This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3005522, IEEE Transactions on Circuits and Systems for Video Technology

Learning to Acquire the Quality of Human Pose Estimation

Lin Zhao, Jie Xu, Chen Gong, *Member, IEEE,* Jian Yang, *Member, IEEE,* Wangmeng Zuo, *Senior Member, IEEE,* and Xinbo Gao, *Senior Member, IEEE*

***Abstract*—Making human poses serve high-level computer vision tasks such as action recognition, recognizing the quality of estimated poses is of critical importance. Conventionally, the mean confidence of each keypoint is used as pose quality in most human pose estimation frameworks. However, because different types of keypoint are not identical in visibility and size, they should not contribute equally, which produces biased quality scores. In the paper, we propose end-to-end human pose quality learning, which adds a quality prediction block alongside pose regression. The proposed block learns the object keypoint similarity (OKS) between the estimated pose and its corresponding ground truth by sharing the pose features with heatmap regression. The predicted OKS correlates well with pose quality, making the selection of reliable poses straightforward. Moreover, utilizing the learned quality as pose score improves pose estimation performance during COCO AP evaluation, because it ranks more accurate ones high among all pose detections. We conduct extensive experiments based on the three most popular human pose estimation frameworks, including Hourglass, SimpleBaseline and HRNet. Adding the proposed quality learning block is able to consistently bring nearly 1 percent AP improvement on all the frameworks.**

***Index Terms*—human pose estimation, prediction quality acquisition, end to end learning.**

I. Introduction

Human poses or named as human skeletons convey significant information, they facilitate many high-level computer vision tasks: human action recognition [1], [2], video anomaly detection [3], [4], person re-identification [5], [6], to name but a few. Nowadays, Deep Convolutional Neural Networks (DCNN) have remarkably advanced the performance of human pose estimation, making utilizing pose detections from images [7] rather than 3D sensors [8] possible. However, it is quite important to make sure the used poses achieve high accuracy, as inaccurate pose estimations may inevitably affect the following reasoning in high-level tasks. Fig. 1 gives an example of detected poses in images using modern DCNN-based models. As can be seen, we will get both positive and negative results, and certainly the accurate poses need to be selected. Thus, awareness of the quality of each prediction is essential for developing practical applications of pose estimation.

Fig. 1. Illustration of the necessity of acquiring the quality of pose detections. The same backbone network of pose estimation [9] is utilized to detect poses in the figure. The first line shows all pose detection results without selection. Alongside pose estimation, we add the proposed OKS module (OKS-net) on the backbone to predict the quality of detected poses. Using the produced quality score by OKS-net, accurate poses are able to be selected, as given in the second line. For comparison, the third line shows the results filtered by using the conventional heatmap post-processing score. Clearly, the scores predicted by OKS-net can better represent the true quality of the detected poses. We mark some bad pose detections that are filtered out by OKS-net score in the first line to better exhibit the advantage of OKS-net.

Although present-day pose estimation frameworks enjoy phenomenal success on large scale benchmark datasets [10], [11], most of them are not good at properly judging the quality of their pose predictions. The typical process of human pose estimation just regresses heatmaps of keypoints, and the pixel getting the maximum heat value will be chosen as keypoint. Conveniently, the maximum value in heatmap is considered as the confidence of the predicted keypoint, and the quality

L. Zhao, J. Xu, and J. Yang are with the PCA Lab, the Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and the Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. (e-mail: linzhao@njust.edu.cn, jiexu@njust.edu.cn, csjyang@njust.edu.cn)

C. Gong is with the PCA Lab, the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; and is also with the Department of Computing, Hong Kong Polytechnic University, Hong Kong SAR, China. (e-mail: chen.gong@njust.edu.cn)

W. Zuo is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China. (e-mail: cswmzuo@gmail.com)

X. Gao is with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. (e-mail: gaoxb@cqupt.edu.cn)

Corresponding authors: Lin Zhao (linzhao@njust.edu.cn) and Jian Yang (csjyang@njust.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3005522, IEEE Transactions on Circuits and Systems for Video Technology

Fig. 2. Examples of misalignment between heatmap post-processing scores and the actual pose quality. GT-OKS represents the ground truth pose quality, which is obtained by calculating the object keypoint similarity (OKS) between the estimated pose and its corresponding ground truth. OKS-net pred gives the quality score predicted by the proposed OKS-net.

of a pose is computed by taking the mean of its keypoints' confidence. Quality scores obtained from this widely-used post-processing technique are adopted as detection scores for computing precision-recall (PR) curves and average precision (AP) in evaluation. Nonetheless, these scores may not correctly reflect the real quality of the corresponding pose predictions. Fig. 2 presents some cases of misalignment between post-processing scores and the actual pose quality.

