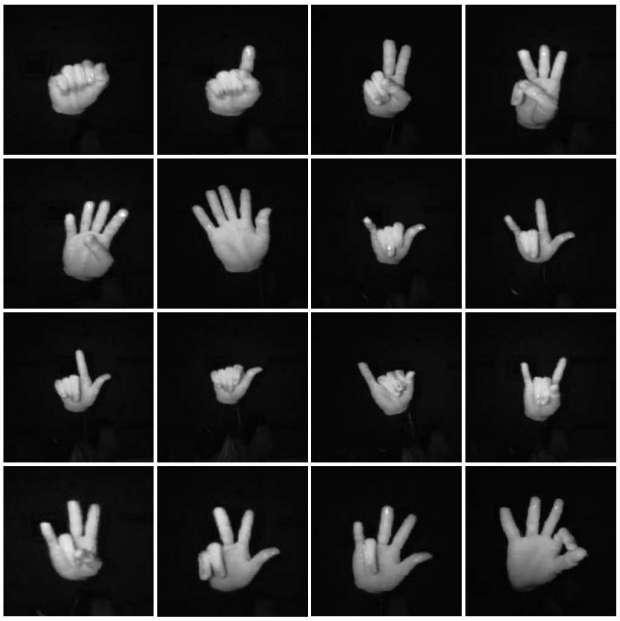


Fig. 5. The 16 basic hand poses in the THU-Bi-Hand dataset.

B. Glimpse of the THU-Bi-Hand Dataset

Similar to [21], we defined 16 basic hand poses (See Fig. 5). In order to better cover the articulation space, hand pose changing between two basic poses was also captured. As mentioned in Section IV-A, the THU-Bi-Hand dataset consists of two parts: (1) Basic poses: 16 kinds, almost fully cover all hand poses used in HCI. For each subject, about 1000 or more frames were captured for each basic pose. (2) Transforming poses: contains hand poses while subjects transformed their hand pose from one basic pose to another, including transforming poses between each pair of adjacent basic poses, about 800 frames for each pair, and more than 10000 frames of transforming poses between those pairs of basic poses which are not adjacent.

THU-Bi-Hand contains samples of ten subjects with different hand shapes, while the numbers of subjects in the datasets of [20] and [21] are only one and eight respectively. Totally about 447k frames of the left and right images with locations of six joints annotated were captured, which is almost 3 times larger than the ThuHand17 dataset [21]. In the experiments, we use half of the samples of two subjects and all samples of another subject for testing, while the remaining samples (including half of the samples of the two subjects and all samples of the remaining seven subjects) are used for training. The training set and the test set contain about 357k and 90k frames respectively. The subject for testing has a different hand shape from the subjects for training, so we can test the generalization ability of different methods. We must point out that, in [21], some hand poses in the test set have a larger range of rotation and translation than the training set, but the hand shapes in the test set also appear in the training set, unfavorable for testing models' generalization ability.

The THU-Bi-Hand dataset has a large variety of hand poses,

a large range of hand movements (rotation and translation), a large diversity of hand shapes and a large size of data samples. THU-Bi-Hand can be very beneficial for promoting research in hand pose estimation, especially fingertip detection from binocular images. Fig. 6 shows some example images (after preprocessing) and corresponding masks in THU-Bi-Hand.

C. Dataset Analysis

We compare the THU-Bi-Hand dataset and the hand pose benchmark in [23], by visualizing the global viewpoints, hand articulation and hand shapes of both datasets with t-SNE visualization [47], [48] and PCA projection.

1) Hand Viewpoint Space: During data sampling, the subjects were asked to explore the global viewpoint space as large as possible by moving their hands in all the directions, under the constraint that the hand region captured by the camera is neither too small nor too big. They were also asked to rotate their hands randomly, as long as the rotation angles were smaller than 90 degrees from the natural state facing the camera. As shown in Fig. 7 (left), the THU-Bi-Hand dataset covers much larger hand viewpoint space than the stereo dataset in [23].

2) Hand Articulation Space: In THU-Bi-Hand, all subjects were asked to perform 16 basic hand poses and abundant transforming poses between pairs of basic poses. Compared with [23], the THU-Bi-Hand dataset explores much larger hand articulation space (See Fig. 7 (middle and right)).

3) Hand Shape Space: There are totally 10 different hand shapes in the THU-Bi-Hand dataset, while [23] only contains one hand shape. Fig. 8 shows the 2D PCA projections of hand shapes in these two datasets.

V. Experiments and Discussions

In this section, Bi-Pose-REN, Chen et al. [20], TSBNet [21] and Bi-REN [22] are first evaluated on the ThuHand17 and THU-Bi-Hand datasets as the benchmarks for further research on fingertip localization from stereo images. Then extra experiments on the THU-Bi-Hand dataset are conducted to investigate the effectiveness of different modules in Bi-Pose-REN.

