SRHandNet: Real-time 2D Hand Pose Estimation with Simultaneous Region Localization

Yangang Wang, *Member IEEE,* Baowen Zhang, and Cong Peng *Member IEEE*

Abstract—This paper introduces a novel method for real-time 2D hand pose estimation from monocular color images, which is named as **SRHandNet**. Existing methods can not time efficiently obtain appropriate results for small hand. Our key idea is to simultaneously regress the hand region of interests (RoIs) and hand keypoints for a given color image, and iteratively take the hand RoIs as feedback information for boosting the performance of hand keypoints estimation with a single encoder-decoder network architecture. Different from previous region proposal network (RPN), a new lightweight bounding box representation, which is called **region map**, is proposed. The proposed bounding box representation map together with hand keypoints heatmaps are combined into the unified multi-channel feature maps, which can be easily acquired with only one forward network inference and thus improve the runtime efficiency of the network. Our proposed SRHandNet can run at 40fps for hand bounding box detection and up to 30fps accurate hand keypoints estimation under the desktop environment without implementation optimization. Experiments demonstrate the effectiveness of the proposed method. State-of-the-art results are also achieved out competing all recent methods.

Index Terms—real-time hand pose estimation, bounding box representation, inference feedback

1 Introduction

Real-time hand pose estimation is very important in the area of articulated object pose estimation and it is a key step for many practical applications, such as human-machine interactions, virtual reality (VR), augmented reality (AR) and *etc*. In the past few years, hand pose estimation has been greatly developed with the introduction of commodity depth sensors [1]. Nevertheless, the restricted sensing distances as well as low resolution depth maps remain the problem to be unsolved in its full generality. Recently, with the rapid development of Convolutional Neural Networks (ConvNets), hand pose estimation from color images has been discussed by a lot of research [2]. Even so, achieving this goal is still challenging due to the flexible hand fingers movements, self-occlusions and appearance ambiguities in color images. It is yet far from being completely solved for real-time hand pose estimation.

This work focuses on the problem of estimating, in real-time, the 2D hand pose from monocular color images.

- *This work was supported in part by the National Key R&D Program of China (No. 2018YFB1403901), in part by the National Natural Science Foundation of China (No. 61806054, 61703203), in part by the Natural Science Foundation of Jiangsu Province (No. BK20180355, BK20170812) and in part by the Shenzhen Science and Technology Innovation Committee (STIC) (No. JCYJ20180306174459972).*
- *Yangang Wang is now with the School of Automation, Southeast University, Nanjing, China, 210096. He is also a member of Shenzhen research institute of Southeast University. Before that, he was an associate researcher at Microsoft Research. E-mail: ygwangthu@gmail.com. Personal website: http://www.yangangwang.com. Corresponding author: Yangang Wang.*
- *Baowen Zhang is with the School of Automation, Southeast University, Nanjing, China, 210096.*
- *Cong Peng is with the School of Automation, Nanjing University of Aeronautics and Astronautics, Nanjing, China, 211106.*

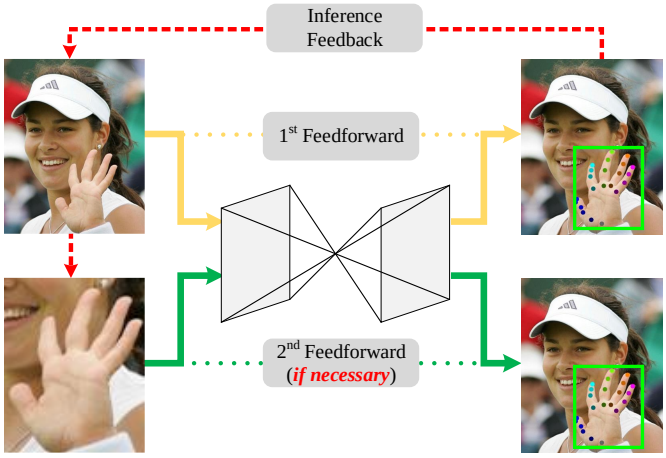


Fig. 1. Given a single color image, SRHandNet can estimate the hand keypoints as well as the hand bounding box with only one network. In order to boost the performance of small hand pose estimation, we iteratively take the hand RoIs as feedback information and redo the inference with the same encoder-decoder network. The proposed SRHandNet is time efficient for realtime applications.

Specifically, the 2D hand pose is described by 21 pre-defined keypoints. Our goal is to estimate these keypoints in real-time with a conventional RGB camera or the monocular color images input. Inspired by several state-of-the-art neural networks for human pose estimation [3], we propose **SRHandNet**, which takes the encoder-decoder network architecture as the backbone of our network. It is noted that all the existing methods for accurate pose estimation are time-consuming. And for small hand, they can not obtain appropriate pose estimation results.

A naïve and straightforward solution for small hand pose estimation is to obtain the hand bounding box at first,

and consequently take the extracted hand region of interests (RoIs) as tight hand images input to perform the hand pose estimation, *e.g.*, Simon *et al*. [2] used the human body to guide the localization of hands and then cropped the hand images for obtaining accurate hand pose estimation. In general, a neural network based bounding box regression, such as Faster-rcnn [4], Mask-rcnn [5] or ConerNet [6], would be more appropriate for detecting the hand in color images. Nonetheless, the sequential two steps, *i.e.*, region detection and then pose estimation, are not efficient for the scenario of real-time hand pose estimation. Moreover, existing region detection networks can not obtain appropriate results for small hand since it is still an open problem for small object detection.

To solve the above puzzle, our key idea is to simultaneously regress the hand regions and keypoints from a given image, and iteratively take the hand RoIs as feedback information for boosting the performance of hand keypoints estimation with a single encoder-decoder network architecture, as shown in Fig. 1. Different from previous region proposal network (RPN), we propose a novel lightweight bounding box representation, which is called **Region Map** as described in Sec. 3.1. The proposed bounding box representation are combined with hand keypoints heatmaps to formulate the 25-channel feature maps for efficient training and inference. SRHandNet takes the advantage of feedback to enhance the 2D hand pose estimation accuracy for small hand on the fly. The 2D hand pose estimation is then performed in a cycle way other than one pass of the network, as described in Sec. 3.3. Benefited from the proposed bounding box representation, the region localization and hand keypoints detection could be easily implemented in a unified framework. Besides, the network inference can be further accelerated by whether adopting the cycle detection or not according to the size of hand in the input.

