

# FastHand: Fast monocular hand pose estimation on embedded systems

Shan An<sup>a</sup>, Xiajie Zhang<sup>b</sup>, Dong Wei<sup>b</sup>, Haogang Zhu<sup>a,\*</sup>, Jianyu Yang<sup>c</sup>, Konstantinos A. Tsintotas<sup>d</sup>

<sup>a</sup> School of Computer Science and Engineering, Beihang University, Beijing, 100191, China

<sup>b</sup> Tech. and Data Center, JD.COM Inc, Beijing, 100108, China

<sup>c</sup> School of Rail Transportation, Soochow University, Suzhou, 215006, China

<sup>d</sup> Department of Production and Management Engineering, Democritus University of Thrace, Xanthi, 67132, Greece

## ARTICLE INFO

### Keywords:

Hand pose estimation  
Landmark localization  
Hand detection  
Encoder–decoder network  
Heatmap regression

## ABSTRACT

Hand pose estimation is a fundamental task in many human–robot interaction-related applications. However, previous approaches suffer from unsatisfying hand landmark predictions in real-world scenes and high computation burden. This paper proposes a fast and accurate framework for hand pose estimation, dubbed as “FastHand”. Using a lightweight encoder–decoder network architecture, FastHand fulfills the requirements of practical applications running on embedded devices. The encoder consists of deep layers with a small number of parameters, while the decoder uses spatial location information to obtain more accurate results. The evaluation took place on two publicly available datasets demonstrating the improved performance of the proposed pipeline compared to other state-of-the-art approaches. FastHand offers high accuracy scores while reaching a speed of 25 frames per second on an NVIDIA Jetson TX2 graphics processing unit.

## 1. Introduction

Gesture recognition is an essential component needed for a robust Human–Robot Interaction (HRI) [1–6]. Since robot malfunctions may occur due to the noisy sensor measurements or the operator’s speech disabilities, an alternative technique has to be adopted to allow the agent to understand the human’s intentions and, consequently, react accordingly [7]. However, gesture recognition is a challenging problem in the robotics community as the hand is defined through a small area compared to the human body. Moreover, it exhibits a high degree of freedom and similarity, i.e., among the finger joints’ visual appearance.

The timing required for a robot to interact with its operator constitutes another considerable challenge in HRI since real-time responses are necessary for a reliable application. Skeleton-based methods [8–11], generally follow two processing steps. At first, they detect hand landmarks, a task widely known as hand pose estimation, and subsequently, they recognize gestures according to these points. Some approaches make use of depth cameras to locate the 3D hand landmarks [12–16]. Nevertheless, these modules are expensive for small and low-cost platforms. Therefore, it is urgent to study hand pose estimation using a monocular camera as the primary sensing modality.

Based on the technique they use to tackle the problem, modern methods can be divided into top-down and bottom-up ones. The former firstly detect the hand and afterward locate its landmarks. On the contrary, bottom-up frameworks initially detect the landmarks, and next, they cluster them to form a hand. Both strategies present advantages

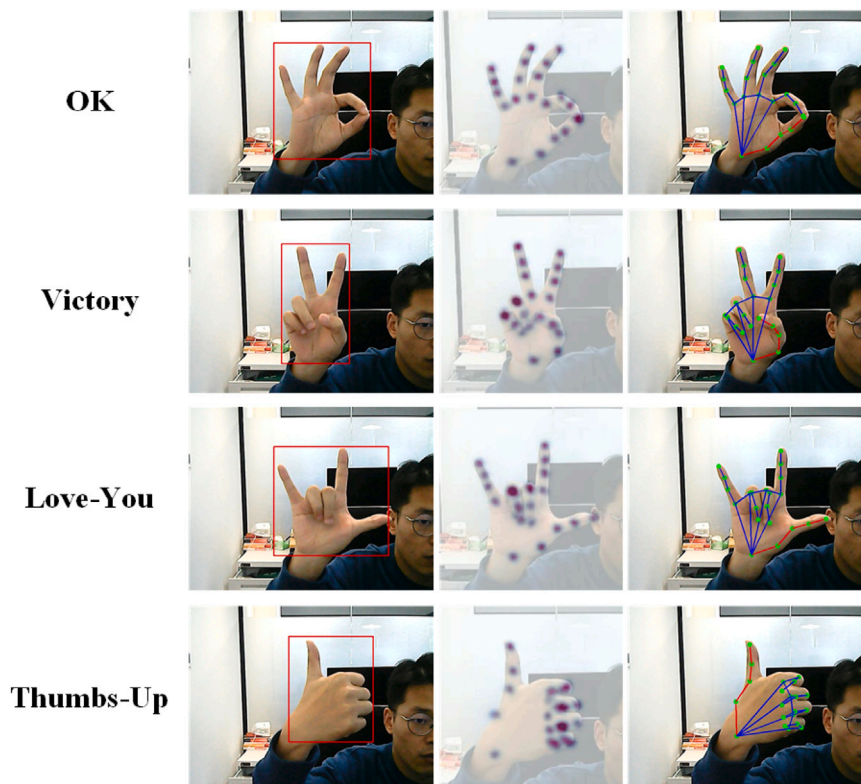
and disadvantages. The top-down methods repeatedly run on each hand bounding box to localize landmarks, resulting in the linearly increased run-time. The bottom-up approaches detect landmarks from the whole image but have a low recall rate because the complex background causes false positives.