The reason for heatmap post-processing scores being inappropriate to make pose quality can lie in twofold: (1) First, values in heatmaps represent the likelihood of being keypoint. They naturally act for classification confidence but not localization confidence. There is a gap between these two concepts. Both studies in object detection [12] and instance segmentation [13] show that localization accuracy is not well correlated with classification confidence. (2) Second, keypoints may have different influences on pose quality according to their types. Obviously, invisible keypoints should not take into account. Besides, due to the disparity in size between different keypoints, for example hip and eye, accuracy requirements diverge. Thus, the average confidence of keypoints is incompetent to be pose quality.

In this paper, we introduce end to end pose quality learning, which adds a network block to the existing human pose estimation frameworks. The attached block does not aim to boost the accuracy of heatmap regression, but just attempts to acquire the quality of the corresponding pose predictions. The proposed quality learning strategy shares the same pose feature for heatmap regression, thus it is compatible with any framework for single-person pose estimation. We will demonstrate its effectiveness on the three most popular pose networks: Hourglass [14], SimpleBaseline [15] and HRNet [9].

We name the proposed quality learning network OKS-Net, which directly predicts the Object Keypoint Similarity (OKS) between detected keypoints and the ground truth keypoints of a pose. OKS is used in the keypoint evaluation by COCO [10] that mimics Intersection-over-Union (IoU) in object detection, thus it is a natural measure of pose quality. We conduct experiments thoroughly with the proposed OKS-net, and the results show that it is able to produce quality score correlating very well with the ground truth OKS. This learned pose quality brings major benefits to advance human pose estimation:

1) OKS-net provides quality predictions that have a clear physical interpretation. According to the definition of the OKS, they represent the degree of similarity to the ground truth pose. Thus, we can directly judge the localization accuracy of detected poses from the predicted quality score. By contrast, scores given by heatmap post-processing only serve as the ranking keyword in evaluation.
2) Noticeable improvement can be observed on the challenging COCO benchmark, using OKS-net predictions to make pose score. Experimental results on Hourglass, SimpleBaseline and HRNet all show nearly 1% AP improvement, regardless of the size of the input image. We will make the source code and models of OKS-net publicly available.

II. Related Work

A. Human Pose Estimation

Human pose estimation is a long-researched task in the community of computer vision. Methods based on the probabilistic graphical models [16], [17] and the pictorial structure models [18], [19] dominate the traditional solutions. Nowadays, deep models far surpass the performance of traditional ones. Deeppose [20] marks a watershed method for human pose estimation. Since then the power of deep learning has been gradually exploited to better model part appearance [21] and body structure [22]. Also, the inference process of graphical models has been integrated into deep models [23] for better performance. Generally, deep models for human pose estimation can be classified into bottom-up [24] and top-down [25] two types, according to whether human body bounding boxes are needed.

Bottom-up methods. All keyoints in an input image are detected at one time in bottom-up methods, then keypoints belonging to the same person are clustered to form separate poses. Deepcut [26] applies CNN-based detectors to locate keypoints, and Integer Liner Program is utilized for grouping. Openpose [7] becomes the state-of-the-art method by introducing part affinity fields, which makes keypoints grouping much easier. Because of its real-time performance and good implementation, this method is quite popular in the community. Associative embedding [27] is a contemporary model with Openpose that shares similar ideas.