With the proposed method, we show remarkable real-time 2D hand pose estimation results. Our method can pursue high efficient end-to-end hand pose estimation results and run at 40fps for hand bounding box detection and up to 30fps for hand keypoints estimation under ordinary desktop environment with a single GTX1080Ti GPU. Although our method only focuses on the 2D hand pose estimation other than 3D, we believe that 3D pose estimation can be greatly benefited and directly lifted from 2D pose estimation. The code, dataset and all relevant stuff will be available on the project website **www.yangangwang.com**.

2 Related Work

There are lots of work focusing on the hand pose estimation from color images. In the following, we will briefly introduce some related works.

Hand Pose Estimation. Research on hand pose estimation lasts for a long time and numerous approaches have been proposed in decades [7]. Despite the efforts, hand pose estimation is still a challenging task due to the similarities among fingers, significant self-occlusions, flexible hand poses and *etc*. For the past few years, inspired by the depth based human pose estimation [8], hand pose estimation has made great progresses on the introduction of commodity depth sensors [9], [10], [11], [12], [13]. However, current depth sensors are always restricted in the indoor environments, which limits the scope of usage with depth sensors for hand pose estimation. In this paper, we focus on the 2D hand pose estimation from RGB images.

Similar as many other computer vision tasks, hand pose estimation and more generally, pose estimation from RGB images are also benefited by the Convolutional Neural Networks (CovNets), which is first introduced in DeepPose [14]. It is well known that deep learning requires large number of training data. However, hand datasets with high-quality in-the-wild color images are seriously inadequate compared against human pose datasets. In order to solve this challenge, Wang *et al*. [15] proposed a dataset named OneHand10K, which includes both the hand masks and visible hand keypoints. Simon *et al*. [2] proposed a multiview bootstrapping method to train the hand keypoint detectors, which allows the generation of large annotated datasets using a weak initial detector.

Beyond the challenge of inadequate training data, network architecture is another main factor for the performance of hand pose estimation. In these years, a majority of different network architectures, *e.g.*, [16], [17], [18], [19], [20] are proposed for human pose estimation. However, compared with human body pose estimation, hand always cover small region in the color images. This indicates that directly shifting the networks for human pose estimation might generate inaccurate hand pose estimation results.

In this paper, the dataset with high-quality in-the-wild hand color images is downloaded from [15] for enabling easy training. As for the network, we investigated the performance of different networks for human pose estimation in MPII [3] and found that the top ones were almost the varieties of the Hourglass network [21], which used the encoder-decoder architecture. We finally choose the encoder-decoder network architecture as our network backbone to perform the 2D hand pose estimation.

Recently, some pioneers attempt to estimate the 3D hand pose from monocular color images. Panteleris *et al*. [22] combined the deep learning based 2D hand pose estimation with the power of generative techniques to achieve real-time monocular 3D hand pose estimation in unrestricted scenarios. Iqbal *et al*. [23] proposed a novel CNN structure to estimate the 2.5D heatmaps and obtain the 3D hand keypoints coordinates from the 2.5D heatmaps. Cai *et al*. [24] proposed an end-to-end 3D hand pose estimation convolutional neural network by leveraging a depth regularizer to enhance the accuracy of 3D hand pose estimation. Our proposed SRHandNet has the potential to improve the performance of these approaches.

Region Localization. We adopt the hand region localization to lift the performance of 2D hand pose estimation. Considering the computational cost, our hand region localization is performed as bounding box regression other than semantic segmentation at the pixel level. Typically, semantic segmentation for object region representation plays an important role in many computer vision tasks, such as medical image processing [25], [26], [27], autonomous driving [28], [29] and *etc*. Previous works [15] demonstrate that hand semantic segmentation could benefit for the pose estimation in the framework of deep learning. Nevertheless, obtaining

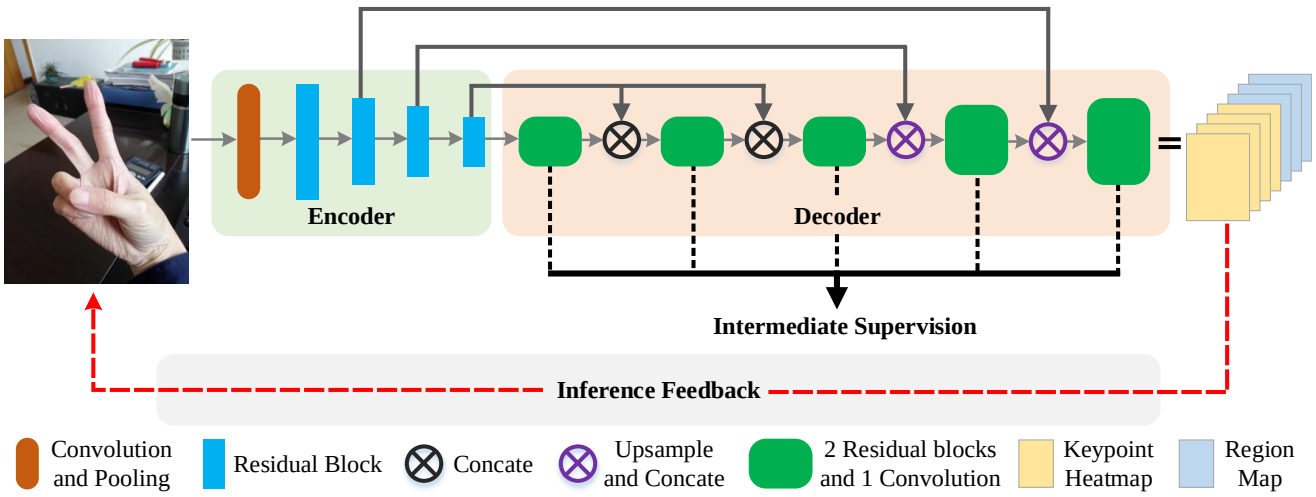


Fig. 2. Overview of the proposed SRHandNet. We use an encoder-decoder architecture as the backbone of our network to perform the 2D hand pose estimation. In the training stage, the intermediate supervision is adopted. In the inference stage, we perform the cycle detection according to the size of hand. See the texts of Sec. 3.2 for more details.

majorities of pixel-level hand segmentation dataset for network training is difficult and hugely expensive.