Furthermore, hand pose estimation pipelines are divided into two categories according to their output. Methods that utilize coordinate regression are placed in the first [17–19], whereas algorithms that extract heatmaps belong to the second [20–25]. More specifically, approaches that fall into the first category directly predict the landmark coordinates from images. However, this is not accurate and has limitations when encountering occlusions and large poses. Frameworks belonging to the second category infer each hand joint’s likelihood heatmap and then generate their coordinates. These techniques offer high accuracy scores; nevertheless, they present the disadvantage of a slower execution time, making them impracticable for resource-constrained mobile platforms. This is mainly due to their used network structure since they do not follow a lightweight architecture.

Considering the reasons above, we propose a lightweight and top-down pose estimation technique, dubbed “FastHand”, in this paper. Our method first detects the hand using a monocular camera, then crops the Region Of Interest (ROI), and then localizes its 2D landmarks based on a heatmap regression (Fig. 1). An encoder–decoder network design is employed. Regarding the encoder, a low amount of

\* Corresponding author.

E-mail address: [haogangzhu@buaa.edu.cn](mailto:haogangzhu@buaa.edu.cn) (H. Zhu).



**Fig. 1.** An illustrative example of the proposed hand pose estimation pipeline. The first column shows the input frames where the detected hand bounding-boxes are indicated. The second column depicts the predicted heatmaps painted in the cropped image, while in the third column, the predicted 2D coordinates for the 21 hand landmarks are demonstrated.

parameters are maintained while learning high-level information by using  $1 \times 1$  convolution, skip connection [26], and depthwise separable convolutions [27]. Concerning the decoder, an improvement in feature map resolution is achieved through a deconvolution process, which preserves the high resolution of the hand and avoids the false recognition of the background. Moreover, the heatmap regression accuracy is also improved. Extensive experiments are conducted on two publicly available datasets demonstrating our method's superior performance compared to other state-of-the-art frameworks. The main contributions can be summarized below:

- A novel, lightweight encoder–decoder network, entitled “Fast-Hand”, is proposed for hand pose estimation based on heatmap regression. The proposed framework is able to perform on a mobile robot in real-time.
- A 2D hand landmark dataset is generated based on the Youtube3D Hands image-sequence [28]. The generation code<sup>1</sup> is made available to facilitate future studies.
- Extensive experiments are performed to demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows. Section 2 reviews the related study about the hand pose estimation task. In Section 3, the proposed method is described, while Section 4 evaluates the system's performance and reports the comparative results. Finally, our paper concludes in Section 5.

## 2. Related work

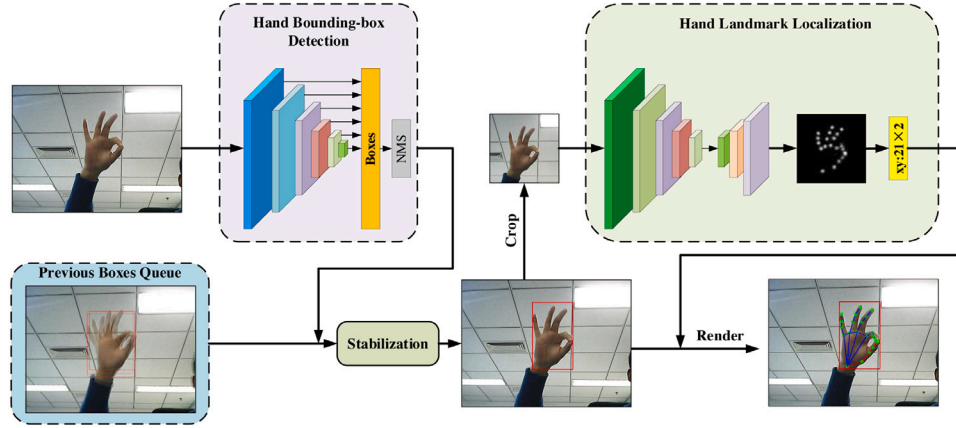
### 2.1. Coordinate regression-based approaches

FreiHAND [17] proposed the first large-scale, multi-view, and 3D annotated real dataset, wherein a Convolutional Neural Network (CNN) was trained [17]. Subsequently, this framework could predict hand poses from the incoming RGB images with a certain ability of generalization. NSRM [18] adopted a cascaded multi-task architecture aiming to jointly learn the hand structure and its landmarks' representation [18]. A novel CNN with a “self-attention” module was proposed in [19]. This network is lightweight and is capable of deployment on embedded systems. A deep regression network, combined with a specific label distribution learning network, is proposed to learn hand landmarks from 3D point cloud [29]. “MediaPipe Hands” [30] first determines the position of hands with a palm detector and then inputs the bounding-box containing hands to the landmark model to regress 21 2.5D hand landmarks directly.

### 2.2. Heatmap regression-based approaches

Methods based on heatmap regression showed an improved performance regarding the final accuracy than the one achieved via the coordinate regression. In [21], a deep network was proposed to estimate the 3D hand pose originated from RGB images. This framework detected the hand's location, and then it cropped the selected region. Subsequently, an encoder–decoder network extracted the 2D score maps for landmarks definition. Finally, 3D hand poses were estimated from these points by another network trained a priori on 3D data. A multi-view training method [25] used Convolutional Pose Machines (CPM) [31] as a detector and achieved similar performance to RGB-D systems. In [32], two stages determine the network's functionality.

<sup>1</sup> <https://github.com/AnshanTJU/Youtube3DHands-2D>.



