EURASIP Journal on Advances
in Signal Processing

RESEARCH **Open Access**

Hand pose estimation based on improved NSRM network

Shiqiang Yang[1*], Duo He[2], Qi Li[3], Jinhua Wang[4] and Dexin Li[5]

*Correspondence:
yangsq@126.com

[1] School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, South Jinhua Road, Baiyin 710048, Gansu Province, People's Republic of China
[2] School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, South Jinhua Road, Yulin 710048, Shaanxi Province, People's Republic of China
[3] School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, South Jinhua Road, Zhumadian 710048, Henan Province, People's Republic of China
[4] School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, South Jinhua Road, Zhengzhou 710048, Henan Province, People's Republic of China
[5] School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, South Jinhua Road, Yantai 710048, Shandong Province, People's Republic of China

Abstract

Hand pose estimation is the basis of dynamic gesture recognition. In vision-based hand pose estimation, the performance of hand pose estimation is affected due to the high flexibility of hand joints, local similarity and severe occlusion among hand joints. In this paper, the structural relations between hand joints are established, and the improved nonparametric structure regularization machine (NSRM) is used to achieve more accurate estimation of hand pose. Based on the NSRM network, the backbone network is replaced by the new high-resolution net proposed in this paper to improve the network performance, and then the number of parameters is decreased by reducing the input and output channels of some convolutional layers. The experiment of hand pose estimation is carried out by using public dataset, the experimental results show that the improved NSRM network has higher accuracy and faster inference speed for hand pose estimation.

Keywords: Deep learning, Hand pose estimation, NSRM, NHRNet

1 Introduction

1.1 Background and significance

Hand pose estimation can infer 2D or 3D positions of hand keypoints in the input image, which has a wide range of application potential in virtual reality [1], human–computer interaction [2, 3] and other fields [4, 5]. Due to the high flexibility of hand joints, local similarity and severe occlusion, hand pose estimation remain to be further studied.

In recent years, more and more methods have emerged in the field of hand pose estimation, including multi-view RGB systems [6, 7], depth-based methods [8–10] and monocular RGB methods [11, 12], etc. Therefore, the accuracy, speed and other performance of hand pose estimation have been continuously improved. Although the 3D hand pose estimation [3, 5, 13, 14] has attracted more and more attention, the 2D hand pose estimation [2, 15, 16] is still an essential research direction. A large number of 3D hand pose estimation algorithms rely on the corresponding 2D algorithms [3, 17], which obtain the result of estimation by mapping the features of 2D space to 3D space. With the emergence of Deep Convolutional Neural Network (DCNN), human pose estimation has also made significant progress, and more and more excellent networks are emerging in this field, such as Convolution Pose Machine (CPM) [18], Residual Network [19] and

Stacked Hourglass Network (SHG) [20]. These methods implicitly access information of the body part and performed 2D pose estimation sub-module [7, 11, 12]. However, although DCNN has good representation ability, it cannot capture the complex structural relationship between the keypoints of the human hand, so it is difficult to deal with the serious hand occlusion problem.

Nonparametric Structure Regularization Machine (NSRM) [21] adopts cascading multi-task architecture to jointly learn keypoints and hand structure representation, and uses synthetic hand mask to guide keypoints structure learning. As an effective method in pose estimation field, High-Resolution Net (HRNet) [22, 23] maintains high-resolution transmission in the whole network. It consists of multiple branches with different resolutions, the low-resolution branch captures the context information, and the high-resolution branch stores spatial information. HRNet can generate high-resolution feature maps with rich semantics by using the multi-scale fusion between branches.

For better solving the problem of inaccurate location of keypoints caused by occlusion, interference and complex structural relationship of the human hand in 2D hand pose estimation, the main contributions of this paper are threefold:

1. We introduce an attention mechanism to the HRNet framework to obtain a more accurate and efficient model NHRNet.
2. We decrease the input–output channel of the convolutional layer in stages of NSRM, which aims to reduce the number of parameters of the network.
3. We improve the backbone network of NSRM by replacing VGG-19 with NHRNet to decrease the positioning error of keypoints.

The rest of this paper is organized as follows. Section 1.2 discusses related work; Sect. 2 describes our proposed network; Sect. 3 shows our experimental details; Sect. 4 discusses the experimental results; finally, Sect. 5 concludes the paper.

1.2 Related work

Research on hand pose estimation has developed rapidly in recent years. Hand pose estimation is a crucial task in computer vision, and its goal is to detect important skeletal points of the hand. Existing hand pose estimation methods can be roughly divided into three classes: generative method, discriminative method and hybrid method.

Generative method usually builds a deformable hand model and defines an objective function to compare the similarity between the collected image and the hand model. In order to optimize the objective function, the parameters of the hand model are adjusted iteratively to fit the input image, so as to obtain the optimal solution. Sridhar et al. [24] proposed a Gaussian mixture model of hands and established a new objective function to estimate hand posture. The model requires only relatively few computational resources which makes hand pose estimation very fast. Romero et al. [25] proposed a hand model called hand Model with Articulated and Non-rigid defOrmations (MANO). The model can learn from examples based on a linear manifold of pose synergies and a compact mapping from hand poses to pose blend shape corrections.

Discriminative method does not need to know the size of the hand, and does not need to perform motion constraints. It learns a mapping relationship between the observed features and the predicted output through the training data, so as to realize the estimation of the hand pose in the current image. Kong et al. [26] proposed a network architecture called Rotation-invariant Mixed Graphical Model Network (R-MGMN) to address the task of 2D hand posture estimation in monocular vision. Fang et al. [27] proposed a joint graph inference module depended on Graph Convolutional Network (GCN) to build complex dependencies among joints while enhancing the representation capability of each pixel. The offset of each pixel to the joint is estimated in the domain, and the position of the human hand joint is estimated based on the weighted average of all pixel predictions. Goodfellow et al. [28] proposed a Generative Adversarial Network (GAN) based on the core idea of synthesizing data in the skeleton space for outputting hand poses. Kourbane et al. [29] proposed a new 2D hand pose estimation method, which has multi-scale heatmap regression performance, and adopts hand skeleton as additional information to constrain the regression problem.