Bounding box regression, on the other hand, is often represented as object recognition in literature [30]. There are many works focusing on this problem and deep learning are recently the main streams against three widely known object detection competitions, *i.e.*, PASCAL VOC [31], ILSVRC [30] and COCO [32]. However, all of these competitions focus on generic object detection and none of them have specific tasks for hand detection. Still, we can shift existing deep learning network structures to solve the problem of hand region localization from color images. It is noted that, among all the deep learning networks, the representative one is region proposal network (RPN) [33]. RPN uses two-steps processing for object detection and several region proposals are selected at first by the network. Fast-rcnn [34], Faster-rcnn [4] and Mask-rcnn [5] are the recent representative methods to improve the detection accuracy as well as computational efficiency for object detection with region proposals. However, the two-steps processing has heavy computational cost during training and inference, which is not suitable for realtime applications.

Targeting to achieve lower computational cost without reducing the detection accuracy, Law *et al*. [6] proposed a novel one-step network architecture named as CornerNet. However, the introduced corner pooling is very complex for easy usage and general adaption. In this paper, we propose a simple yet effective bounding box representation. Our main idea is to represent the bounding box as 3-channel feature maps, thus we can enable the simultaneous training of joint heatmaps and region localization. We use the idea of fully convolutional network [35] to train the bounding box feature maps. All the details are presented in Sec. 3.1 and experiments demonstrate the performance of the proposed method.

Our method takes the advantage of feedback. The forward inferred hand regions are utilized to extract the RoIs and we can redo the hand pose inference with the same network, which is called cycle detection in Sec. 3.3. It is worth mentioning that the proposed cycle detection is similar to an attention model, yet we do not train this mechanism as previous networks [36]. It is noted that the network of Iterative Error Feedback [37] also takes the idea of feedback for pose estimation. Their method learns hierarchical feature extractors over the pose joint space by incorporating the top-down feedback. Different from their feedback procedure in the training, we take the feedback in the inference stage.

3 Method

An overview of the proposed SRHandNet is illustrated in Fig. 2. Given a color image $\mathbf{I}$ of the size $w \times h$ with a hand, our goal is to estimate the 2D positions of all the $K = 21$ keypoints of the hand. The 2D hand pose is represented as $\mathbf{p} = \{p_k\}_{k \in K}$, where $p_k = (x_k, y_k) \in \mathbb{R}^2$ is the 2D pixel coordinate of the k-th keypoint in image $\mathbf{I}$ and the order of these keypoints is defined as [15]. We use a heatmap to describe the k-th keypoint p_k, which is denoted as $\mathbf{H}_k$. The benefits of regressing a heatmap rather than a pixel coordinate directly have been discussed in literature [38]. In short, the heatmap representation is robust against data noisy. Each heatmap encodes the keypoint location via a 2D Gaussian distribution, where the mean is $\mu = p_k$ and the variance is $\sigma_k \in \mathbb{R}$. It is noted that if the k-th hand keypoint is not visible (*e.g.*, the keypoint is out of range or occluded), $\mathbf{H}_k$ is set by 0.

We use the idea of fully convolutional neural network to perform the hand pose estimation. For boosting the accuracy of small hand without losing the runtime efficiency, a novel bounding box representation, which is called **Region Map**, is proposed. The region map as well as hand keypoints heatmaps are integrated together and fed into the proposed network for an end-to-end training. In such case, we can simultaneously localize the hand region and estimate all the keypoints $\mathbf{p}$ of the hand with only one forward pass. The hand region localization information then can be utilized to extract the RoIs (region of interest) of the hand. We iteratively use the extracted RoIs as feedback information

(red dashed line in Fig. 2) to finetune the 2D hand pose by performing the cycle detection.

3.1 Region Map

Bounding box is very important for accurate hand pose detection. Previous bounding box representations, such as RPN [4] or CornerNet [6], are heavy for training or inference. These representations may degrade the runtime efficiency of 2D hand pose estimation, which is not good for real-time scenario. In this paper, we propose a novel bounding box representation, which is called **Region Map**. Fig. 3 illustrates the proposed region map for 2D hand bounding box representation. For each hand, our region map has 3 channels and the same size of hand keypoints heatmaps. Each individual channel encodes the center of hand, the width ratio as well as the height ratio of the bounding box with respect to the input image separately. The proposed bounding box representation are integrated with hand keypoints heatmaps to formulate the n-channel feature maps for an easy end-to-end training, where the n-channel feature map is illustrated as the orange and blue maps in the rightmost part of Fig. 2.

There are many ways to describe the bounding box. Other than training the corners of bounding box as CornerNet [6], which is not computationally efficient, the bounding box is represented by the **center** and **size** in this work. During training, the bounding box of each hand can be directly obtained from labels. It is worth noting that the training dataset [15] only contains visible 2D hand keypoints. For simplicity, we first compute the two corners of the bounding box from labeled visible 2D hand keypoints $\{p_k\}_{k \in K}$ as follows,

$$p_{lt} = \min_{k \in K} p_k; \tag{1}$$

$$p_{rb} = \max_{k \in K} p_k. \tag{2}$$

Where p_{lt} is the left-top corner and p_{rb} is the right-bottom corner of the hand bounding box.

And then, the center and the size of bounding box is computed as

$$c = p_{lt} + (p_{rb} - p_{lt}) * 0.5; \tag{3}$$

$$s = (p_{rb} - p_{lt}) * \alpha, \tag{4}$$

where c is the center of the hand and s is the size of the computed bounding box from labeled visible 2D keypoints. It is noted that we use a scale factor α to relax the computed compact bounding box since the labeled visible 2D hand keypoints always shrink the hand.