**Fig. 2.** An overview of the proposed framework. Firstly, through the MobileNet-SSD network, which is based on MobileNetV2 [38] and Single Shot multi-box Detector (SSD) [39], we detect the human hand in the incoming RGB image. Subsequently, aiming to stabilize the Region Of Interest (ROI), this area is combined with the previous visual sensory information. Next, we crop the ROI and feed it into the landmark localization module. The encoder-decoder network is used to predict the landmarks' heatmap and compute its 2D position.

Firstly, the hand mask prediction took place, while the pose prediction followed. For a given RGB image, the hand bounding box and landmarks were obtained simultaneously through a single forward pass of an encoder-decoder network in [33]. Afterward, the hand region was used as feedback to determine whether a cycle detection was performed. The monocular 3D hand tracking performed well even in challenging occlusion scenarios through the synthetic generation of the training data [22]. Cai et al. [23] adopted a weakly-supervised pipeline trained on RGB-D images and predicted 3D joints from the incoming color frames. Utilizing a CNN for depth and heatmap estimation, 2.5D poses are obtained in [24]. Adaptive Graphical Model Network (AGMN) [34] integrated a CNN with a graphical model for accurate 2D estimation. Similarly, a Rotation-invariant Mixed Graph Model Network (R-MGMN) performed 2D gesture estimation through a monocular RGB camera [35]. The authors in [36] proposed the InterHand2.6M dataset and InterNet for cases where two hands are used. Boukhayma1 et al. introduced the first end-to-end deep learning method, which was formed by an encoder-decoder structure placed in series [37]. 2D landmark information was utilized to facilitate the learning of a 3D hand pose. A graph CNN-based method was proposed to estimate 3D hand joints and reconstructed its 3D mesh of the surface in [20]. Our network is based on heatmap regression because of its high accuracy.

### 3. Proposed method

An overview of the proposed pipeline is outlined in Fig. 2. First, the input image is fed into the hand detection network to get the corresponding bounding box. Then, we obtain a stabilized box through the information provided by the previous boxes. Afterward, this ROI is placed on the landmark localization network for computing the heatmaps. Finally, 2D positions of the 21 hand landmarks are extracted.

#### 3.1. Hand detection and stabilization

##### 3.1.1. Hand detection

Top-down approaches firstly detect the ROI, i.e., the hand, in the visual sensory information and then localize its landmarks on the chosen area. In this way, a computational complexity reduction is achieved, which is our main focus. Therefore, we select to utilize the MobileNet-SSD framework to detect human hands, which is based on MobileNetV2 [38] as the network's backbone and Single Shot multi-box Detector (SSD) [39] for the detection.

##### 3.1.2. Stabilization of the bounding-box

As we adopt the top-down approach, the detection box's inevitable jitters occur, affecting the landmarks' localization. Aiming for more accurate prediction scores, we added a fast and straightforward method to stabilize the hand bounding box. Since early perceived frames present a higher correlation with the current box, we perform a weighted average according to the exponentially decreasing weight. This way, we exploit the correlation between the current box and the previous ones as the weight's value is higher for neighboring images. The current bounding-box coordinates  $P_{cur}$  are calculated as:

$$P_{cur} = \sum_{k=0}^n P_k \times \frac{e^{-k}}{\sum_{j=0}^n e^{-j}}, \quad (1)$$

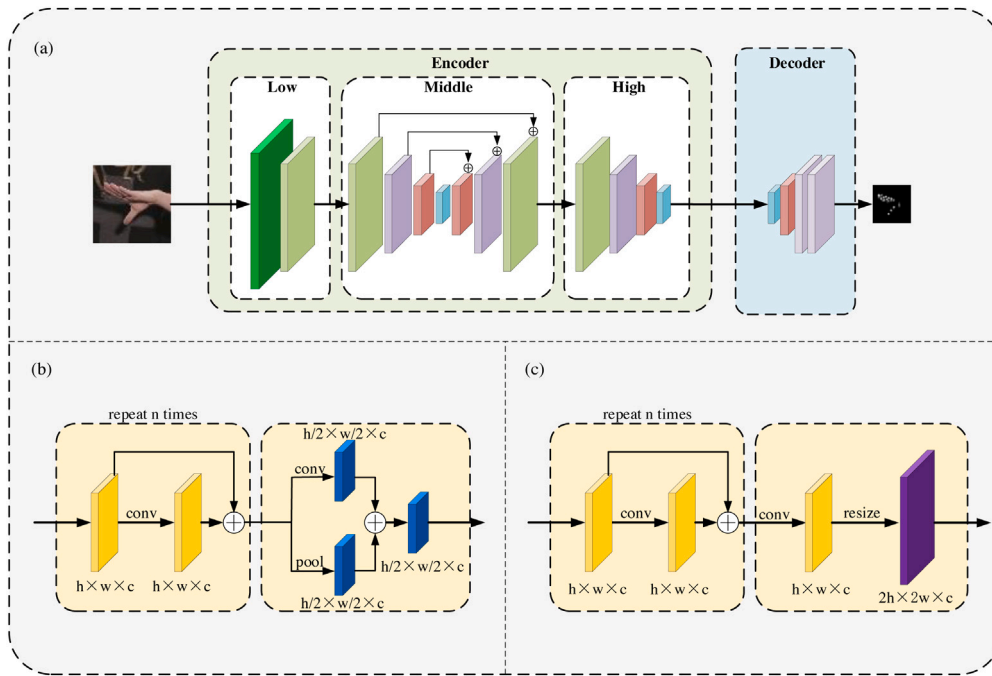
where  $k$  represents the previous frame's index (e.g., 0 denotes the current frame, while 1 is the previous one) and  $n$  corresponds to the multitude of the examined frames. This value was set heuristic to 6 ( $n = 6$ ).  $P_k$  are the coordinates of the  $k$  bounding-box. Finally, the last term of Eq. (1), i.e.,  $\frac{e^{-k}}{\sum_{j=0}^n e^{-j}}$ , is the exponentially decreasing weight.