Hybrid method is a combination of generative method and discriminative method. Zhang et al. [30] proposed a unified optimization framework to jointly track hand pose and object motion. The model first segments the hand and object by a Deep Neural Network (DNN), and predicts the current hand pose based on the previous pose with a pretrained Long Short-Term Memory (LSTM) network, then reconstructs the target model with a nonrigid fusion technique. Chen et al. [31] proposed the Spherical Part Model (SPM) to represent the hand pose. The model can more precisely estimate the hand posture depending on the prior knowledge of the hand.

2 Methods

2.1 Principle and framework of NSRM

NSRM for 2D hand pose estimation is applied to learn keypoints representation by synthesizing hand masks. A new representation method for hand joint probability is emerged; the synthetic mask can be obtained from the keypoints without additional data annotation. The hand model contains 21 keypoints and 20 limbs connected by keypoints, which as shown in Fig. 1.

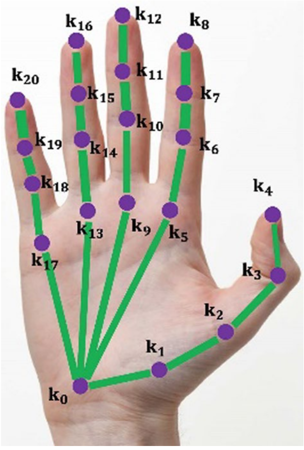


Fig. 1 Keypoints and limbs of the hand

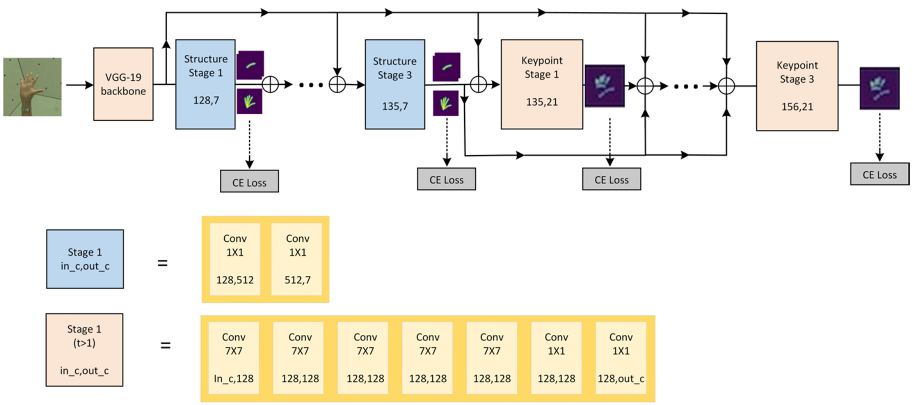


Fig. 2 NSRM framework

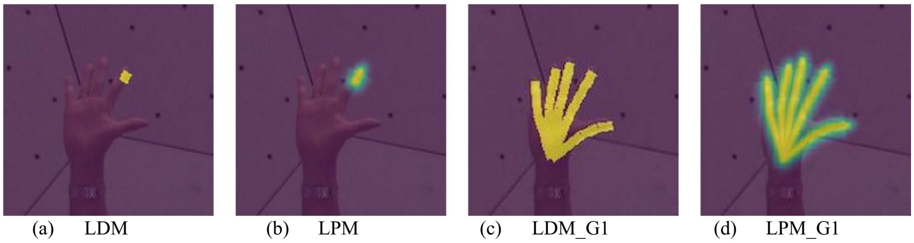

(a) LDM (b) LPM (c) LDM_G1 (d) LPM_G1

Fig. 3 Two mask representations

The NSRM network structure (G1_6) is shown in Fig. 2. Firstly, feature extraction is carried out on the backbone network to obtain the feature map of the hand. Secondly, input the feature map into the model to learn the limb structure and obtain the hand structure representation. Then, the feature map and the structure characterization are fused. Finally, learn hand posture and output keypoint coordinates.

2.1.1 Representation and composition of limb mask

For limb L between any two keypoints i and j, two masks are defined: Limb Deterministic Mask (LDM) and Limb Probabilistic Mask (LPM), as shown in Fig. 3a, b.

The pixels of L are those in the fixed width rectangle centered on the line segment $\overline{p_i p_j}$ between keypoints i and j, i.e.,

$$\begin{cases} 0 \leq \left(p - p_j\right)^T \left(p_i - p_j\right) \leq \left\|p_i - p_j\right\|_2^2 \\ \left|\left(p - p_j\right)^T u^{\perp}\right| \leq \sigma_{\mathrm{LDM}} \end{cases} \tag{1}$$

where $u^{\perp}$ is a vector perpendicular to $\overline{p_i p_j}$, and σ_{LDM} is a hyperparameter that controls limb width.

LDM represents that if a point p is in the rectangular area where a limb L is located, the pixel value of p is defined as 1, otherwise, it is defined as 0, i.e.,

$$S_{\mathrm{LDM}}(p|L) = \begin{cases} 1 \ \text{if} \ \ p \in L \\ 0 \ \text{otherwise} \end{cases} \tag{2}$$

where $p \in L$ is any pixel in the image.

LDM works poor in the actual scene due to its rough processing method, therefore LPM is proposed to solve this problem. Instead of representing a point with absolute 0 and 1 pixels, LPM represents it as the ratio of the distance from pixel p to line $\overline{p_ip_j}$ to a Gaussian threshold σ^2_{LPM}, i.e.,

$$S_{\text{LPM}}(p|L) = \exp\left(-\frac{D\left(p, \overline{p_ip_j}\right)}{2\sigma^2_{\text{LPM}}}\right) \tag{3}$$

where $D\left(p, \overline{p_ip_j}\right)$ is the distance between the pixel p and the line segment $\overline{p_ip_j}$, and σ_{LPM} is the hyperparameter that controls the diffusion of Gaussian distribution.