After that, we use a 2D Gaussian distribution with $\mu = c$ and the variance σ_r to obtain the 1^{st} channel of region map. The procedure is similar to the heatmap generation of hand keypoints. For the 2^{nd} and 3^{rd} channel of region map, we fill the same value s_x/w and s_y/h in a square region separately, and all other locations are filled with 0. Here, s_x and s_y are the x component and y component of the bounding box size s, respectively. The center of the square is c, the width and height of the square is $3\sigma_r$. Note that s_x/w and s_y/h both have the bound, *i.e.*, $0 \leq s_x/w \leq 1$ and $0 \leq s_y/h \leq 1$. This property gives us the same order of magnitude among the 3 channels of the proposed region map, which is convenient for training. In our current implementation, $\alpha = 1.3$ and $\sigma_r = 3$.

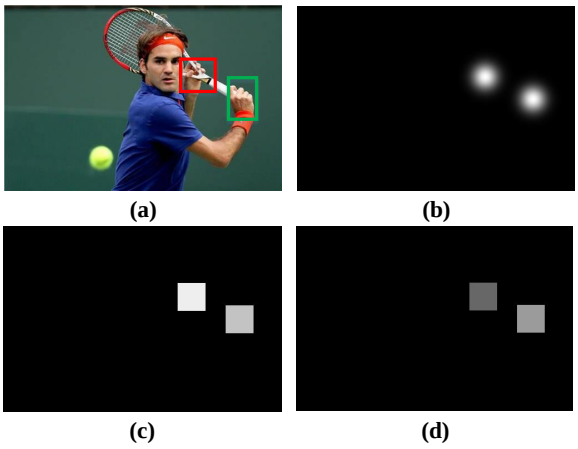


Fig. 3. The proposed region map. (a) is the original color image; (b) is the heatmap of the bounding box center; (c) is the map which encodes the width ratio between the bounding box and image width and (d) is the map which encodes the height ratio between the bounding box and image height. All the 3-channel maps are combined as the proposed **Region Map**. See texts for more details.

The proposed bounding box representation, *i.e.*, **Region Map**, is lightweight and has good performance for 2D hand pose estimation as demonstrated in the experiment section and supplemental videos. We hope that the proposed bounding box representation would inspire and promote the performance of time efficiency for other tasks, such as human pose estimation, small face detection and, *etc*.

3.2 Network Architecture

The proposed network is a closely connected fully convolutional network [35] and easy to train. We use an encoder-decoder architecture as the backbone to perform the 2D hand pose estimation, which is shown in Fig. 2. In recent years, there are many improvements for encoder-decoder network architectures aiming at different computer vision tasks, *e.g.*, [39], [40] use it to do semantic segmentation. One typical example is the hourglass network [21], which is first introduced for 2D human pose estimation. The original hourglass network is stacked by multiple hourglass modules for performance boosting. We found that the hourglass module is very memory consumption and hard to train for real-time 2D hand pose estimation. Compared with the original hourglass network, we perform several improvements for real-time 2D hand pose estimation in this paper.

Firstly, we abandon the use of several exhausting conv-deconv stages. Our network has only one encoder-decoder stage for runtime efficiency consideration. In the encoder part, the feature resolutions of the network are reduced 4 times along the way (352, 176, 88, 44, 22). We use stride 2 for the first feature resolution reduction and other feature resolution reduction are performed by maxpooling operations. As for the decoder part, the features are upsampled by the nearest neighbor upsampling operations with 2 times, which finally generates the output feature maps with the size of 88. Similar as previous encoder-decoder network structures in literature, we also do the skip connection

between the same scale of encoder and decoder feature maps. These feature maps are simply concatenated for the subsequent processing.

We make extensive use of residual modules proposed in ResNet [18]. Our network finally applies convolutional operations with the size of less than 3×3 completely except for the first feature resolution reduction, which is performed with a 7×7 convolutional operation. The keypoint heatmaps and region maps are combined as the output feature maps with the channel size of 25, where there are 21 channels for hand keypoints heatmaps, 3 channels for region maps and 1 for background heatmaps. For training the proposed network, we use the L_2 loss. Beyond that, we also perform the intermediate supervision in the 3 scales of decoder part, *i.e.*, (22, 44, 88), as shown in Fig. 2. The 25-channel output feature maps are upsampled and concatenated with the feature maps in the encoder part for further keypoints heatmaps and region maps estimation.

In particular, it is not trivial for performing the intermediate supervision in training. Newell *et al*. [21] claim that the intermediate supervision for only one encoder-decoder might be ignorant of critical global cues and is not good. Recently, Ke *et al*. [41] proposed a method to perform the supervision via downsampling the groundtruth heatmaps. However, we found that this strategy often has the divergent training results and finally fail the training. In their intermediate supervision strategy, the groundtruth heatmaps are generated from 2D Gaussian distributions and the Gaussian heatmaps have the influential range reduction with the power of 2 among scales. Suppose the heatmaps in the scale with the size of 88 are generated from 2D Gaussian distribution with the variance of σ, then the heatmaps in the scale with the size of 22 have only the $\sigma/4$ influential range, which is very small and hard to train.

We conducted a number of experiments and found that the training can be stably converged when the variance of 2D Gaussian distribution for generating the heatmaps is larger than a threshold for a specific size of heatmaps. That is

$$\sigma \geq \tau(m), \tag{5}$$

where $m \in (22, 44, 88)$ is the size of heatmaps, τ is a function of the groundtruth heatmap size. Instead of downsampling the groundtruth heatmaps for different scales, we finally chose $\sigma = 3.0$ for all the 3 scales to generate the heatmaps from 2D Gaussian distributions.

3.3 Cycle Detection

The proposed network in Sec. 3.2 performs well for the hand with regular size. However, we found that the 2D pose estimation accuracy dropped dramatically for small hands. In fact, it is still an open problem and challenging for the network to detect small objects. Most of the existing networks address this issue by taking the image pyramid information, such as training the face detectors with multiscale feature extraction [42] or feature pyramid network (FPN) [43], which is improved by the RoIAlign operation [5]. In this paper, we propose a novel strategy named **Cycle Detection**. Our key idea is to perform the 2D hand pose estimation in a cycle way other than one pass of the network. Benefited from the proposed region map representation, the region localization and hand keypoints detection could be implemented in a unified framework.