Top-down methods. Human bounding boxes are a prerequisite for top-down methods, which can use the ground truth or results from person detectors [28]. Thus, this kind of methods mainly deals with the problem of single-person pose estimation [29]. Various skills and strategies have been proposed to polish up deep models on getting more accurate poses, and they can roughly fit into three categories according to the main structure. Hourglass [14] and its follow ups [30] constitute the largest category. The vanilla Hourglass attracts much attention due to its elegant structure and good performance. Based on this structure, a multi-context attention mechanism [31] is added to boost performance. Feature pyramids [32] are introduced to replace the basic residual block for larger receptive field. Adversarial learning [33] is also utilized for training more robust models. ResNet [34] based models fall into another important category. These models utilize ResNet to extract high-level pose features, then design different strategies to transform low-resolution features to high-resolution heatmaps. Bilinear-upsampling and multi-scale fusion are employed in cascaded pyramid network [35], afterwards channel-wise and spatial-wise attention [36] are put to use for enhancing feature maps. SimpleBaseline [15] proves that simple transpose convolution is enough to regress heatmaps. In addition, pyramidical gather-excite context is proposed in [37] to extend ResNet based pose estimation models for human parsing [38]. HRNet [9] is a recently proposed deep structure, which differs all the above models by maintaining high-resolution features throughout the whole process. This model is efficient in computation. Nevertheless, it achieves state-of-the-art performance.

Because top-down methods can zoom in on each individual person, they are less sensitive to the size of input images. Besides, the latest object detectors are able to provide quite robust human bounding boxes. Generally, top-down methods are able to present more accurate pose estimation results than bottom-up ones. In this paper, we are interested in learning the quality of pose predictions within the top-down framework. Given a cropped image containing one person, our proposed model gives the prediction of human pose and offers a score indicating its quality at the same time.

B. Detection Quality Acquisition

Getting a better idea of prediction quality has been an interesting topic in the area of object detection for many years. Earlier works mainly focused on correcting the score of detections to get higher AP in evaluation. SoftNMS [39] refines score using the overlap between detected boxes, rather than eliminating low score boxes. Fitness NMS [40] classifies box IoU and corrects score by learning from the ground truth. Rescoring hard positive samples is proposed in [41] by using a decoupled classification refinement network.

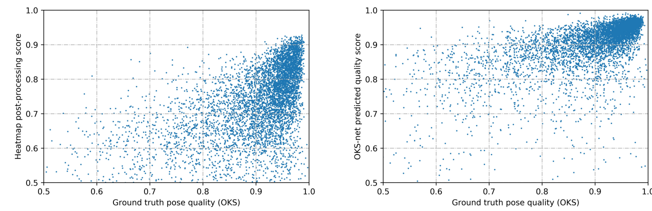
The above works only focus on the refinement of scores for ranking detections in evaluation. Nonetheless, whether the score can reflect the actual quality of detection is neglected. IoU-Net [12] shows conventional CNN-based detectors misalign classification confidence with localization accuracy, and proposes to regress box IoU directly. Mask Scoring RCNN [13] borrows the same idea to learn mask score for instance segmentation.

Our method shares the same purpose with these works, however, it is designed for human pose estimation. We argue that the conventional strategy used in existing methods is not able to get the optimal pose score. The proposed end to end learning framework provides pose quality based on physical interpretation and improves the performance of human pose estimation.

III. METHOD

A. Motivation

To do human pose estimation, regressing heatmap rather than the position of keypoint is widely adopted in the community. Because heat value depicts the likelihood of being keypoint, it is natural to choose the location with the highest value. Nevertheless, the likelihood is also used as keypoint confidence, and the score of pose hypothesis is determined by the average of all keypoints' confidence. This heatmap post-processing score is utilized by most of the existing pose estimation frameworks, however, it is problematic to represent pose quality. In COCO keypoint evaluation, the OKS is defined as a similarity measure which plays the same role as the IoU in object detection. Its formulation is given as:



(a) Scores from vanilla HRNet (b) OKS-net predicted scores

Fig. 3. The distribution of pose scores over the OKS. (a) Scores obtained by heatmap post-processing lack the ability to represent the actual pose quality as they are biased. Many pose detections with the high OKS get low scores. (b) We propose OKS-net to predict the OKS for each pose estimation, which produces scores much closer to the ground truth OKS.