The mask representation can be divided into four types, named LDM_G1 (as shown in Fig. 3c), LPM_G1 (as shown in Fig. 3d), LDM_G1_6 and LPM_G1_6. G1 represents the whole hand with a single mask, while G6 divided limbs into six parts, it takes the form of a palm and five fingers, as shown in Fig. 4. Furthermore, G1 captures the entire hand structure, while G6 pays more attention to detail in local areas of the hand. G1_6 represents the integration of G1 and G6 versions.

2.1.2 Loss function

The loss function of structural modules adopts cross-entropy loss, i.e.,

$$\begin{aligned} L_{\text{S}} = & \sum_{t=1}^{T_{\text{S}}} \sum_{g \in G} \sum_{p \in I} S^*\left(p|g\right) \log \underset{t}{\hat{S}}\left(p|g\right) \\ & + \left(1 - S^*\left(p|g\right)\right) \log \left(1 - \underset{t}{\hat{S}}\left(p|g\right)\right) \end{aligned} \tag{4}$$

where T_{S} is the number of stages of structural learning, $\underset{t}{\hat{S}}\left(p|g\right)$ is the predicted value of structural module, $S^*\left(p|g\right)$ is the mask after the combination of limbs, and G is the number of groups (usually G1_6 is 7, G1 is 1).

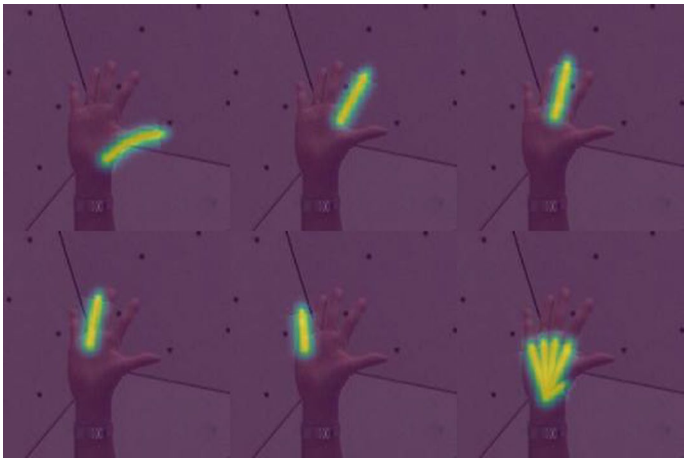


Fig. 4 LPM_G6

The confidence position KCM of keypoint k is defined as a 2D Gaussian distribution with marked keypoints centered on standard deviation σ_{KCM}, i.e.,

$$C^{*}(p|k) = \exp\left\{-\frac{\|p - p_{k}^{*}\|_{2}^{2}}{2\sigma_{\mathrm{KCM}}^{2}}\right\} \tag{5}$$

The loss function of the position prediction module adopts the sum-of-squared-error loss, i.e.,

$$L_{K} = \sum_{t=1}^{T_{K}}\sum_{k=1}^{K}\sum_{p\in I}\left\|C^{*}(p|k) - \underset{t}{\hat{C}}(p|k)\right\|_{2}^{2} \tag{6}$$

where T_{K} is the number of stages of pose learning, and $\underset{t}{\hat{C}}(p|k)$ is the predicted value of position module at pixel p and keypoint k of stage t.

The total loss function is:

$$L = \begin{cases} L_{K}+\lambda_{1}L_{S}^{G1} \ \mathrm{G1} \\ L_{K}+\lambda_{1}L_{S}^{G1}+\lambda_{2}L_{S}^{G6} \ \mathrm{G6} \end{cases} \tag{7}$$

where λ_{1}, λ_{2} are hyperparameters for controlling relative weight.

2.2 Improved NSRM network

By observing the NSRM network based on LPM_G1_6 mask, it can be found that the backbone network is the classic VGG-19. The quality of backbone network architecture will directly affect the strength of feature extraction capability, so its importance is self-evident. Although VGG-19 network can effectively extract the feature information of human hand keypoints, the resolution of feature maps obtained by VGG-19 network is very low and the spatial structure is lost, which leads to incomplete or invalid image features extraction. Moreover, the function map of 128 channels is generated through VGG-19 network, which leads to the increase in the number of parameters in the NSRM network and consumes more computing resources. Excellent HRNet pose estimation network can connect feature maps of different resolutions in parallel instead of simply concatenating them, which makes the whole network structure maintain high-resolution representation. The different resolution representations of the network at the same stage are fused repeatedly to enhance the capture of image information and improve the prediction accuracy. Therefore, the original backbone network VGG-19 is replaced by the HRNet network.

2.2.1 NHRNet model

The HRNet network mainly consists of four stages, and its structure is shown in Fig. 5. At the end of each stage, the network will be connected with the next stage through an exchange unit. After the end of the last stage, the output of the feature map is obtained. The HRNet network has four parallel sub-networks, and the resolution of each subnet is 1/4, 1/8, 1/16 and 1/32 of the input image, respectively. The network first reduces the image resolution to 1/4 of the input image through 3 × 3 convolution with stride 2, then enters the first stage which belongs to the first sub-network. The first stage includes four