Specifically, we train the proposed network without special considerations. During the inference stage, we first do the network forward pass to obtain the 25-channel output feature maps $\mathbf{H}$. The hand keypoints positions and center of the hand can be obtained through non-maximal suppression (NMS) algorithm [44]. Suppose the center of hand is c, the ratios of width γ_x and height γ_y are computed through

$$\gamma_x = \max(\min(\frac{1}{|N|} \sum_{p \in N(c)} \mathbf{H}_x(p), 1), 0), \tag{6}$$

and

$$\gamma_y = \max(\min(\frac{1}{|N|} \sum_{p \in N(c)} \mathbf{H}_y(p), 1), 0). \tag{7}$$

Here, $\mathbf{H}_x$ and $\mathbf{H}_y$ are the region map of bounding box width ratio and height ratio respectively, N is the set of neighboring pixels in position c, and we define the neighboring pixels in a square region. In our current implementation, the size of the square region is 5×5. $|N|$ is the number of neighboring pixels.

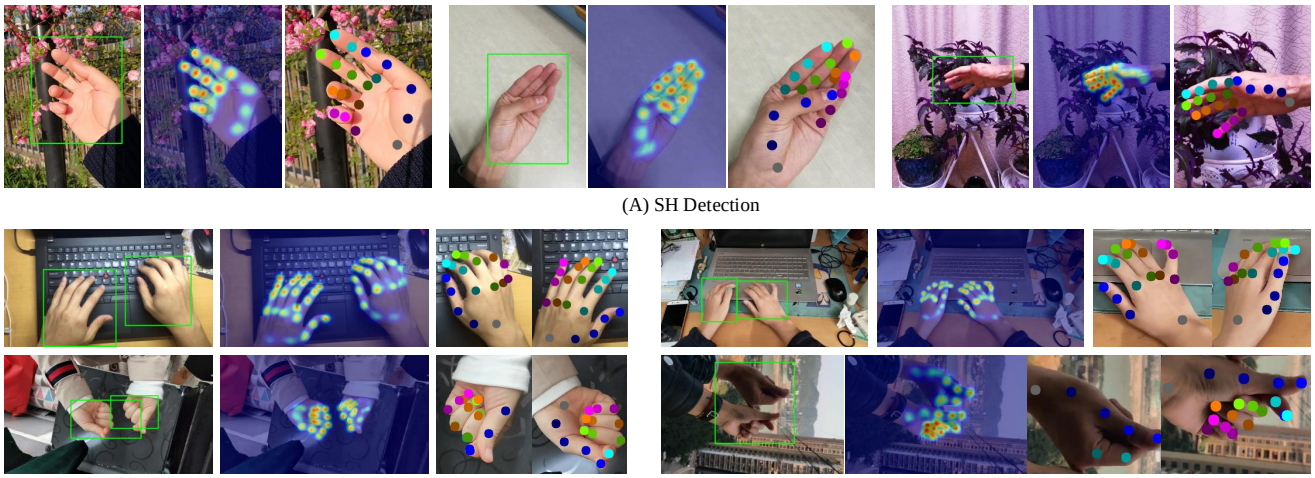
After that, we could extract RoIs of the input image with the rectangle of the width $(\gamma_x * w)$ and height $(\gamma_y * h)$ at the position of c. Then, we redo the network inference by scaling the extracted image content to fit the input size of the network and obtain the final 2D hand pose, which is shown as the red dashed line in Fig. 2. We call this procedure as cycle detection since we use the forward output region maps as the feedback information for 2D hand pose performance boosting. With the benefits of the proposed SRHandNet, we can simultaneously obtain the hand keypoints positions and localize the hand region with one forward network pass. This also gives us an opportunity to perform the cycle detection whether or not. Beyond that, the proposed cycle detection can also have benefits for pose estimation with more than one hand. All the RoIs of image contents can be packed as a batch and it is very fast and convenient to perform the network inference with only one time for a batch of scaled image data.

In our current implementation, the cycle detection is only performed for small hand, where the small hand is defined as $\gamma_x \leq 0.3$ and $\gamma_y \leq 0.3$. It means that the cycle detection is only active when the hand covers the area of input image less than $1/9$. This strategy is only designed for runtime efficiency. In fact, we could always perform the cycle detection (**Full Cycle Detection**) if we do not care about the execution time. We find that full cycle detection could improve the accuracy of 2D hand pose estimation a bit. Sec. 4.3 describes it for more details.

4 Experiments

We conducted several experiments to evaluate the performance of the proposed SRHandNet along with comprehensive ablations on the dataset in [2] and the dataset in [15]. The experiments demonstrate that our method can achieve appealing and high efficient 2D hand pose estimation results compared against the state-of-the-art methods.

Fig. 4. Selected image frames under desktop environment. SRHandNet can run up to 30fps accurate hand keypoints detection and simultaneous hand region localization.



(A) SH Detection

(B) DH Detection

(C) SMH Detection

(D) Synthetic Hand Detection

Fig. 5. Typical results of SRHandNet. We show some typical results including the examples with single hand (SH) detection, double hand (DH) detection, small hand (SMH) detection as well as the synthetic hand detection. Our method can correctly detect the 2D hand keypoints from color images. For each instance, we show the hand bounding box, the hand keypoints heatmap and the enlarged 2D hand pose, which is computed by NMS.

4.1 Implementation Details

We implemented the proposed SRHandNet with Caffe2 [45]. In this subsection, the most important implementation details are described. Other remaining details could be found on the project website, where the source code and trained model are both available on **www.yangangwang.com**.

Training. To train the proposed network, we used the dataset provided in [15], where there are 10K images for training and about 1.5K images for validation. All the images were reshaped into the size of 352×352, padding with $(128, 128, 128)$ if necessary. Meanwhile, the hand keypoints positions were also rescaled according to the scale factor of each training image. To guarantee the performance of training, we performed the data augmentation on the fly. The pixel values were randomly adjusted by gamma correction

TABLE 1
Hand keypoints estimation results (PCK@0.2 score) on the validation dataset with three categories. We compare our method with/without the cycle detection. Moreover, the recent OpenPose [2] and Mask-rcnn [5] are both reported. Similar as bounding box estimation, we train all of them on the same training dataset.