A. Experimental Setup

Bi-Pose-REN was implemented with Caffe [49] using C++. Stochastic gradient descent (SGD) was adopted with the mini-batch size of 128, a weight decay of 0.0005 and a momentum of 0.9. For experiments on ThuHand17, the learning rate started from 0.001 and was divided by 10 on iteration 100k, 160k and 200k, and the model was trained for total 240k iterations. As for THU-Bi-Hand, the learning rate was divided by 10 on iterations 300k, 500k and 600k, and the model was trained for totally 700k iterations. For Bi-Pose-REN, we trained the model for two iterations, and used the final model of the second iteration to test for one iteration.

TABLE I
Comparison with state-of-the-arts on ThuHand17 and THU-Bi-Hand. Our Bi-Pose-REN outperforms other methods.

Method	ThuHand17	THU-Bi-Hand
Chen et al. [20]	16.84mm	18.12mm
TSBNet [21]	10.91mm	13.27mm
Bi-REN [22]	8.98mm	9.47mm
Bi-Pose-REN (Ours)	**8.08mm**	**9.17mm**

B. Evaluation Metrics

The performance is evaluated via two metrics by following [21]: 1) *Average 3D distance error* is the Euclidean distance between the 3D coordinate predictions in the Leap Motion coordinate system and the ground-truths (in millimeters). 2) *Percentage of success frames* is the percentage of correctly predicted frames where all 3D distance errors of the six hand joints are smaller than a threshold.

C. Benchmarks of Fingertip Localization

We compare Bi-Pose-REN with previous work including Chen et al. [20], TSBNet [21] and Bi-REN [22], to demonstrate its effectiveness. The average 3D distance error and the percentage of success frames on the THU-Bi-Hand dataset are shown in Fig. 9. Table I presents the quantitative mean error for all the joints (the rightmost three bars labelled as "$Mean$") on both datasets.

The mean error of Bi-Pose-REN on the ThuHand17 dataset is 8.08mm, reduced from 16.84mm, 10.91mm and 8.98mm compared with Chen et al. [20], TSBNet [21] and Bi-REN [22], respectively (with about 52%, 26% and 10% improvements). As for the THU-Bi-Hand dataset, Bi-Pose-REN outperforms Chen et al. [20], TSBNet [21] and Bi-REN [22] by 49.39%, 30.90% and 3.17%, respectively. As shown in Fig. 9, Bi-Pose-REN localizes each joint in smaller average errors than Chen et al. [20], TSBNet [21] and Bi-REN [22] on THU-Bi-Hand. Except the wrist, all the joints have their average errors smaller than 10mm for Bi-Pose-REN.

Moreover, Bi-Pose-REN produces higher percentage of success frames than Chen et al. [20] and TSBNet [21] consistently, no matter with large or small thresholds. Compared with [22], Bi-Pose-REN is slighly better with small thresholds, and it is comparable with [22] with large thresholds. Specifically on the THU-Bi-Hand dataset, with a threshold of 20mm, the percentage of success frames of Bi-Pose-REN is higher than 75%, while TSBNet [21] about 45%, Chen et al. [20] lower than 20% and Bi-REN [22] about 75%. In summary, Bi-Pose-REN not only localizes the wrist and fingertips more accurately than previous methods, but also surpasses others in predicting frames under various levels of accuracy requirement.

Bi-Pose-REN runs at 55fps on an NVIDIA GeForce 1080TI GPU in inference phase (6.6ms for initialization plus 11.6ms for refinement), which is promising for real-time applications.

D. Ablation Studies

For ablation studies of our Bi-Pose-REN, we first introduce the modules and then evaluate different methods.