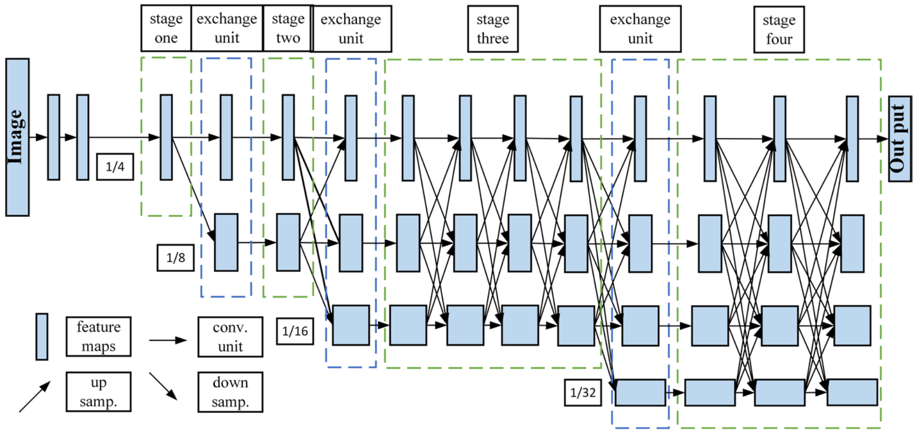


Fig. 5 HRNet framework

Bottleneck residual units with 64 output channels (represented by one unit in Fig. 5). The input feature map is generated by convolution module which contains convolution layer, Batch Normalization (BN) layer and Rectified Linear Unit (ReLU) layer, and the newly generated feature map is fused with the input feature map as the output feature map. There are 1, 4 and 3 resolution blocks in the second, third and fourth stages, respectively, each resolution block contains 4 Basic residual units (represented by one unit in Fig. 5). The resulting feature map is generated by two convolution modules (including convolution layer, BN layer and ReLU layer), and the new feature map is fused with the resulting feature map as the output feature map.

HRNet network performs cross-resolution feature fusion among sub-networks by exchange units. The resolution of the new sub-network is reduced to half of the previous one, but the number of channels is doubled. Cross-resolution feature fusion can be achieved by 2 (4 or 8) times down-sampling, convolution unit and 2 (4 or 8) times up-sampling. Multiple multi-resolution scale fusion is realized by exchange units, so that each high to low resolution representation receives the information of other parallel features repeatedly, finally obtains the high-resolution feature with rich information.

By introducing the attention mechanism, the network focuses its attention on the input information that is more critical to the current task, and even filters out the information that is irrelevant, so as to improve the efficiency and accuracy of the network. The idea of SA module proposed by SANet (Shuffle Attention Networks) [32] is convenient to implement and easy to load into the existing model framework. The architecture of SA module is shown in Fig. 6. SA module combines spatial attention and channel attention, which aims to capture the pixel-level pairwise relationship and channel dependency together. SA module first groups the channel dimensions and then performs parallel operations for each group. For each group, SA performs global average pooling (F_{gp}) and group normalization (GN) operations, respectively, to obtain channel-wise and spatial-wise information. Then all the groups are aggregated and the channel shuffle operator is used to realize the information communication between each group.

On the basis of the HRNet network, an attention channel is increased by introducing the SA module after the second convolution module of the Basic residual unit. The attention mechanism of HRNet is introduced to make the network pay more attention

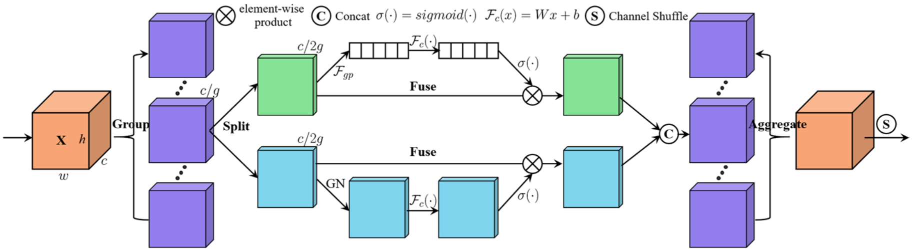


Fig. 6 SA-module

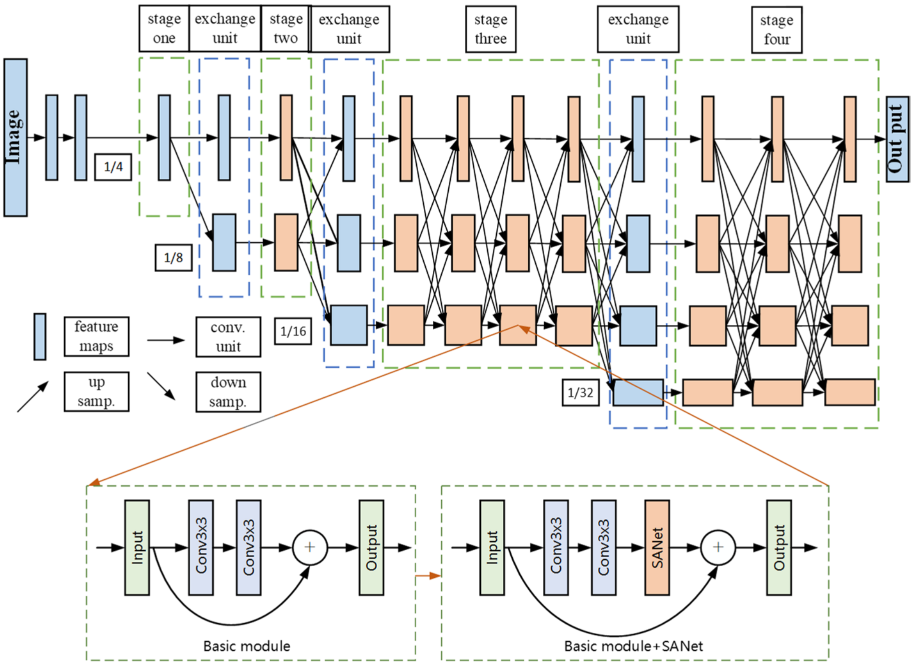


Fig. 7 The framework of NHRNet

to the skeleton point of human hand, then the NHRNet (NewHRNet) network model is constructed. The framework of NHRNet is shown in Fig. 7. The orange squares represent improved Basic modules. Take one of the squares as an example, the details of which are shown in Fig. 7. The input form of the Basic module is $x \in \mathbb{R}^{c\times w\times h}$, and the Basic module adopts the residual mapping, representing the required underlying mapping as $H(x)$, then $H(x) = F(x) + x$, and F is the two 3×3 convolution function used in the Basic module. After adding SA, the formula becomes the following form: $H(x) = SA(F(x)) + x$.