Model	Time	Single Hand (SH)		Double Hand (DH)		Small Hand (SMH)	
		DIP	mean	DIP	mean	DIP	mean
OpenPose [2] *(downloaded)*	70ms	0.60	0.58	0.59	0.60	0.02	0.02
OpenPose [2] *(trained)*	70ms	0.81	0.76	0.64	0.65	0.02	0.02
Mask-rcnn [5] *(trained)*	140ms	0.87	0.82	0.78	0.79	0.27	0.25
Ours (without cycle detection)	**15ms**	0.86	0.85	0.68	0.70	0.33	0.34
Ours (with cycle detection)	**30ms**	**0.94**	**0.94**	**0.82**	**0.84**	**0.52**	**0.55**

and Gaussian blur. After that, the pixel values were all subtracted by 128 and normalized with 256. We also randomly did the homography by adjusting the 4-corners of the input image. Other standard data augmentation techniques such as random rotation, random scaling, random cropping and random horizontal flipping were also utilized.

We used stochastic gradient decent to optimize the training loss. The mini-batch was set to 15. Momentum was set to 0.9. The learning rate was set to 2×10^{-5} and fixed for the whole training stage. The training iterations were set to 300K. The parameters of our network were randomly initialized with no pre-training on any other external datasets. The training loss reduced from about 6.0K to the final 0.25K. We trained our network on a single GTX1080Ti GPU, which costed about 2 hours per 10K iterations.

Inference. During inference, the non-maximal suppression (NMS) was used to obtain the hand keypoints and hand center positions within a 5×5 square region on the last feature maps of our network. The output feature maps had the size of 88×88. A threshold τ was set for the confidence, which means the detected hand keypoints and center positions are regarded as good detection only when they are larger than the threshold τ. In our current implementation, $\tau = 0.2$. The final hand keypoints positions were then multiplied by 4.

All the testing images were resized into 352×352 and padded with $(128, 128, 128)$ if necessary. Similar as image processing in the training stage, the pixel values of testing images were also subtracted by 128 and normalized with 256. Currently, we have tested our method under desktop environment with one GTX1080Ti GPU and our implementation can run at 40**fps** for hand bounding box regression and more than **20**fps for 2D hand pose estimation with full cycle detection. Typically, since small hand is less common in the close shot desktop applications, we can perform cycle detection only for small hand and our system can run up to **30**fps accurate 2D hand pose estimation in this scenario. It is noted that our design is not optimized for speed and better speed/accuracy would be achieved, *e.g.*, by varying the image sizes, performing the NMS with CUDA and *etc*..

4.2 Main Results

The proposed SRHandNet can run up to 30fps accurate hand keypoints detection and simultaneous region localization under desktop environment with a single GPU. Some of the selected color image frames from supplemental videos are shown in Fig. 4. The detected hand region (visualized as green rectangle) and 2D hand keypoints are both presented. Furthermore, we also evaluated our method on the validation dataset, which was collected from [2] and [15]. The evaluation dataset contains 1300 images of different hand gestures from in-the-wild real color images. To investigate the performance of different methods for different types of hand poses, we divided the evaluation dataset into three categories, including 500 images single hand (SH), 500 images double hand (DH) and 300 images small hand (SMH). The SH and DH are the hands captured in a close shot, while SMH contain the hands which cover very small area in the color image space. In particular, we show the hand pose estimation results with different lighting, cluttered background, the occlusions, similar color variations, as well as the indoor and outdoor environment. Our method can perform good results for these cases. Some typical results are shown in Fig. 5, where the detected hand bounding box, the heatmaps of hand keypoints as well as the obtained 2D hand keypoints by NMS are visualized. All of them are overlaid on the original color images. Notably, there are numerous synthetic hands in [2]. We tested our method on these hands although we did not train on the synthetic ones. Our method also works well for the 2D synthetic hand pose estimation as shown in the last row of Fig. 5. The whole evaluation dataset is also available on the project website.

4.3 Evaluations

To make a quantitative comparison, we report the standard Averaged Precision (AP) [46] for bounding box regression and Percentage of Correct Keypoint (PCK) for 2D hand pose [47]. Specifically, AP is the average over all the bounding box IoU, and we computed AP_{50} for the threshold 0.5. The PCK is a measurement between the predicted keypoint location and the groundtruth keypoint location within a threshold σ. For the k-th hand keypoint p_k, we denote it by PCK_σ^k and approximate it on the validation dataset $\mathbb{D}$ as

$$\text{PCK}_\sigma^k = \frac{1}{||\mathbb{D}||} \sum_{\mathbb{D}} \mathbf{1}\left(\frac{||p_k^{pt} - p_k^{gd}||_2}{\max(w, h)} \leq \sigma \right), \tag{8}$$

where p_k^{pt} is the predicted position and p_k^{gd} is its groundtruth location, $\mathbf{1}(\cdot)$ is the indicator function, w and h are the width and height of the groundtruth bounding box, respectively. It is noted that σ is measured as a normalized distance in our implementation, *i.e.*, pixel distances for each example are normalized by the groundtruth bounding box size. For

TABLE 2
Bounding box estimation results on the validation dataset with three categories. We compare our method with recent successful object detection method, *i.e.*, Faster-rcnn [4] and Mask-rcnn [5]. All of them are trained on the same training dataset and our method can high efficiently obtain the comparable bounding box results.

	SH/$_{AP_{50}}$	DH/$_{AP_{50}}$	SMH/$_{AP_{50}}$
Faster-rcnn [4]	67.71	33.84	6.79
Mask-rcnn [5]	80	35.05	13.24
Ours	**89.8**	**43.5**	**27.37**

fair comparison, all the networks were trained on the same training dataset. One exception is that we also downloaded the model given by OpenPose [2] for more detailed comparison. Tab. 1 shows some quantitative results in three evaluation datasets. We report the selected PCK values with $\sigma = 0.2$ for DIP fingers (*i.e.*, the 7-th, 11-th, 15-th and 19-th keypoints) and the averaged mean PCK values for all fingers. Our method has the superior results, both on time efficiency and accuracy, compared with the relevant methods.