#### 3.2. Landmark localization

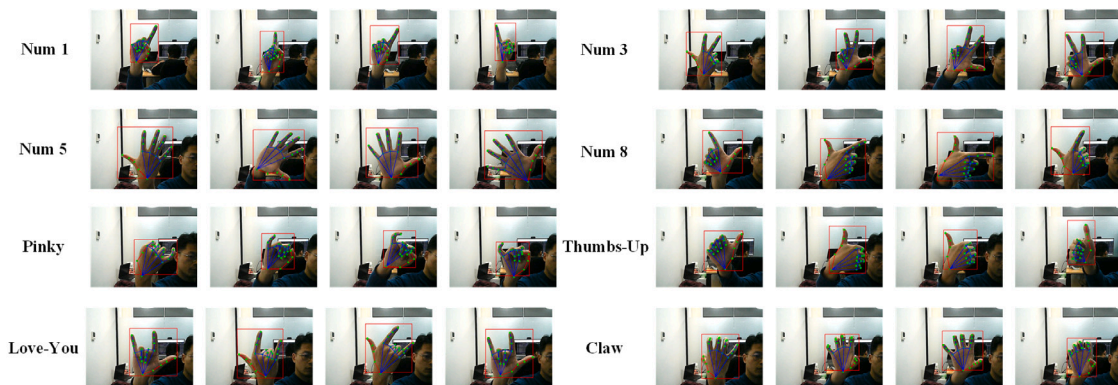
For the landmarks' detection, we design a heatmap regression method following the encoder-decoder architecture outlined in Fig. 3. At first, resize the image obtained from the previous steps to  $256 \times 256$ , and then feed it to the proposed network to extract high-level information. The feature map resolution is restored to  $64 \times 64$  through the decoder to generate the heatmaps, which are processed to extract the 2D coordinates of the landmarks.

##### 3.2.1. The encoder network

The encoder network utilizes  $1 \times 1$  convolution, skip connection [26], and depthwise separable convolutions [27], since we aim to a lightweight solution. The  $1 \times 1$  convolution appears where channel compression is required while maintaining the same resolution. Almost all convolution layers are replaced by depthwise separable convolutions. The network structure follows the design of "MediaPipe Hands" [30]. It preserves a small number of parameters while learning more high-level information. Three parts define its structure, namely the Low, the Middle, and the High (Fig. 3(a)). More specifically, low-level features are extracted via the first part, composed of repeated residual blocks ( $\times 4$ ) and downsampling operation. In the Middle part, we make full use of the network parameters through the encoder-decoder network and skip connection. The model's low and high-level



**Fig. 3.** The illustration of the proposed landmark localization network. The encoder–decoder structure is shown in division (a). The encoder network has three parts (the Low, the Middle, and the High), which encodes the high-level information into low-resolution feature maps using a small number of parameters and computations. In the decoder, the three deconvolution operations improve the output resolution. The components in (b) are the downsampling operation between all feature maps with reduced resolution in the encoder. The convolutions performed in (c) are the upsampling operations of the encoder’s middle part to improve the resolution.



**Fig. 4.** Visualization results obtained by the proposed FastHand. Different hand poses on real-world scenarios are tested.

features can be learned without losing information, which makes the model training faster and less prone to overfitting. In the encoder, a downsampling operation is utilized to reduce the resolution with residual blocks repeated four times, as illustrated in Fig. 3(b). Then, in the decoder part, the resolution can be improved through the upsampling module, which consists of residual blocks, a convolution, and a resize operation by interpolation, as shown in Fig. 3(c). At last, in the High part, more high-level information is extracted while the downsampling module reduces the resolution, as shown in Fig. 3(b).

### 3.2.2. The decoder network

The decoder network aims to restore the resolution of the feature map. Since the human hand presents a different structure from the face or other human parts, presenting higher degrees of freedom, we improve the feature map resolution through a deconvolution ( $\times 3$ ) performed to preserve spatial position information. This technique offers a better modeling capability than a direct, coordinate regression

and can provide improved accuracy. Furthermore, when heatmaps are used as ground truth to predict the landmarks, they must have high resolution. In our method, the encoder’s output is  $8 \times 8$ . As a result, we use the decoder to increase the resolution to  $64 \times 64$ . This way, we manage to retain more position-related information. Three layers of deconvolutions are used to restore the feature map size. As a result, the proposed pipeline outputs  $64 \times 64 \times 21$  heatmaps. Finally, the 2D coordinates of the 21 hand landmarks are derived by extracting the peak points at each channel.