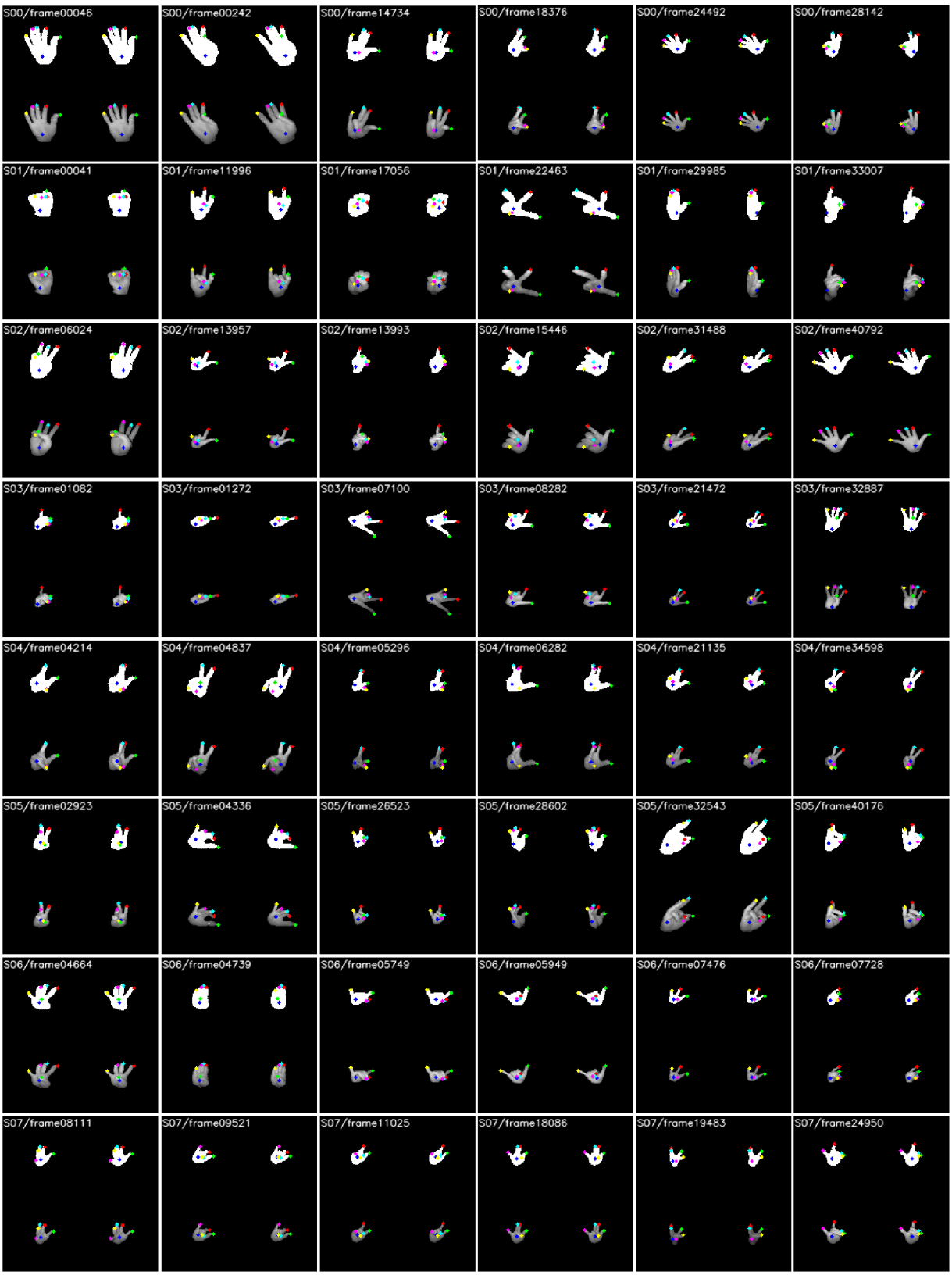


Fig. 6. Some examples from THU-Bi-Hand. Cropped masks and stereo images (both left and right) with annotations after preprocessing are shown.

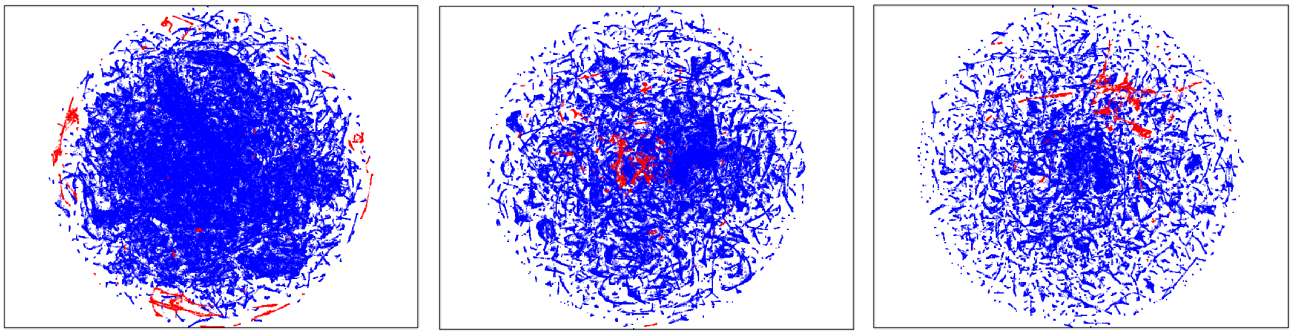


Fig. 7. 2D t-SNE visualization of the hand pose space. THU-Bi-Hand and the dataset in [23] are presented in blue and red, respectively. Left: global viewpoint space. Middle: hand articulation space. Right: combination of global viewpoint and hand articulation coverage. THU-Bi-Hand covers much larger hand pose space compared with the dataset in [23].

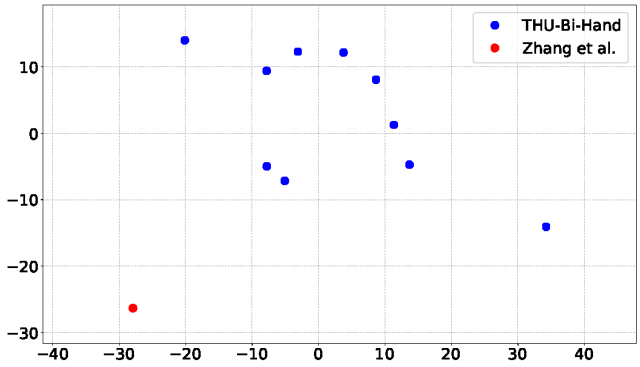


Fig. 8. 2D PCA visualization of hand shapes. THU-Bi-Hand contains 10 hand shapes while Zhang et al. [23] includes only one.