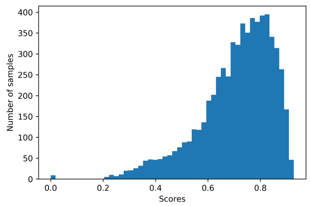
$$OKS = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (1)$$

Here, $\exp(-d_i^2/2s^2k_i^2)$ is keypoint similarity, d_i is the Euclidean distance between each detected keypoint and its matched ground truth, s is the object scale, k_i is a keypoint-specific constant that controls falloff, v_i is the visibility flag of the ground truth.

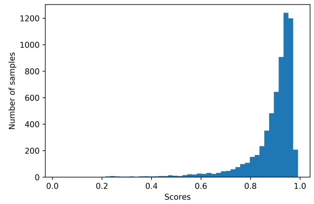
For each pose hypothesis, the OKS yields a similarity ranging between 0 and 1, this makes the optimal score and naturally reflects the actual quality. Fig. 3(a) shows the distribution of vanilla scores from HRNet [9] over the OKS. The experiment is conducted on the COCO 2017 validation set, and the ground truth bounding box is utilized. As can be seen, heatmap post-processing scores are heavily biased towards the higher OKS. This manifests that heatmap post-processing is not able to produce the optimal score. To better illustrate this problem, we compare the histogram of the scores from HRNet with the OKS in Fig. 4. There is also large disparity of value distribution between them. It is clear pose scores obtained by heatmap post-preprocessing may not align with the OKS. At least, they cannot truly reflect the actual quality of estimated poses, which is important in practical applications of human pose estimation.

Inaccurate scores also harm the performance in AP evaluation. In the task of object detection or segmentation, the classification confidence is usually utilized as the score [41]. IoU-Net [12] and Mask Scoring RCNN [13] prove that the correlation of the classification confidence with the ground truth score is very weak (as shown in Fig. 5(a)), and this affects AP in evaluation. Nevertheless, we see from Fig. 3(a) that the heatmap post-processing scores are not so poorly correlated to the ground-truth scores. Though the scores are

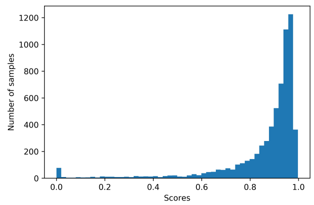
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3005522, IEEE Transactions on Circuits and Systems for Video Technology



(a) Heatmap post-processing

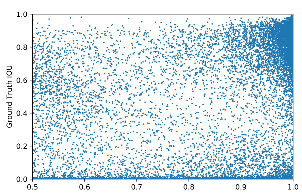


(b) OKS-net predictions

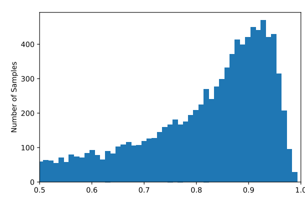


(c) Ground truth OKS

Fig. 4. Comparison of the histograms of heatmap post-processing scores, OKS-net predicted scores and the ground truth OKS. (a) The value distribution of heatmap post-processing scores is quite different from the ground truth. (b) OKS-net produces scores with much closer distribution.



(a) IoU vs. classification confidence



(b) Histogram of IoU

Fig. 5. The scoring problem in object detection or segmentation. (a) The correlation between the IoU and the score (classification confidence). (b) The histogram of the IoU of bounding boxes with the matched ground-truth. The experiment is conducted on Mask-RCNN [42], which is used in IoU-Net [12] and Mask Scoring RCNN [13].

biased, they may not do harm to AP evaluation. Using the error diagnosis tool of multi-instance pose estimation by Ronchi et al. [43], we carry out detailed study into the errors caused by the heatmap post-processing scores. We find that the scoring errors are mainly sensitive to the occlusion of keypoints. Fig. 6 presents the PR (Precision Recall) curves obtained by the heatmap post-processing scores and the ground-truth scores on four occlusion benchmarks respectively. While there is little scoring error on the benchmark of no occlusion, it is especially huge on the poses containing only less than five visible keypoints. Thus in pose estimation the heatmap post-processing scores introduce errors in AP evaluation, but the main reason is because it does not take care of the occlusion of poses, which is quite different from the task of object detection or segmentation. This means that we should carefully deal with occlusion to assign correct scores to estimated poses.