2.2.2 *Improved backbone network*

For improving the accuracy of neural network for human hand keypoint detection, this paper improves the NSRM network. This paper replaces the backbone network VGG-19 with NHRNet. The number of generated channels is decreased from 128 to 32, and the size of the obtained feature map is also changed due to the replacement of the backbone network. Therefore, the output channel of the first convolution layer in the first stage is changed to 128, and from the second stage to the sixth stage, the output channel of the first convolution layer and the input channel of the seventh convolution layer are changed to 32, so that the network performance can be improved and the number of network parameters can be reduced. The improved network structure is shown in Fig. 8. In front of the green arrow are the modules adopted in the NSRM network, and behind the green arrow are the improved modules.

The backbone network used in the improved network is NHRNet, and a 32-channel feature map is generated through NHRNet. The entire network architecture has six stages, each stage contains five convolution layers of 7 × 7 kernel size and two convolution layers of 1 × 1 kernel size (except the first stage). The first three stages are used for learning synthetic mask representation, and the last three stages are used for learning posture representation. The improved model's process is as follows: Firstly, all the hand images are resized to 256 × 256 and fed into the model, so as to generate 64 × 64 feature representation map for the mask method. As shown in Fig. 8, each stage consists of a range of convolution layers with $k \times k$ kernel size, and the specific input–output feature map channel number and convolution kernel size are shown in the rectangular box below. The output of structure stage 3 is fed to each keypoint stage. The output of the two kinds of network stages is a tensor with the size of (batchsize, 3, 7, 64, 64) and a tensor with the size of (batchsize, 3, 21, 64, 64), respectively, which is expressed as the position score of each six sets of masks and the confidence score of each point in the image.

3 Experiments

CMU Panoptic Hand dataset [4] was selected for model training and testing, and the effectiveness of the proposed model was verified by comparing with the existing excellent models.

3.1 Dataset and evaluation index

CMU Panoptic Hand dataset contains 14,817 human images, each of which has 21 annotated keypoints of the right hand. Since this paper focuses on the hand pose estimation rather than the hand target detection, the images are clipped according to the 2.2 times of the largest size of the boundary box containing every hand keypoint. The clipped

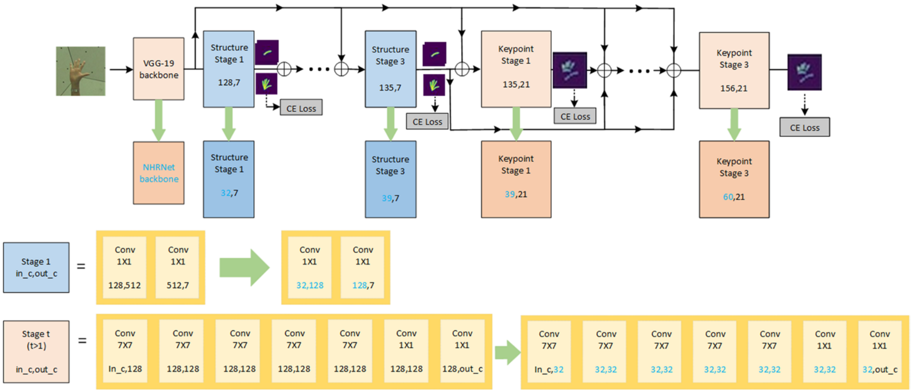


Fig. 8 Improved NSRM framework

dataset is arbitrarily split into three subsets for training (80%), validation (10%) and testing (10%), with 11,853, 1482 and 1482 hand targets, respectively.

The quality of detector d_0 is defined as the Probability of Correct Keypoint (PCK): PCK is defined as the proportion of the keypoints correctly estimated, that is, the proportion of the normalized distance between the detected keypoints and the corresponding groundtruth location less than the set threshold σ. For an especial keypoint p, this paper uses $PCK_{\sigma}^{P}(\mathrm{d}_0)$ [7] to represent it approximately on a test set T as follows:

$$\mathrm{PCK}_{\sigma}^{\mathrm{P}}(d_0) := \frac{1}{|T|}\sum_{T}\delta\left(\left\|x_{\mathrm{p}}^{\mathrm{f}} - y_{\mathrm{p}}^{\mathrm{f}}\right\|_2 < \sigma\right) \tag{8}$$

where $x_{\mathrm{p}}^{\mathrm{f}}$ represents the predicted p-th joint location of the detector, $y_{\mathrm{p}}^{\mathrm{f}}$ represents its true joint location, $\delta(\cdot)$ is an indicator function, and σ is the set threshold. Because the dataset used in this paper does not explicitly provide the size of the hand, we select to normalization with respect to the dimension of the tightest hand bounding box, and mean PCK (mPCK) with threshold $\sigma = \{0.04, 0.06, 0.08, 0.10, 0.12\}$.

3.2 Train and test settings

The model was trained on the training set, and tested on the validation set and test set. The experiment adopted Windows10 system and a P106-100 graphics card with 6 GB memory. The CPU version is Intel (R) Core (TM) i5-4460 CPU @ 3.20 GHz. The software environment is python3.7 + pytorch1.2.0 + cuda10.0 + cudnn7.4.0.