Region detection. For validating the proposed region detection strategy, we compared our bounding box representation with the recent successful object region detection methods, *i.e.*, Faster-rcnn [4] and Mask-rcnn [5]. We downloaded the code from the website[1] and trained the model with the same training dataset. Since Mask-rcnn needs the hand mask to train, we obtained the hand mask by handcraft. Tab. 2 shows the comparison results. From the table, we could find that our method has superior performance in all the three categories of evaluation datasets, compared with the other two methods. The big improvements may come from our special designed on-the-fly data augmentation strategies. In particular, we found that the data augmentation of scaling with tiny factors as well as performing the homography is very important for boosting the hand pose training, which is beyond the scope of this paper. Fig. 6 visualizes a few selected frames from the supplemental video. From this figure, we can find that the proposed SRHandNet can obtain comparable results but need only 20ms per-frame for hand pose estimation. Overall, the experiment demonstrates that the proposed region map is effective for hand bounding box representation. And more importantly, it is lightweight.

Cycle detection. In order to validate the effectiveness of the proposed cycle detection, we compared the hand keypoints estimation results with/without the cycle detection and the comparison results are shown in Tab. 1. From the table, we can find that for all the three evaluation datasets, cycle detection can all improve the PCK values. Specifically, for small hands, the PCK values are improved more than 0.2 compared with the non-cycle detection strategy. For the single hands and double hands, performing the cycle detection could improve the accuracy of pose estimation from 0.85 to 0.95 and 0.70 to 0.84, respectively. The improvements are not significant as small hands. We checked the estimation results and found that the main reason might come from the close shot hand images, which are not sensitive to the hand

1. https://github.com/facebookresearch/Detectron

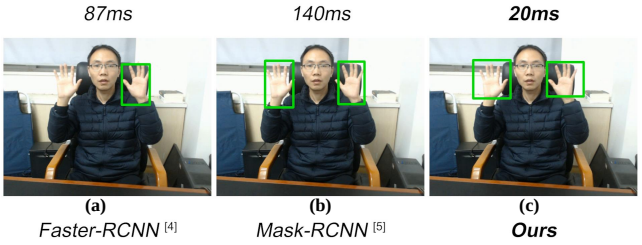


Fig. 6. Comparison result of region detection. We compared SRHandNet with recent Faster-rcnn [4] and Mask-rcnn [5]. Our SRHandNet can obtain comparable results but is more time efficient. (a) is the bounding box detection result with Faster-rcnn; (b) is the result with Mask-rcnn and (c) is result of our method.

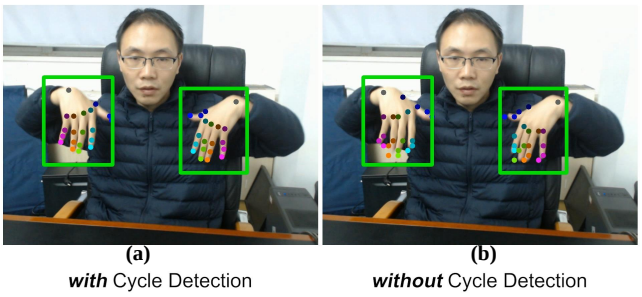


Fig. 7. Detection result with / without cycle detection. With cycle detection, we can obtain more accurate hand keypoints estimation results.

bounding box estimations. We also performed real video with our method and two selected frames are shown in Fig. 7. From this figure, we can clearly find that the locations of hand keypoints are more accurate in the green rectangles of Fig. 7(a). That is, with cycle detection, we can obtain better hand keypoints estimation results. In summary, the cycle detection is more important for hand pose estimation when the size of hand is regular or small. Since our network has the ability to estimate the bounding box of the hand for the input color images, it is easier to adopt the cycle detection whether or not on the fly, thus to improve the runtime efficiency.

Comparison with state-of-the-art. Beyond that, we compared our method with the recent OpenPose [2] and Mask-rcnn [5], where the later one can be also used for pose estimation as described by the authors. Mask-rcnn was

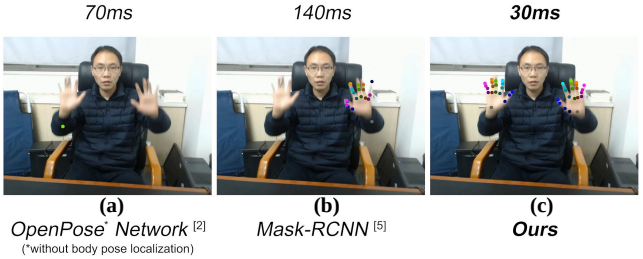


Fig. 8. Comparison result with state-of-the-art. SRHandNet can obtain faster and more accurate results compared with OpenPose [2] and Mask-rcnn [5]. (a) is the result with the OpenPose network trained by our dataset; (b) is the detection result with Mask-rcnn and (c) is the result of our method.

TABLE 3
Hand keypoints estimation comparison results (average mean PCK@0.2 score) on several existing hand datasets.

	GANH [49]	STB [50]	RHD [48]
OpenPose [2]	0.26	0.37	0.26
Mask-rcnn [5]	0.37	0.47	0.35
Model in [48]	0.22	0.32	**0.80**
Ours	**0.41**	**0.53**	0.46

trained by the given code on our training dataset. For training OpenPose, we implemented the network by ourselves through referring to the given network in [2]. Moreover, we also downloaded the trained OpenPose model for thorough study. Fig. 8 shows the selected frames from supplemental video and the proposed SRHandNet can obtain faster and more accurate results. We also report the quantitative comparison results on the evaluation dataset as described in Tab. 1. From this table, we can find that our method has more than $0.05 \sim 0.1$ and $0.1 \sim 0.2$ overall PCK values compared with Mask-rcnn and OpenPose, respectively. It is noted that OpenPose fails to the pose estimation for small hands. The reason might come from the designed large output feature maps, which can not capture the information for small hands. All the experiments demonstrate the superior performance of our method. Furthermore, we compared our method on more existing datasets and the results are shown in Tab. 3, where the PCK values are reported. All the experiments demonstrate the superior performance of our method, except for [48] on their own dataset, which might be over-trained.