We propose OKS-net to produce scores for human pose estimation, which learns from the OKS to actually reflect pose quality. A channel attention mechanism is specially utilized in OKS-net to adaptively perceive the occlusion of different poses, hence occluded poses can also get right scores. The proposed OKS-net is a module analogous to the heatmap regressor, and can work with any pose estimation framework that estimates heatmap. Without loss of generality, Fig. 3(b) and Fig. 4(b) show the distribution and histogram of the scores given by OKS-net based on the HRNet framework, which exhibit similar characteristics comparing with the ground truth OKS.

B. Frameworks of Human Pose Estimation

Since the proposed pose quality learning works with the existing human pose estimation frameworks, we first give a brief introduction to them. Given an input image $\mathbf{I}$ of size $W \times H \times 3$, the goal is to estimate K heatmaps of size $W' \times H'$, where heatmap $\mathbf{H}_k$ represents the likelihood of the kth keypoint. The location of keypoint, (x_k, y_k), is obtained by heatmap post-processing. We follow this standard methodology for pose estimation and aim to give the quality of pose predictions.

To produce the heatmap of keypoint, three stages are mainly conducted in a convolutional network, which we dub image pre-processing, feature extraction, and heatmap regression. Fig. 7 illustrates this widely-adopted pipeline [14], [15], [9]. Image pre-processing typically decreases the input resolution to a quarter by two strided convolutions. Feature extraction usually processes the input low-level feature through a high-to-low and low-to-high procedure and outputs features containing both high-level and low-level information about keypoints. In Fig. 7, this feature is termed pose feature. Finally, heatmap regression produces heatmaps by several simple convolutions. Obviously, feature extraction is the core of the framework. Here, we introduce three most popular network designs of pose feature extraction. The proposed pose quality learning can work well with all the three networks.

Hourglass: Each hourglass [14] is a symmetric structure that performs bottom-up and top-down processing. High-level features in the up layers are combined with low-level features in the bottom layers through skip connections. The hourglass module is repeated several times to form a stacked network.

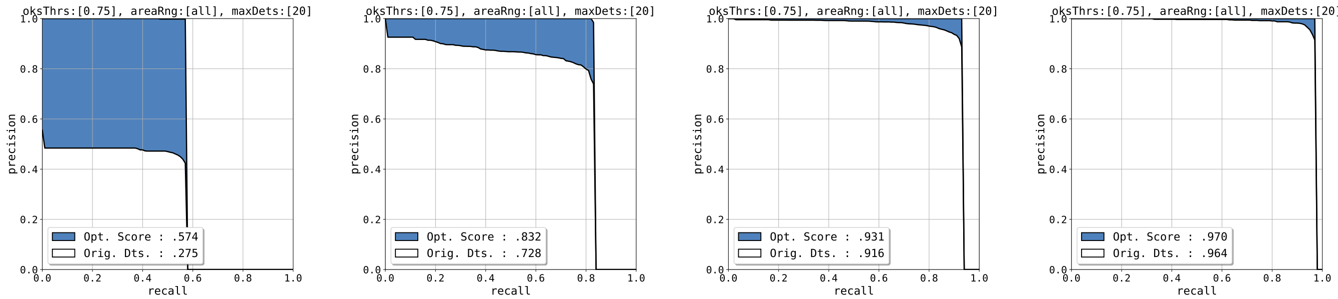
SimpleBaseline: High-level features are extracted by ResNet in SimpleBaseline [15], and transposed convolutions are utilized to up-sample the high-level feature maps to the resolution of heatmaps. This model demonstrates better performance than cascaded pyramid network [35], which also uses ResNet as the backbone.

HRNet: This model differs from both the above ones that it does not have a separate low-to-high process. In HRNet [9], several high-to-low resolution sub-networks are connected in parallel. Features produced by the sub-networks are repeatedly fused to generate reliable high-resolution representations.

We will not make any change to these networks but focus on learning pose quality within the frameworks.