1) Module Introduction:

Feature Extraction Module. DenseNet [24] is exploited for feature extraction in Bi-Pose-REN. For comparison, the baseline in [36] (basic CNN, in the two-stream style) is used to extract features in other methods; see Fig. 10. Average pooling is used in DenseNet as [24], while max pooling is used in the basic CNN, following [36].

Regression Module. Region ensemble methods are proved to be effective in hand pose estimation from depth images [35]–[37]. To demonstrate the performances of different models on the THU-Bi-Hand dataset in the best way, we employ different region ensemble methods for fingertip localization from stereo images. REN is used in some methods while Pose-REN in others. While using REN, as in [36], nine regions are extracted in each stream. The regions of the left and right streams are concatenated and fed into two FC layers. Afterwards, the nine outputs are fused by concatenation and regress the hand pose.

We present several methods by using different modules, as listed in Table II.

2) Module Effects:

Feature Extraction. Compared with Basic-CNN, DenseNet-CNN produces a smaller mean error for all the six joints (see Fig. 11). As shown in Table III, the mean error of all joints in DenseNet-CNN is 10.51mm (9.47% better than Basic-CNN). Exploiting DenseNet for feature extraction also leads to higher percentage of success frames.

TABLE II
DIFFERENT METHODS OF LOCALIZING FINGERTIPS FROM STEREO IMAGES. ALL THESE METHODS EXTRACT FEATURE MAPS IN TWO STREAMS.

Method	Feature extraction	Regression
Basic-CNN	Basic CNN	FC layers
DenseNet-CNN	DenseNet	FC layers
DenseNet-REN	DenseNet	REN
Bi-Pose-REN	DenseNet	Pose-REN

TABLE III
ABLATION STUDIES OF BI-POSE-REN ON THE THU-BI-HAND AND THUHAND17 DATASETS.

Method	THU-Bi-Hand	ThuHand17
Basic-CNN	11.61mm	10.76mm
DenseNet-CNN	10.51mm	10.20mm
DenseNet-REN	9.47mm	8.98mm
Bi-Pose-REN	**9.17mm**	**8.08mm**

Regression Module. DenseNet-REN performs 9.90% better than DenseNet-CNN in per joint error (9.47mm versus 10.51mm, see Fig. 11 and Table III). Bi-Pose-REN outperforms DenseNet-REN by 3.17%. That is, the pose guided structured region ensemble method beats the region ensemble network without guidance, while it performs the worst without region ensemble. Similar patterns can be observed on the ThuHand17 dataset.

Initialization Network. We explore the robustness of Bi-Pose-REN over different pose initializations during inference. Different methods including Basic-CNN, DenseNet-CNN, Chen et al. [20] and TSBNet [21] were used to provide the initial pose in inference phase. The results of different initializations and refined results are shown in Fig. 12 and Table IV. It can be seen that our method boosts the performances of initializations. Even with some rather poor initializations (Chen et al. and TSBNet), the refined results are quite competitive. With better initializations (Basic-CNN and DenseNet-CNN), the final results are similar. The results demonstrate the robustness over initializations of our Bi-

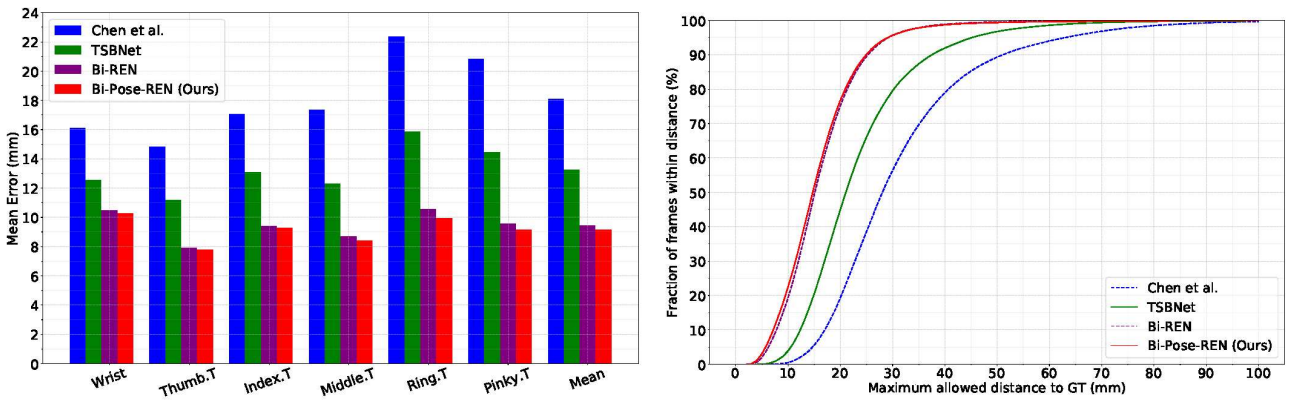


Fig. 9. Performance of Bi-Pose-REN, Chen et al. [20], TSBNet [21] and Bi-REN [22] on THU-Bi-Hand. Left: average 3D distance error. Right: percentage of success frames.