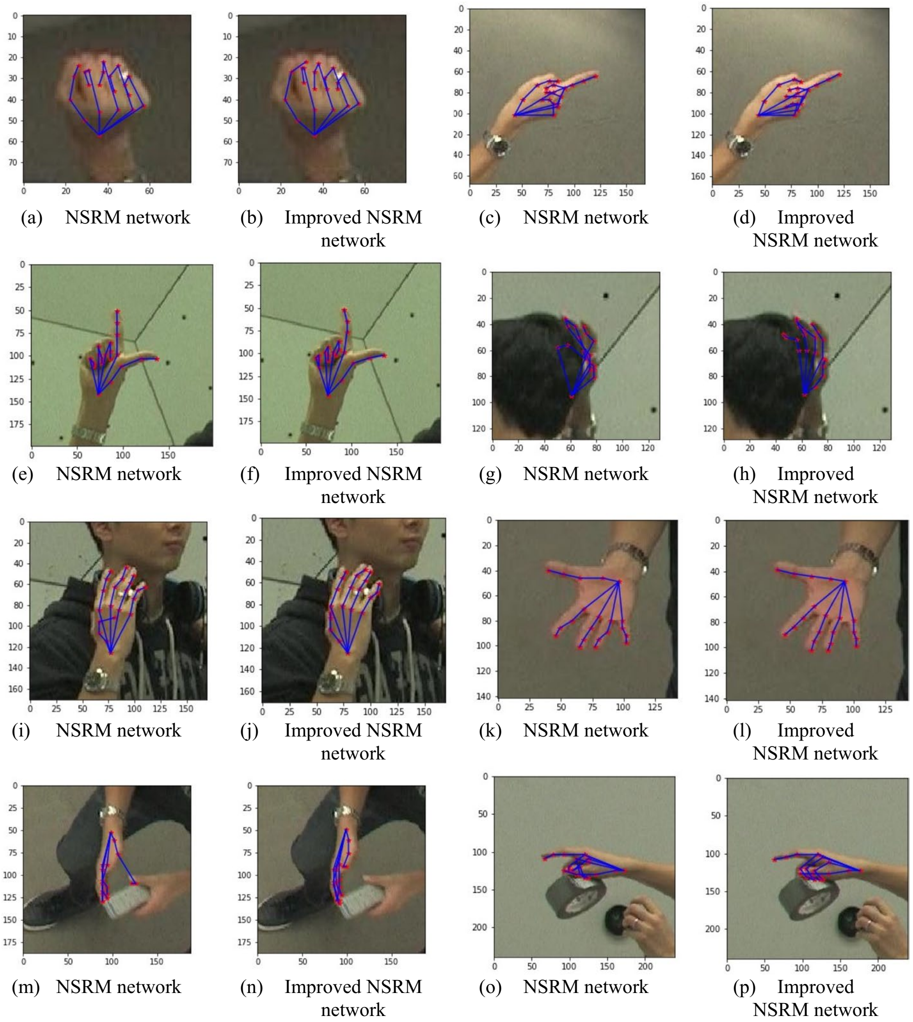
In this study, the Adam optimizer was used to train the model better, we set λ_1 to 0.1, λ_2 to 0.02 and decay by a ratio of 0.1 every 20 epochs. The batchsize is set to 8, the total number of training epochs is set to 80, the initial learning rate is set to 1e−4, and other parameters are set to the default value.

4 Results and discussion

4.1 Comparison with NSRM

In order to compare the effectiveness of the original NSRM network and the improved NSRM network in detecting hand keypoints, eight hand pictures in different states in the test set were selected for testing, respectively, and the results are shown in Fig. 9.

Where Fig. 9a, c, e, g, i, k, m, o shows the effect pictures of the NSRM network detection, and Fig. 9b, d, f, h, j, l, n, p shows the improved NSRM network detection. There are 4 keypoints for each finger and 1 keypoint at the wrist, for a total of 21 keypoints. In Fig. 9a, b, the right hand in a clenched fist is detected. By comparison, it is found that some of the keypoints of the thumb and middle finger are not accurately detected in Fig. 9a, while in Fig. 9b, each keypoint of the hand can be detected well by using the improved NSRM network. In Fig. 9c, d, a right hand with one finger exposed was detected. Both networks are able to roughly infer the locations of the 21 keypoints on the hand, but the predicted locations of the occluded keypoints are still slightly different from the true locations. In Fig. 9e, f, a right hand with two fingers exposed was detected, and in both images of the hand, two networks are able to roughly infer the 21 keypoints of the hand. In Fig. 9g, h, the right hand is obscured by the head, and only three fingers exposed can be detected. In Fig. 9i, j a right hand with four fingers exposed is detected, and the greatest ambiguity in these two images arises



(a) NSRM network (b) Improved NSRM network (c) NSRM network (d) Improved NSRM network

(e) NSRM network (f) Improved NSRM network (g) NSRM network (h) Improved NSRM network

(i) NSRM network (j) Improved NSRM network (k) NSRM network (l) Improved NSRM network

(m) NSRM network (n) Improved NSRM network (o) NSRM network (p) Improved NSRM network

Fig. 9 Test results

from the position of the thumb predicted by both networks. In Fig. 9k, l a right hand with an open hand showing the front is detected, and in two images with no occlusions of the hand, both networks are able to infer the positions of the 21 keypoints of the hand well. In Fig. 9m, n, two hands are detected in an image. By comparison, it can be seen that the NSRM network cannot detect the vertex position of the right thumb, while the improved NSRM network can detect this keypoint. In Fig. 9o, p, the right hand in the case of holding a tape is detected. By comparison, it is found that the detection of some keypoints of the little and middle fingers in Fig. 9o is biased, while the improved NSRM network in Fig. 9p can accurately detect each keypoint of the hand. Combining the various cases above, it can be found that both networks can detect the 21 keypoints of the right hand well in the no occlusion or less occlusion cases. In the case of the hand is occluded more or the hand is in a clenched state, the improved NSRM network is a better predictor of the occluded keypoints than the NSRM network. The improved NSRM network in this paper is to be superior to the original network for hand pose estimation.