## 4. Experiments

In this section, extensive experiments are conducted to demonstrate the effectiveness of FastHand. Due to the lack of the 2D hand pose dataset, we generate a large-scale dataset using image-sequences from YouTube3D Hands [28]. Two widely-used datasets, STB [40] and RHD [21], are selected for the method’s evaluation. Subsequently,

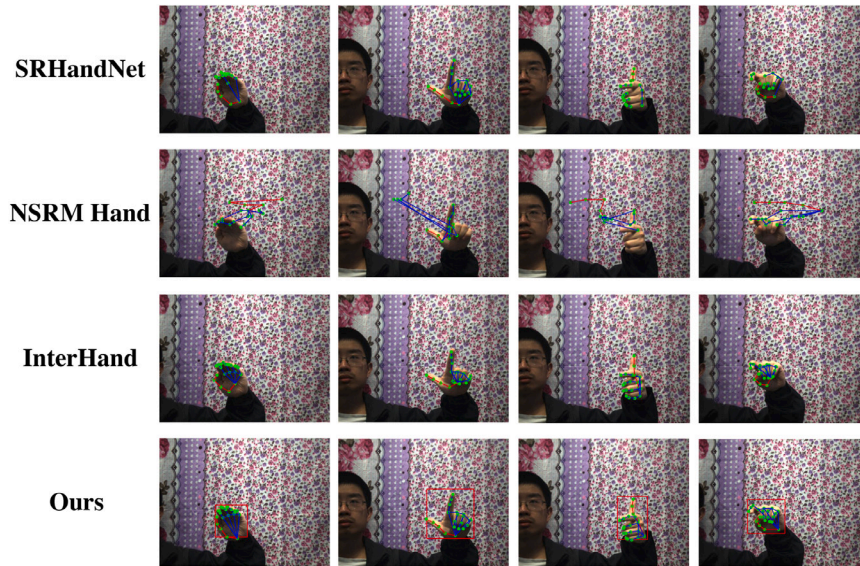


Fig. 5. Qualitative results of the proposed method compared with other approaches on the STB dataset [40].



Fig. 6. Qualitative results of the proposed method compared with other approaches on the RHD dataset [21].

we introduce the experimental settings, which include the training parameters and evaluation metrics. Finally, we provide quantitative and qualitative results to fully demonstrate the overall performance of our method.

#### 4.1. 2D Hand pose dataset generation

There are several limitations of the existed 2D hand pose dataset. The datasets, such as STB [40], MHP [41], HO3D [42], only have

limited scenes and hand movements. On the other hand, synthetic datasets are always unreal. It is difficult to collect and annotate a 2D hand pose because it has high degrees of freedom, self-occlusion, and mutual occlusion.

We generate a 2D hand pose dataset based on the YouTube3D Hands [28] dataset, which is a large-scale image-sequence including hand actions. It provides 109 videos in total. 102 of them are the training sets, while the rest are the test ones. As this dataset offers the 3D

**Table 1**  
Properties of the used datasets.

Dataset	Usage	Description	Image resolution	# Joints	# Images
YouTube2D Hands	Training	Real-world	256 × 256	21	47 125
GANeratedHands [22]	Training	Synthetic	256 × 256	21	141 449
Test set of STB [40]	Test	Real-world	640 × 480	21	6000
Test set of RHD [21]	Test	Synthetic	320 × 320	21	2727

vertex coordinates of hands, we need to extract the 2D information of the hand landmarks. Through the average of 10 vertices around each hand skeleton, we get the 3D coordinates of the 21 hand landmarks. Subsequently, we project them onto the 2D plane to obtain the 2D points. The generated 2D hand pose dataset is referred to as YouTube2D Hands. As a final note, the images are resized to 256 × 256 and augmented by 10 times using random scaling and cropping. Besides, we use the GANeratedHands dataset [22] for the network’s training since, through the utilization of real-world images, we cannot cover all gestures, and the annotation in real data is always inaccurate. A brief description of each dataset is shown in Table 1.

## 4.2. Experimental settings

### 4.2.1. Training process

The training are conducted on four NVIDIA Tesla P40 GPU. The entire network was trained for a total of 50 epochs, while we decreased the learning rate exponentially after the first three ones. The initial learning rate was set to  $10^{-2}$  and the batch size was set to 64. The Adam optimizer [43] was utilized as well. Finally, the network’s implementation was based on TensorFlow [44].

### 4.2.2. Evaluation metrics

Sum of Squares Error (SSE), end-point error (EPE), and Probability of Correct Keypoint (PCK) within a normalized distance threshold, are used as evaluation metrics:

$$SSE = \frac{\sum_{s=1}^D (\sum_{i=1}^{21} ((\frac{y_{si} - \hat{y}_{si}}{\max(w,h)})^2))}{D}, \quad (2)$$

$$EPE = \frac{\sum_{s=1}^D (\sum_{i=1}^{21} \left\| \frac{(y_{si} - y_{s0}) - (\hat{y}_{si} - \hat{y}_{s0})}{\max(w,h)} \right\|)}{21 \times D}, \quad (3)$$

$$PCK_{\sigma}^i = \frac{1}{D} \sum_{s=1}^D \mathbb{1} \left( \frac{\|y_{si} - \hat{y}_{si}\|}{\max(w,h)} \leq \sigma \right), \quad (4)$$

$$PCK_{\sigma} = \frac{\sum_{i=1}^{21} PCK_{\sigma}^i}{21}, \quad (5)$$

where the term  $y_{si}$  is the ground truth of landmark  $i$ , and  $\hat{y}_{si}$  is the system’s predicted coordinate.  $i$  represents the landmark indexes  $i = [1, \dots, 21]$  and  $s$  represents the index of hand sample.  $y_{s0}$  and  $\hat{y}_{s0}$  are the ground truth root joint and the predicted root joint.  $D$  denotes the number of samples in the dataset, while  $w$  and  $h$  are the original images’ width and height, respectively. In Eq. (4),  $\mathbb{1}(\cdot)$  is the indicator function and  $\sigma$  is the threshold. If  $L2$  distance between the predicted landmark and the ground truth is less than  $\sigma$ , the indicator function is set to 1, otherwise to 0.  $PCK_{\sigma}^i$  represents the  $PCK_{\sigma}$  metric of landmark  $i$  at the ratio of  $\sigma$ , which is an empirical value of 0.2.