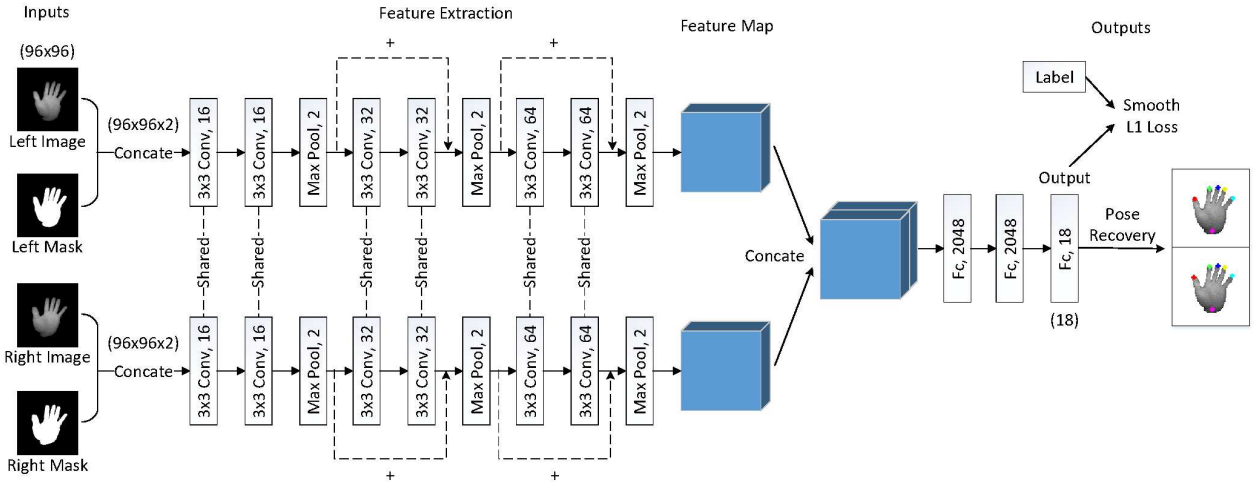


Fig. 10. The two-stream basic CNN. "$k \times k$ $Conv, c$" denotes a $k \times k$ convolutional layer with c output channels ($stride = 1, same\ padding$). "$Max\ Pool, 2$" means a 2×2 max pooling layer ($stride = 2, no\ padding$). "FC, n" is an n-dimensional fully connected layer. The sizes of images and vectors are in parentheses.

TABLE IV
EFFECTS OF DIFFERENT INITIALIZATION METHODS OF BI-POSE-REN ON THU-BI-HAND AND THUHAND17. THE AVERAGE 3D DISTANCE ERRORS (IN MILLIMETERS) OF REFINED AND INITIAL POSES ARE SHOWN OUTSIDE AND INSIDE PARENTHESES, RESPECTIVELY.

Init Net	THU-Bi-Hand	ThuHand17
Chen et al. [20]	11.10 (18.12)	9.80 (16.84)
TSBNet [21]	9.69 (13.27)	8.41 (10.91)
Basic-CNN	9.23 (11.61)	8.25 (10.76)
DenseNet-CNN	**9.17 (10.51)**	**8.08 (10.20)**

Pose-REN. Note that the model used above was trained with DenseNet-CNN initialization.

Iterations in Pose-REN. Bi-Pose-REN are cascaded frameworks with iterations. We explore the effect of the number of iterations by iteratively testing the Bi-Pose-REN model. The model was trained by using the samples with DenseNet-CNN initialization. The results of refined mean error of all the joints obtained from using different Init Nets during inference are presented in Fig. 13. It can be seen that Bi-Pose-REN converges very fast, as the results after only one or two iterations are adorable.

E. Qualitative Results

Some qualitative results of Bi-Pose-REN, Chen et al. [20], TSBNet [21] and Bi-REN [22] on the THU-Bi-Hand dataset are shown in Fig. 14. The predictions of Bi-Pose-REN are very close to the ground-truths and quite promising in difficult cases like side viewpoints and occluded fingers, while Chen et al., TSBNet and Bi-REN are poorer in these cases.

VI. CONCLUSION

In this paper we proposed a large-scale binocular hand pose dataset called THU-Bi-Hand, which contains about 447k frames of stereo images with accurate 3D location annotations of the wrist and five fingertips captured from 10 different