Table 1 Comparison of the improved NSRM model with other models on the test set

σ_{PCK}	**0.04 (%)**	**0.06 (%)**	**0.08 (%)**	**0.10 (%)**	**0.12 (%)**	**mPCK (%)**	**Parameter file size (M)**
CPM [3]	55.25	73.23	81.45	85.97	88.80	76.94	-
LDM-G1 [12]	59.20	75.98	83.45	87.28	89.81	79.14	138
LDM-G1_6 [12]	59.16	76.32	83.63	87.46	90.03	79.32	139
LPM-G1 [12]	59.81	76.82	84.16	87.86	90.26	79.78	138
LPM-G1_6 [12]	59.73	76.86	84.43	88.23	90.87	80.03	139
Ours	66.68	79.72	85.99	89.54	91.73	82.74	114

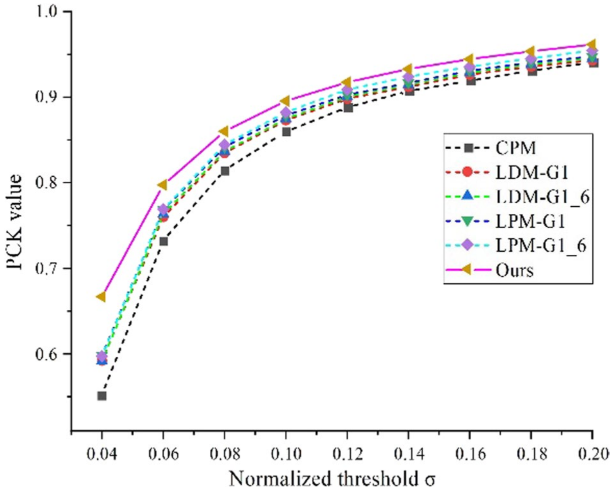


Fig. 10 PCK curve of different networks

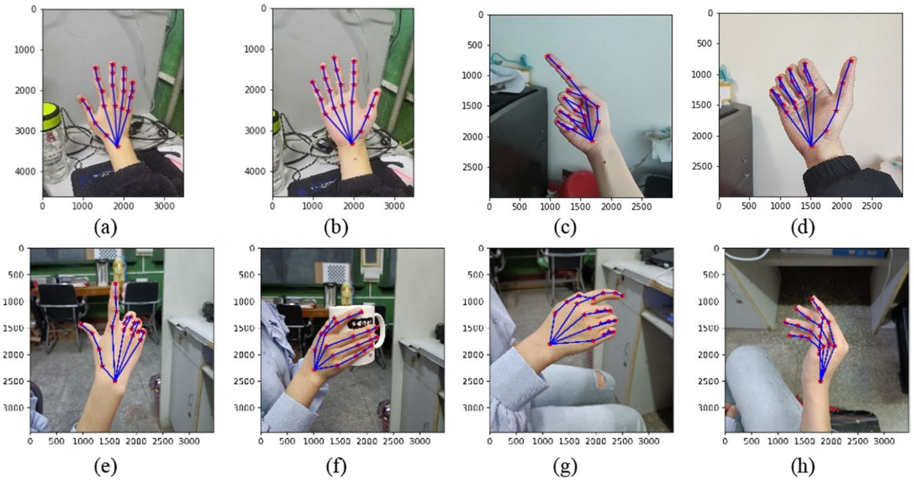
The improved NSRM model was evaluated on validation set and test set. The performance results of the improved NSRM model and other most advanced models on the test set are listed in Table 1. The PCK curves of different networks are shown in Fig. 10.

The PCK of the improved NSRM network in this paper is higher than that of the original network and other networks under different thresholds. When the threshold is 0.04, the PCK of the improved NSRM network is 6.95% higher than that of the NSRM network, the average PCK of the improved NSRM network is 2.71% higher than that of the NSRM network, the parameters of the improved NSRM network is 25M less than those of the NSRM network, indicating that the improved NSRM network is slightly better than other models.

4.2 Picture test in real scene

In order to detect the generalization ability of the improved NSRM network, eight images in real scenes were selected to detect hand keypoints using the improved NSRM network. The detection effect is shown in Fig. 11.

Figure 11a shows the back of the hand with the five fingers in the open state. Figure 11b shows the front side of the hand for a state with five fingers open. Figure 11c shows a state that the index finger expanded and the other fingers bent. Figure 11d shows a state



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 11 Test results in real-world scenarios

that the thumb is extended and the other four fingers are half closed. Figure 11e shows the back of the hand with the thumb and index finger straight and the other three fingers half closed. Figure 11f shows the state of holding the cup, in which the thumb is blocked. Figure 11g shows the back of the hand with the thumb half occluded, the index finger straight, and the other three fingers half closed. Figure 11h shows the side of the hand with a state of the fingers are slightly open, and there is a slight occlusion. It can be seen that the improved NSRM network can identify approximate positions of the keypoints of the hand in the eight images.

5 Conclusion

Human hand joints have high flexibility, local similarity and serious occlusion, which are supposed to have big impacts on hand posture estimation. In order to adopt to the complex hand posture and establish the structural relationship between the hand joints, this study replaced its backbone network with NHRNet based on the NSRM network, and reduced the input and output channels of some convolution layers to achieve more accurate and faster hand posture estimation.

On the CMU Panoptic Hand dataset, the PCK of the improved NSRM model under different thresholds is higher than that of other networks. Compared with the NSRM network, the PCK increases by 6.95% and the average PCK increases by 2.71% when the threshold is 0.04, and the improved NSRM network reduces the number of parameters. In the comparison experiment of the test set and the experiment in the real scene, it can also be seen that the improved NSRM network can identify the hand keypoints in different states. Therefore, the improved NSRM is an excellent hand pose estimation model.

Abbreviations

NSRM	Nonparametric Structure Regularization Machine
NHRNet	New High-Resolution Net
DCNN	Deep Convolutional Neural Network
CPM	Convolutional Pose Machine
SHG	Stacked Hourglass
HRNet	High-Resolution Net
MANO	hand Model with Articulated and Non-rigid defOrmations

R-MGMN	Rotation-invariant Mixed Graphical Model Network
GCN	Graph Convolutional Network
GAN	Generative Adversarial Network
DNN	Deep Neural Networks
LSTM	Long Short-Term Memory
SPM	Spherical Part Model
LDM	Limb Deterministic Mask
LPM	Limb Probabilistic Mask
BN	Batch Normalization
ReLU	Rectified Linear Unit
SANet	Shuffle Attention Networks
PCK	Probability of Correct Keypoint

Acknowledgements
The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of the manuscript.

Author contributions
All authors have contributed equally. All authors have read and approved the fnal manuscript.