### 4.3. Comparison with the state-of-the-art techniques

We compare the proposed pipeline against four state-of-the-art methods, namely SRHandNet [33], NSRM Hand [18], InterHand [36], and MediaPipe Hands [30]. The first three methods have open source code, which we ported to embedded devices to test their speed. As shown in Table 2, our FastHand ranks first in most metrics on the STB and RHD datasets. The proposed approach achieves a  $SSE$  score of

**Table 2**

Comparative results of the baseline methods against the proposed method. Green denotes the best, while blue is second best.  $\uparrow$  means the higher the better, while  $\downarrow$  means the lower the better.

Dataset	Metric	SRHandNet [33]	NSRM hand [18]	MediaPipe hands [30]	Inter-Hand [36]	Our method
STB [40]	$SSE \downarrow$	–	0.7078	0.9435	0.4853	0.3490
	$EPE \downarrow$	–	0.1326	0.1522	0.1302	0.1317
	$PCK@0.2 \uparrow$	0.8526	0.7246	0.7032	0.8245	0.8948
RHD [21]	$SSE \downarrow$	–	2.5613	1.9929	2.0413	0.6368
	$EPE \downarrow$	–	0.1953	0.2133	0.2630	0.0986
	$PCK@0.2 \uparrow$	0.5317	0.7177	0.6927	0.3910	0.8661

**Table 3**

Comparison in inference runtime (Frames Per Second — FPS) of different pose estimation methods.

Device	SRHand-Net [33]	NSRM hand [18]	InterHand [36]	Our method
NVIDIA GeForce 940MX GPU	21.06 FPS	2.42 FPS	14.24 FPS	31.33 FPS
NVIDIA Jetson TX2 GPU	19.16 FPS	3.65 FPS	7.77 FPS	25.05 FPS

**Table 4**

Comparison in terms of model size of different pose estimation methods.

Model	SRHand-Net [33]	NSRM hand [18]	MediaPipe hands [30]	InterHand [36]	Our method
Size (MB)	71.90	139.7	3.9	541.7	13.0

0.3490 on the STB dataset, surpassing the second-best (0.4853). For the  $EPE$  metric on the STB dataset, our method is slightly higher than InterHand [36]. One can conclude that our method is competitive against state-of-the-art, yielding accurate locating performance of the hand landmarks.

In Tables 3 and 4, we compare the execution times and the networks’ size, respectively. The execution time reported here is only the time of landmark localization, that is, the model forward time. As one can observe, our model has a size of 13.0 MB, which is less than 1/3 of SRHandNet. MediaPipe Hands is the smallest model because the official implementation has many optimizations, such as quantization. Furthermore, we list the inference runtime for different methods showing that FastHand can execute at 25 Frames Per Second (FPS) on an NVIDIA Jetson TX2 GPU and 31 FPS on an NVIDIA GeForce 940 MX GPU of a laptop outperforming each other method, without adopting any acceleration. Note that since the official implementation of MediaPipe Hands is based on TFlite<sup>2</sup> with considerable acceleration operations, for a fair evaluation, it is not involved in the inference runtime comparison.

### 4.4. Qualitative evaluation

We show the visualization results of the proposed pipeline in Fig. 4. Different perspectives of the hand pose are shown in each column. Our framework can accurately detect the landmarks of open hands, such as “Num 5”, while performs favorably in cases where occlusions occur, such as the fingertips for the pose of “thumbs-up” and “claw”. The high-accuracy hand landmark localization is due to our decoder design, which extracts high-resolution feature maps. In our experiments, we found that our method can recognize the hand gestures of numbers 0 to 9 and other basic ones, including the self-occluding poses. In Figs. 5 and 6, we show the results obtained by FastHand and the other methods on the STB and RHD datasets. As can be seen, our network predicts accurate coordinates outperforming the rest of the approaches.

<sup>2</sup> <https://www.tensorflow.org/lite/guide>.

The hand pose estimation videos can be found at: <https://youtu.be/r9CKDhXXs3Y>.

## 5. Conclusions

This paper proposes a fast 2D hand pose estimation pipeline, referred to as “FastHand”. The core of our framework is based on a lightweight encoder–decoder network. Through hand detection and stabilization, the ROI is cropped and fed into our network for landmark localization. The proposed network is trained over our large-scale 2D hand pose dataset and a synthetic one while evaluated on two publicly available datasets. Compared to several state-of-the-art techniques, the proposed approach shows a favorable performance while maintaining a real-time speed (over 25 FPS) on a low-cost GeForce 940MX GPU and an NVIDIA Jetson TX2 GPU. Our future plans include the utility of our method in complex robot applications and the utilization of Part Affinity Fields [45] aiming to differentiate between left hand and right hand.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

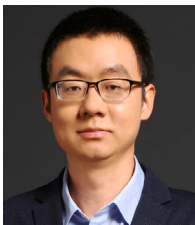
## Acknowledgments

This work was supported by grants from the National Key Research and Development Program of China (Grant No. 2020YFC2006200), the Major Program of the National Natural Science Foundation of China (Grant No. 91846201), the National Natural Science Foundation of China (Grant No. 72071069 and No. 61773272).