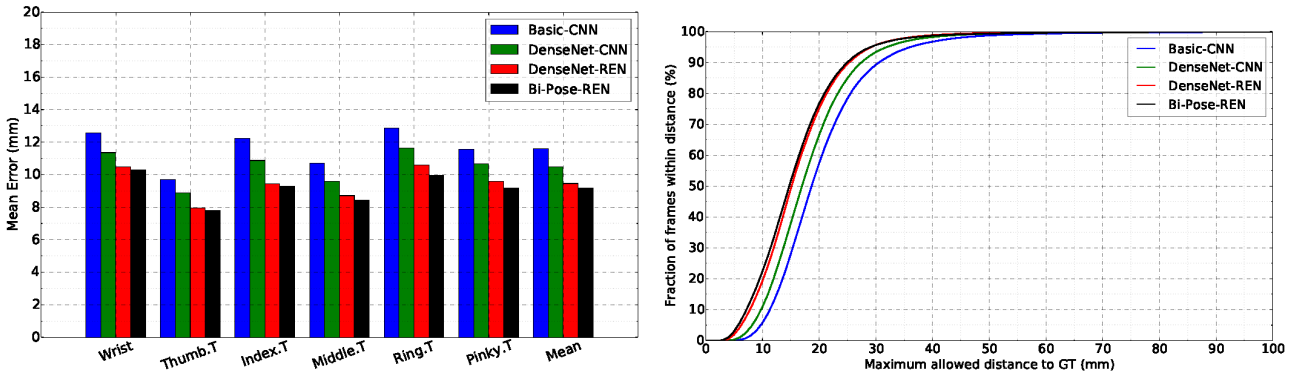


Fig. 11. Module effects of Bi-Pose-REN on THU-Bi-Hand. Left: average 3D distance error. Right: percentage of success frames.

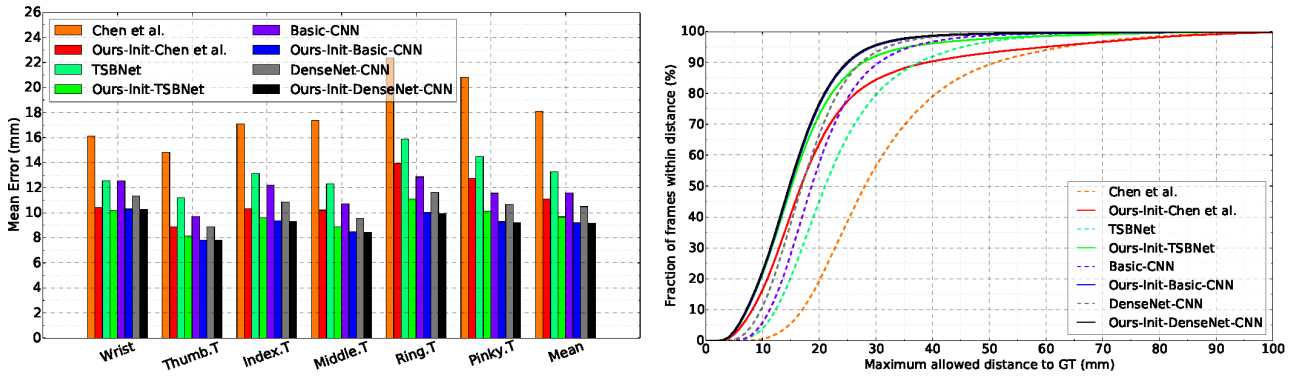


Fig. 12. Effects of different initialization methods of Bi-Pose-REN on THU-Bi-Hand. Left: average 3D distance error. Right: percentage of success frames.

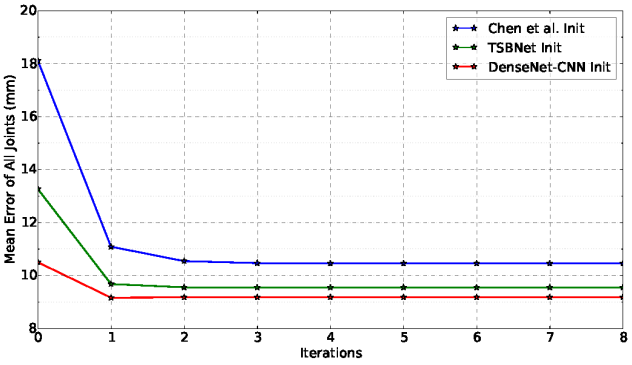


Fig. 13. Effect of the number of iterations in Bi-Pose-REN.

subjects. THU-Bi-Hand covers a large diversity of global viewpoints, hand shapes, hand poses and hand articulations, which has potential benefits to human-computer interaction with hand poses. We also proposed a novel approach termed Bi-Pose-REN to localize fingertips on THU-Bi-Hand. Feature regions are cropped around each joint from the feature maps given an initial hand pose and fused by FC layers to regress a refined pose iteratively. We evaluated several methods including Bi-Pose-REN on the THU-Bi-Hand dataset to provide benchmarks for promoting further research on fingertip localization from stereo images. Hand pose estimation from stereo images directly is then practicable with high accuracy. Future work will focus on fingertip localization from noisy stereo images or monocular RGB images, as well as hand pose estimation with hand-object interaction.

REFERENCES

[1] 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.

[2] J. H. Lee, T. Delbruck, M. Pfeiffer, P. K. Park, C.-W. Shin, H. Ryu, and B. C. Kang, "Real-time gesture interface based on event-driven processing from stereo silicon retinas," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 12, pp. 2250–2263, 2014.

[3] C. Nolker and H. Ritter, "Visual recognition of continuous hand postures," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 983–994, 2002.

[4] L. Zhao, X. Gao, D. Tao, and X. Li, "Learning a tracking and estimation integrated graphical model for human pose tracking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 12, pp. 3176–3186, 2015.