4.4 Discussions and Future work

In Sec. 4.2 and Sec. 4.3, we show the experimental results of our SRHandNet over real-time desktop applications, and we also compare with the recent methods both on the region detection and hand pose estimation. Our method is an adequate solution for hand pose estimation under lightweight scenario with single RGB input. Applications, such as Human machine interaction with mobile camera, integration with voice-driven teleoperation and *etc*., might be flourished by the proposed SRHandNet. Although our method is not restricted into single hand pose estimation, the hand keypoints estimation for more than two hands might be confused with the neighboring joint connections and it can be further improved in the future work.

5 Conclusion

In this paper, we present a novel method for real-time 2D hand pose estimation from monocular color images. In order to efficiently achieve the accurate 2D hand pose in real-time, we regress the hand region and keypoints simultaneously for a given color image and propose a novel lightweight bounding box representation, which is named as Region Map. The detected hand regions are taken as feedback information for boosting the performance of hand keypoints estimation with a single encoder-decoder network. Our method can run at 40fps for hand bounding box detection and up to 30fps for hand keypoints estimation under ordinary desktop environment, which competes all recent methods. We believe that the proposed SRHandNet is one step closer to practical applications with hand pose estimation from monocular color images, such as virtual typing, interaction with mobile cameras and *etc*.

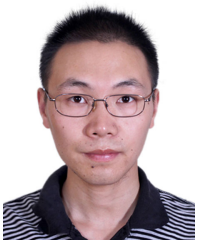
Acknowledgments

We thank the anonymous reviewers to improve this paper. We also thank Wenlin Zhuang and Shiyu Zhao to perform the testing of the proposed real-time desktop system.

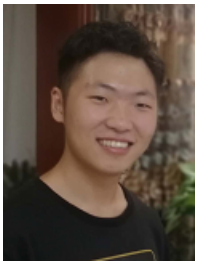
References

[1] 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 *CVPR*, 2017.
[2] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.
[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.
[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.
[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017.
[6] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *ECCV*, 2018.
[7] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *CVIU*, vol. 108, no. 1, pp. 52–73, 2007.
[8] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011.
[9] D. Tang, H. Jin Chang, A. Tejani, and T.-K. Kim, "Latent regression forest: Structured estimation of 3d articulated hand posture," in *CVPR*, 2014.
[10] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *TOG*, vol. 33, no. 5, p. 169, 2014.
[11] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *CVPR*, 2015.
[12] 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 *CVPR*, 2017.
[13] C. Wan, A. Yao, and L. Van Gool, "Hand pose estimation from local surface normals," in *ECCV*, 2016.
[14] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *CVPR*, 2014.
[15] Y. Wang, C. Peng, and Y. Liu, "Mask-pose cascaded cnn for 2d hand pose estimation from single color images," *TCSVT*, vol. 29, no. 11, pp. 3258 – 3268, 2019.
[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *CVPR*, 2017.
[20] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.
[21] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016.
[22] P. Panteleris, I. Oikonomidis, and A. Argyros, "Using a single rgb frame for real time 3d hand pose estimation in the wild," in *WACV*, 2018.
[23] U. Iqbal, P. Molchanov, T. Breuel, J. Gall, and J. Kautz, "Hand pose estimation via latent 2.5d heatmap regression," in *ECCV*, 2018.
[24] Y. Cai, L. Ge, J. Cai, and J. Yuan, "Weakly-supervised 3d hand pose estimation from monocular rgb images," in *ECCV*, 2018.

[25] K. Saini, M. Dewal, and M. Rohit, "A fast region-based active contour model for boundary detection of echocardiographic images," *Journal of digital imaging*, vol. 25, no. 2, pp. 271–278, 2012.
[26] M. Dewal, K. Saini, and M. Rohit, "Assessment of mitral regurgitation severity with intensity based region growing," *International Journal of Hybrid Information Technology*, vol. 8, no. 6, pp. 45–56, 2015.
[27] K. Saini and M. Dewal, M. L.and Rohit, "Automatic jet area detection during mitral regurgitation," *International Review on Computers and Software*, vol. 7, no. 6, pp. 2891–2898, 2012.
[28] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 607–12 616.
[29] W. Li, C. W. Pan, R. Zhang, J. P. Ren, Y. X. Ma, J. Fang, F. L. Yan, Q. C. Geng, X. Y. Huang, H. J. Gong, W. W. Xu, G. P. Wang, D. Manocha, and R. G. Yang, "Aads: Augmented autonomous driving simulation using data-driven algorithms," *Science Robotics*, vol. 4, no. 28, 2019.
[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
[31] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
[33] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
[34] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.
[37] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *CVPR*, 2016.
[38] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *ICCV*, 2015.
[39] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015.
[40] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *PAMI*, no. 12, pp. 2481–2495, 2017.
[41] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *ECCV*, 2018.
[42] P. Hu and D. Ramanan, "Finding tiny faces," in *CVPR*, 2017.
[43] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection." in *CVPR*, 2017.
[44] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *ICPR*, 2006.
[45] "Github reposory for caffe2." https://github.com/pytorch/pytorch, 2018, [accessed 1-Nov-2018].
[46] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *CVPR*, 2016.
[47] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *PAMI*, vol. 35, no. 12, pp. 2878–2890, 2013.
[48] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *ICCV*, 2017.
[49] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular rgb," in *CVPR*, 2018. [Online]. Available: https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/
[50] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang, "A hand pose tracking benchmark from stereo matching," in *ICIP*, 2017.



Yangang Wang received his B.E. degree from Southeast University, Nanjing, China, in 2009 and his Ph.D. degree in control theory and technology from Tsinghua University, Beijing, China, in 2014. He was an associate researcher at Microsoft Research Asia from 2014 to 2017. He is currently an associate professor at Southeast University. His research interests include image processing, computer vision, computer graphics, motion capture and animation.



Baowen Zhang is currently a fourth year undergraduate student at Southeast University. This work was done when he joined in the student research training program of Southeast University. His research interests include computer vision, deep learning and 2D hand pose estimation.

Cong Peng received her B.S. degree from Southeast University, Nanjing, China, in 2010, and the Ph.D. degree in electrical engineering from Beihang University, Beijing, China, in 2016. She is currently a professor with the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her research interests include active vibration control, vibration measurement and computer vision.