## References

- [1] L. Brethes, P. Menezes, F. Lerasle, J. Hayet, Face tracking and hand gesture recognition for human-robot interaction, in: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), Vol. 2, 2004, pp. 1901–1906.
- [2] G. Cicerelli, C. Attolico, C. Guaragnella, T. D’Orazio, A kinect-based gesture recognition approach for a natural human robot interface, *Int. J. Adv. Robot. Syst.* 12 (3) (2015) 22.
- [3] K. Ehlers, K. Brama, A human-robot interaction interface for mobile and stationary robots based on real-time 3d human body and hand-finger pose estimation, in: IEEE 21st Int. Conf. Emerging Technologies and Factory Automation (ETFA), 2016, pp. 1–6.
- [4] O. Mazhar, S. Ramdani, B. Navarro, R. Passama, A. Cherubini, Towards real-time physical human-robot interaction using skeleton information and hand gestures, in: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), 2018, pp. 1–6.
- [5] J.-Y. Chang, A. Tejero-de Pablos, T. Harada, Improved optical flow for gesture-based human-robot interaction, in: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), 2019, pp. 7983–7989.
- [6] I. Kansizoglou, L. Bampis, A. Gasteratos, An active learning paradigm for online audio-visual emotion recognition, *IEEE Trans. Affect. Comput.* (2019).
- [7] S. Goldin-Meadow, The role of gesture in communication and thinking, *Trends Cogn. Sci.* 3 (11) (1999) 419–429.
- [8] J.C. Nunez, R. Cabido, J.J. Pantrigo, A.S. Montemayor, J.F. Velez, Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition, *Pattern Recognit.* 76 (2018) 80–94.
- [9] G. Devineau, F. Moutarde, W. Xi, J. Yang, Deep learning for hand gesture recognition on skeletal data, in: 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG), 2018, pp. 106–113.
- [10] X.S. Nguyen, L. Brun, O. Lézoray, S. Bougleux, A neural network based on SPD manifold learning for skeleton-based hand gesture recognition, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12036–12045.
- [11] J. Liu, Y. Liu, Y. Wang, V. Prinet, S. Xiang, C. Pan, Decoupled representation learning for skeleton-based gesture recognition, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5751–5760.
- [12] I. Oikonomidis, N. Kyriazis, A.A. Argyros, Efficient model-based 3d tracking of hand articulations using kinect., in: The British Machine Vision Conference (BMVC), Vol. 1, 2011, p. 3.
- [13] Z. Ren, J. Yuan, J. Meng, Z. Zhang, Robust part-based hand gesture recognition using kinect sensor, *IEEE Trans. Multimedia* 15 (5) (2013) 1110–1120.
- [14] A. Tagliasacchi, M. Schröder, A. Tkach, S. Bouaziz, M. Botsch, M. Pauly, Robust articulated-icp for real-time hand tracking, in: Computer Graphics Forum, Vol. 34, 2015, pp. 101–114.
- [15] L. Ge, H. Liang, J. Yuan, D. Thalmann, Robust 3d hand pose estimation from single depth images using multi-view cnns, *IEEE Trans. Image Process.* 27 (9) (2018) 4422–4436.
- [16] C. Wan, T. Probst, L.V. Gool, A. Yao, Self-supervised 3d hand pose estimation through training by fitting, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10853–10862.
- [17] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, T. Brox, FreiHAND: A dataset for markerless capture of hand pose and shape from single rgb images, in: Proc. IEEE Int. Conf. Computer Vision (ICCV), 2019, pp. 813–822.
- [18] Y. Chen, H. Ma, D. Kong, X. Yan, J. Wu, W. Fan, X. Xie, Nonparametric structure regularization machine for 2d hand pose estimation, in: IEEE Winter Conf. Applications of Computer Vision (WACV), 2020, pp. 381–390.
- [19] N. Santavas, I. Kansizoglou, L. Bampis, E. Karakasis, A. Gasteratos, Attention! a lightweight 2D hand pose estimation approach, *IEEE Sens. J.* (2020).
- [20] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, J. Yuan, 3d hand shape and pose estimation from a single rgb image, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10833–10842.
- [21] C. Zimmermann, T. Brox, Learning to estimate 3d hand pose from single rgb images, in: Proc. IEEE Int. Conf. Computer Vision (ICCV), 0000, pp. 4903–4911.
- [22] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, C. Theobalt, Generated hands for real-time 3d hand tracking from monocular rgb, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 49–59.
- [23] Y. Cai, L. Ge, J. Cai, J. Yuan, Weakly-supervised 3d hand pose estimation from monocular rgb images, in: Proc. Eur. Conf. Computer Vision (ECCV), 2018, pp. 666–682.
- [24] U. Iqbal, P. Molchanov, T. Breuel Juergen Gall, J. Kautz, Hand pose estimation via latent 2.5 d heatmap regression, in: Proc. Eur. Conf. Computer Vision (ECCV), 2018, pp. 118–134.
- [25] T. Simon, H. Joo, I. Matthews, Y. Sheikh, Hand keypoint detection in single images using multiview bootstrapping, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1145–1153.
- [26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
- [27] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258.
- [28] D. Kulon, R.A. Guler, I. Kokkinos, M.M. Bronstein, S. Zafeiriou, Weakly-supervised mesh-convolutional hand reconstruction in the wild, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4990–5000.
- [29] Y. Xu, X. Wang, 3D hand pose estimation from single depth images with label distribution learning, in: 2020 IEEE International Conference on Embedded Software and Systems (ICES), IEEE, 2020, pp. 1–5.
- [30] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, M. Grundmann, Mediapipe hands: On-device real-time hand tracking, 2020, arXiv preprint arXiv:2006.10214.
- [31] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4724–4732.
- [32] Y. Wang, C. Peng, Y. Liu, Mask-pose cascaded cnn for 2d hand pose estimation from single color image, *IEEE Trans. Circuits Syst. Video Technol.* 29 (11) (2018) 3258–3268.
- [33] Y. Wang, B. Zhang, C. Peng, SrHandNet: Real-time 2d hand pose estimation with simultaneous region localization, *IEEE Trans. Image Process.* 29 (2019) 2977–2986.
- [34] D. Kong, Y. Chen, H. Ma, X. Yan, X. Xie, Adaptive graphical model network for 2d handpose estimation, 2019, arXiv preprint arXiv:1909.08205.
- [35] D. Kong, H. Ma, Y. Chen, X. Xie, Rotation-invariant mixed graphical model network for 2D hand pose estimation, in: IEEE Winter Conf. Applications of Computer Vision (WACV), 2020, pp. 1546–1555.
- [36] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, K.M. Lee, InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image, 2020, arXiv preprint arXiv:2008.09309.
- [37] A. Boukhayma, R.d. Bem, P.H. Torr, 3d hand shape and pose from images in the wild, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10843–10852.
- [38] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.
- [39] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single shot multibox detector, in: Proc. Eur. Conf. Computer Vision (ECCV), 2016, pp. 21–37.
- [40] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, Q. Yang, 3D hand pose tracking and estimation using stereo matching, 2016, arXiv preprint arXiv:1610.07214.
- [41] F. Gomez-Donoso, S. Orts-Escolano, M. Cazorla, Large-scale multiview 3d hand pose dataset, *Image Vis. Comput.* 81 (2019) 25–33.