[5] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics (ToG)*, vol. 33, no. 5, p. 169, 2014.

[6] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, "Depth-based hand pose estimation: data, methods, and challenges," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1868–1876.

[7] D. Tang, J. Taylor, P. Kohli, C. Keskin, T.-K. Kim, and J. Shotton, "Opening the black box: Hierarchical sampling optimization for estimating human hand pose," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3325–3333.

[8] M. Oberweger, P. Wohlhart, and V. Lepetit, "Training a feedback loop for hand pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3316–3324.

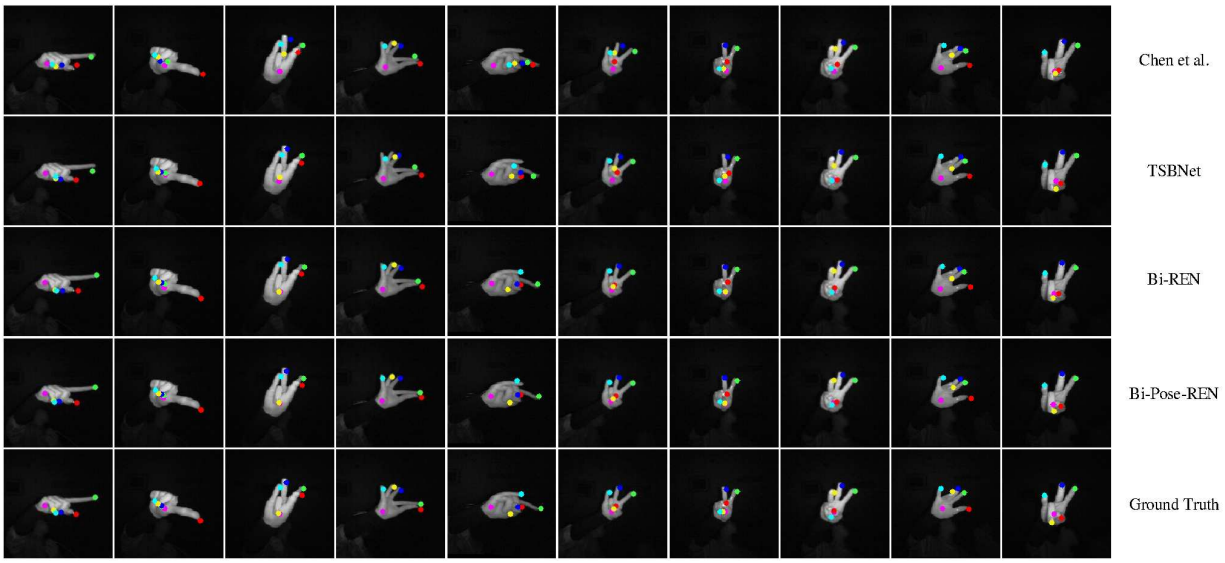


Fig. 14. Qualitative results on the THU-Bi-Hand dataset. Predictions of Chen et al. [20], TSBNet [21], Bi-REN [22] and our Bi-Pose-REN as well as the ground truths are shown in different rows. Bi-Pose-REN predicts promising results even in some difficult cases.

[9] C. Wan, A. Yao, and L. Van Gool, "Hand pose estimation from local surface normals," in *European Conference on Computer Vision*. Springer, 2016, pp. 554–569.

[10] C. Wan, T. Probst, L. Van Gool, and A. Yao, "Crossing nets: Combining GANs and VAEs with a shared latent space for hand pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[11] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Y. Chang, K. M. Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge *et al.*, "Depth-based 3D hand pose estimation: From current achievements to future goals," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[12] 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 Recognition*, vol. 2, no. 3, 2018.

[13] 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 Recognition*, vol. 1, 2017, p. 5.

[14] C. Zhang, G. Wang, H. Guo, X. Chen, F. Qiao, and H. Yang, "Interactive hand pose estimation: Boosting accuracy in localizing extended finger joints," *Electronic Imaging*, vol. 2018, no. 2, pp. 251–1–251–6, 2018.

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

[16] 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.

[17] C. Shi, G. Wang, X. Yin, X. Pei, B. He, and X. Lin, "High-accuracy stereo matching based on adaptive ground control points," *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1412–1423, 2015.

[18] L. Keselman, J. Iselin Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel RealSense stereoscopic depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1–10.

[19] P. Rogister, R. Benosman, S.-H. Ieng, P. Lichtsteiner, and T. Delbruck, "Asynchronous event-based binocular stereo matching," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 2, pp. 347–353, 2012.

[20] X. Chen, G. Wang, and H. Guo, "Accurate fingertip detection from binocular mask images," in *IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2016, pp. 1–4.

[21] Y. Wei, G. Wang, C. Zhang, H. Guo, X. Chen, and H. Yang, "Two-stream binocular network: Accurate near field finger detection based on binocular images," in *IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[22] C. Zhang, G. Wang, X. Chen, and H. Yang, "Bi-stream region ensemble network: Promoting accuracy in fingertip localization from stereo images." in *British Machine Vision Conference workshop on Image Analysis for Human Facial and Activity Recognition (BMVC Workshop)*, 2018.