Funding
This research is supported by Natural Science Foundation of Shaanxi Province, China (Grant No. 2021SF-422).

Availability of data and materials
The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 15 July 2022 Accepted: 3 January 2023
Published online: 12 January 2023

References

1. A.T. Aboukhadra, J. Malik, A. Elhayek, N. Robertini, D. Stricker, THOR-Net: end-to-end Graformer-based realistic two hands and object reconstruction with self-supervision (2022). arXiv preprint arXiv:2210.13853
2. N. Santavas, I. Kansizoglou, L. Bampis, E. Karakasis, A. Gasteratos, Attention! A lightweight 2D hand pose estimation approach. IEEE Sens. J. **21**(10), 11488–11496 (2021). https://doi.org/10.1109/JSEN.2020.3018172
3. W. Cheng, J.H. Park, J.H. Ko, HandFoldingNet: a 3D hand pose estimation network using multiscale-feature guided folding of a 2D hand skeleton, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), pp. 11240–11249. https://doi.org/10.1109/ICCV48922.2021.01107
4. L. Khaleghi, A. Sepas-Moghaddam, J. Marshall, A. Etemad, Multi-view video-based 3D hand pose estimation. IEEE Trans. Artif. Intell. (2022). https://doi.org/10.1109/TAI.2022.3195968
5. B. Doosti, Hand pose estimation: a survey (2019). arXiv:1903.01013
6. H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, Panoptic studio: a massively multiview system for social interaction capture. IEEE Trans. Pattern Anal. Mach. Intell. **41**(1), 190–204 (2019)
7. T. Simon, H. Joo, I. Matthews, Y. Sheikh, Hand keypoint detection in single images using multiview bootstrapping, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), pp. 4645–4653. https://doi.org/10.1109/CVPR.2017.494
8. Z. Zhang, S. Xie, M. Chen, H. Zhu, HandAugment: a simple data augmentation method for depth-based 3D hand pose estimation (2020). arXiv preprint arXiv:2001.00702
9. L. Ge, Y. Cai, J. Weng, J. Yuan, Hand PointNet: 3D hand pose estimation using point sets, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 8417–8426
10. S. Yuan, G.G. Hernando, B. Stenger, G. Moon, J.-Y. Chang, K.-M. Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge, Depth-based 3D hand pose estimation: from current achievements to future goals, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 2636–2645
11. Y. Cai, L. Ge, J. Cai, J. Yuan, Weakly-supervised 3D hand pose estimation from monocular RGB images, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 666–682
12. P. Panteleris, I. Oikonomidis, A. Argyros, Using a single RGB frame for real time 3D hand pose estimation in the wild, in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2018), pp. 436–445

13. J. Cheng, Y. Wan, D. Zuo, C. Ma, J. Gu, P. Tan, H. Wang, X. Deng, Y. Zhang, Efficient virtual view selection for 3D hand pose estimation (2022). arXiv preprint arXiv:2203.15458
14. J. Malik, I. Abdelaziz, A. Elhayek, S. Shimada, S. Aziz Ali, V. Golyanik, C. Theobalt, D. Stricker, HandVoxNet: deep voxel-based network for 3D hand shape and pose estimation from a single depth map, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 7113–7122
15. M. Zhang, Z. Zhou, M. Deng, Cascaded hierarchical CNN for 2D hand pose estimation from a single color image. Multimed Tools Appl. 1–19 (2022)
16. I. Kourbane, Y. Genc, Skeleton-aware multi-scale heatmap regression for 2D hand pose estimation (2021). arXiv preprint arXiv:2105.10904
17. S. Guo, E. Rigall, L. Qi, X. Dong, H. Li, J. Dong, Graph-based CNNs with self-supervised module for 3D hand pose estimation from monocular RGB. IEEE Trans. Circuits Syst. Video Technol. **31**(4), 1514–1525 (2021)
18. S. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4724–4732
19. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778
20. A. Newell, K. Yang, D. Jia, Stacked hourglass networks for human pose estimation, in *European Conference on Computer Vision* (Springer, Cham) (2016), pp. 483–499
21. Y. Chen, H. Ma, D. Kong, X. Yan, J. Wu, W. Fan, X. Xie, Nonparametric structure regularization machine for 2D hand pose estimation, in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020), pp. 370–379
22. K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 5686–5696
23. J. Wang, K. Sun, T. Cheng, B. Jiang, B. Xiao, Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3349–3364 (2021)
24. S. Sridhar, F. Mueller, A. Oulasvirta, C. Theobalt, Fast and robust hand tracking using detection-guided optimization, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3213–3221
25. J. Romero, D. Tzionas, B. Michael, Embodied hands: modeling and capturing hands and bodies together. ACM Trans. Graph. **36**(6), 245:1-245:17 (2017)
26. D. Kong, Y. Chen, H. Ma, X. Yan, X. Xie, Adaptive graphical model network for 2D handpose estimation (2019). arXiv:1909.08205
27. L. Fang, X. Liu, L. Liu, H. Xu, W. Kang, JGR-P2O: joint graph reasoning based pixel-to-offset prediction network for 3D hand pose estimation from a single depth image, in *European Conference on Computer Vision* (Springer, Cham) (2020), pp. 120–137
28. I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks (2014). arXiv:1406.2661
29. I. Kourbane, Y. Genc, Skeleton-aware multi-scale heatmap regression for 2D hand pose estimation (2021). arXiv:2105.10904
30. H. Zhang, Z.H. Bo, J.H. Yong, F. Xu, InteractionFusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions. ACM Trans. Graph. **38**(4), 481–4814 (2019)
31. T.-Y. Chen, P.-W. Ting, M.-Y. Wu, L.-C. Fu, Learning a deep network with spherical part model for 3D hand pose estimation, in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (2017), pp. 2600–2605
32. Y.-B. Yang, SA-Net: shuffle attention for deep convolutional neural networks, in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, IEEE)* (2021), pp. 2235–2239

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.