- [42] S. Hampali, M. Oberweger, M. Rad, V. Lepetit, HO-3D: A multi-user, multi-object dataset for joint 3d hand-object pose estimation, 2019, arXiv preprint [arXiv:1907.01481](https://arxiv.org/abs/1907.01481).
- [43] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [44] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning, in: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, pp. 265–283.
- [45] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7291–7299.



**Shan An** received a Bachelor's degree in Automation Engineering from Tianjin University, China in 2007 and a Master's degree in Control Science from Shandong University, China, in 2010. He is currently a Ph.D. candidate in the field of computer vision at the School of Computer Science and Engineering in the Beihang University, China. He is a senior member of IEEE and a member of ACM. He has served as a program committee member for ACM Multimedia (2019–2021), AAAI (2022), and IJCAI (2021). He is a reviewer of more than twenty highly prestigious journals and conference papers, such as IEEE Transaction on Neural Network and Learning Systems, IEEE Transaction on Multimedia, and Pattern Recognition, CVPR, ICCV, ICRA, etc. His research interests include computer vision in robotics and AR, image retrieval, and image segmentation.



**Xiajie Zhang** received the Bachelor's degree in Automation from the Nanjing University of Information Science & Technology, and the Master's degree in Precision Instrument and Machinery from the University of Chinese Academy of Sciences, China, in 2015 and 2018, respectively. He is currently an algorithm engineer with Tech & Data Center in JD.COM Inc. His research interests include computer vision in robotics, deep learning and embedded systems.



**Dong Wei** received the Bachelor's degree in mathematics and applied mathematics, and the Master's degree in applied statistics from the Northeastern University, China, in 2017 and 2020, respectively. He is currently an algorithm engineer with Tech & Data Center in JD.COM Inc. His research interests include computer vision in robotics, deep learning and image processing.



**Haogang Zhu** received the Ph.D. degree from the University College London, London, United Kingdom. He was Professor of the School of Computer Science and Engineering, Beijing University of Aeronautics and Astronautics from 2015. His main research interest includes Bayesian analysis, machine learning, and image understanding. He worked as the Principle Investigator and Data Scientist on various projects sponsored by research councils and industry leaders such as NIHR UK, Fight for Sight, Heidelberg Engineering, Pfizer, Novartis and Carl Zeiss Meditec. His research has led to several patents and tools/systems used by research institutes and industrial companies. He was awarded a young scientist position in the Chinese Recruitment Program of Global Experts (1000 Talent Plan) in 2015.



**Jianyu Yang** received the B.E. degree from the University of Science and Technology of China in 2006, and the Ph.D. degree from City University of Hong Kong in 2012. He is currently an Associate Professor with the School of Rail Transportation, Soochow University, China. He was a Postdoctoral Research Fellow at the School of Electrical and Electronics Engineering (EEE), Nanyang Technological University (NTU), Singapore, in 2015. His current research interests include human behavior analysis and shape representation. He received the Best Conference Paper Award on IEEE ICAL 2010.



**Konstantinos A. Tsintotas** received a bachelor's degree from the department of automation, Technological Education Institute of Chalkida, Psachna, Greece, a master's degree in mechatronics from the department of electrical engineering, Technological Education Institute of Western Macedonia, Kila Kozanis, Greece, and a Ph.D. in robotics from the department of production and management engineering, Democritus University of Thrace, Xanthi, Greece, in 2010, 2015, and 2021, respectively. He is currently a postdoctoral fellow at the laboratory of robotics and automation, department of production and management engineering, Democritus University of Thrace. His work is supported through several research projects funded by the European Commission and the Greek government. His research interests include vision-based methods for modern, intelligent mechatronics systems. Details are available at <http://robotics.pme.duth.gr/ksintotas>.