[23] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang, "A hand pose tracking benchmark from stereo matching," in *IEEE International Conference on Image Processing*, 2017.

[24] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 2, 2017, p. 3.

[25] D. Tang, H. Jin Chang, A. Tejani, and T.-K. Kim, "Latent regression forest: Structured estimation of 3D articulated hand posture," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3786–3793.

[26] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 824–832.

[27] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim, "BigHand2.2M benchmark: Hand pose dataset and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 2605–2613.

[28] A. Wetzler, R. Slossberg, and R. Kimmel, "Rule of thumb: Deep derotation for improved fingertip detection," *arXiv preprint arXiv:1507.05726*, 2015.

[29] M. Oberweger, P. Wohlhart, and V. Lepetit, "Hands deep in deep learning for hand pose estimation," in *Proceedings of Computer Vision Winter Workshop*, 2015, pp. 21–30.

[30] R. R. Basaru, C. Child, E. Alonso, and G. Slabaugh, "Hand pose estimation using deep stereovision and markov-chain monte carlo," in *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2017, pp. 595–603.

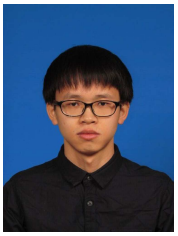
[31] J. Romero, D. Kragic, V. Kyrki, and A. Argyros, "Dynamic time warping for binocular hand tracking and reconstruction," in *IEEE International Conference on Robotics and Automation*. IEEE, 2008, pp. 2289–2294.

[32] P. Panteleris and A. Argyros, "Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo," in *IEEE International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2017, pp. 575–584.

[33] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and robust

hand tracking from depth," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1106–1113.
[34] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3D tracking of hand articulations using Kinect," in *BMVC*, vol. 1, no. 2, 2011, p. 3.
[35] H. Guo, G. Wang, X. Chen, C. Zhang, F. Qiao, and H. Yang, "Region ensemble network: Improving convolutional network for hand pose estimation," in *IEEE International Conference on Image Processing*. IEEE, 2017, pp. 4512–4516.
[36] G. Wang, X. Chen, H. Guo, and C. Zhang, "Region ensemble network: Towards good practices for deep 3D hand pose estimation," *Journal of Visual Communication and Image Representation*, 2018.
[37] X. Chen, G. Wang, H. Guo, and C. Zhang, "Pose guided structured region ensemble network for cascaded hand pose estimation," *Neurocomputing*, 2019.
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[42] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.
[43] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. Huang, "Wide activation for efficient and accurate image super-resolution," *arXiv preprint arXiv:1808.08718*, 2018.
[44] R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 1440–1448.
[45] "Leap Motion," https://www.leapmotion.com/.
[46] "Ascension Trakstar," http://www.ascension-tech.com/.
[47] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
[48] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.
[49] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

Guijin Wang received the B.S. and Ph.D. degree (with honor) from the Department of Electronic Engineering, Tsinghua University, China in 1998 and 2003, respectively, all in signal and information processing. From 2003 to 2006, he has been with Sony Information Technologies Laboratories as a researcher. From Oct. 2006, he has been with the Department of Electronic Engineering, Tsinghua University, China as an Associate Professor. From Jan. - Jun. 2012, he was a visiting researcher in AMP Lab of Cornell. He won the reward (the first prize) of Science and Technology Award of Chinese Association for Artificial Intelligence in 2014, won the reward (the second prize) of Shandong Province Science and Technology Progress in 2014. He was an Associate Editor of IEEE Signal Processing Magazine, a Guest Editor of Neurocomputing, the track chair of ChinaSIP 2015, and the TPC member of ICIP2017. He published over 100 international journal and conference papers, and holds tens of patents with numerous pending. His research interests focus on computational imaging, pose recognition, intelligent human-machine UI, intelligent surveillance, industry inspection, AI for Big medical data, etc.



Cairong Zhang received his B.S. degree from the Department of Electronic Engineering, Tsinghua University, China, in 2017, where he is currently working towards his M.S. degree. His research interests include deep learning, human pose estimation and hand pose estimation.

Xinghao Chen received his B.S. and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, China, in 2013 and 2019, respectively. From Sept. 2016 - Jan. 2017, he was a visiting Ph.D. student with Imperial College London, UK. His research interests include deep learning, hand pose estimation and gesture recognition.

Xiangyang Ji received the B.S. degree in materials science and the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He joined Tsinghua University, Beijing, China, in 2008, where he is currently a Professor in the Department of Automation. His current research interests cover signal processing, image/video processing and machine learning.

Jing-Hao Xue received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998 and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is an Associate Professor in the Department of Statistical Science, University College London. His research interests include statistical classification, high-dimensional data analysis, pattern recognition and image analysis.

Hang Wang received the B.S. degree in automation from Beijing Institute of Technology in 2008. He received the M.S. degree in control science and engineering from Beijing Institute of Technology in 2010. From 2014 to 2015, he was a senior software engineer in Baidu. He is the vice president of Beijing HJIMI Technology Co., Ltd since 2015.