C. Learning to Acquire Pose Quality

The pose quality learning block, namely OKS-net, is analogous to the regressor for estimating heatmap, as shown in Fig. 8. As the OKS naturally reflects pose quality, OKS-net attempts to regress the OKS between detected pose and its corresponding ground truth. The pose feature is utilized as



(a) 0-5 visible keypoints (b) 6-10 visible keypoints (c) 11-15 visible keypoints (d) 16-17 visible keypoints

Fig. 6. The improvement can be obtained by replacing the heatmap post-processing scores with the ground truth scores on the four occlusion benchmarks. The experiments are conducted based on HRNet-W32 [9] using the 256×192 input and ground truth bounding boxes.

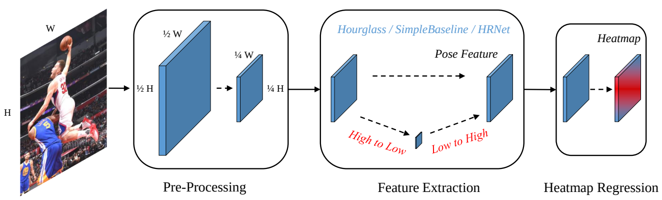


Fig. 7. The widely-adopted pipeline of human pose estimation using deep convolutional neural networks.

the input of OKS-net. An important part of designing OKS-net is to adaptively deal with the occlusion of poses, because occlusion largely affects the scoring error in AP evaluation as analyzed in Section III-A. In order to create the perception of occlusion, we apply the channel attention mechanism [44] to the pose feature. The pose feature is the latest feature before regressing heatmaps, and the main difference with heatmap is the number of channels. Thus the channel information of the pose feature corresponds to the type of keypoints. Several channels of the pose feature are shown in Fig. 9. We can see that each channel has the response at different locations, and some channels may have very weak response at all locations. It is reasonable to consider the channels with low response correspond to occluded keypoints. We therefore utilize the channel attention mechanism to make OKS-net regress the score mainly from visible keypoints. In this way, the impact of occluded keypoints can be significantly diminished. A max-pooling layer and a fully connected layer are utilized on the pose feature to get the re-weighting coefficient of each channel, and the pose feature is re-weighted by multiplying with the coefficients. Then, the channel attention re-weighted features are fed into 3 convolutional layers and 3 fully connected layers to predict the OKS. We set the kernel size and channel number to 3 and 32 respectively for all the convolutional layers. For the fully connected (FC) layers, the output of the first two is set to 1024. The final FC layer outputs the predicted OKS, which is a score ranging between 0 and 1.

OKS-net can be integrated into any single-person pose estimation network that regresses heatmap, and the whole network is capable of end-to-end training and inference. The training procedure is similar to that just estimates heatmap, we only need to add one loss for predicting the OKS. The

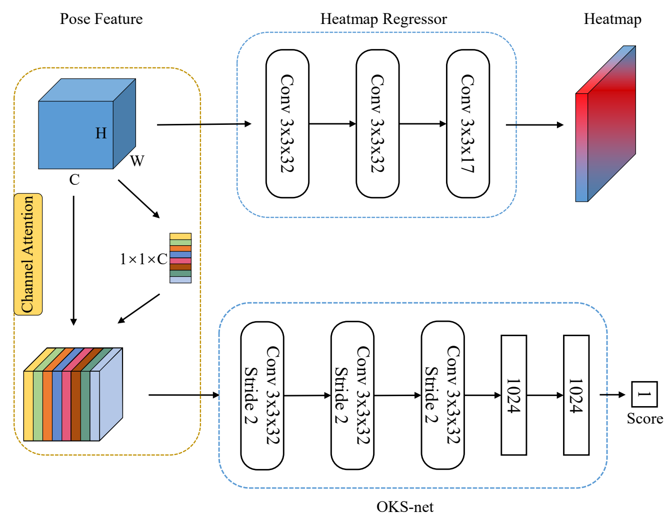


Fig. 8. The architecture of OKS-net. A channel attention mechanism is applied to the pose feature, then 3 convolutional layers with stride 2 are used for down-sampling. Finally, three fully connected layers are applied to produce the pose quality score. For the channel attention mechanism, the $1 \times 1 \times C$ feature is obtained through a max-pooling layer and a fully connected layer, then it is multiplied with the pose feature for re-weighting.

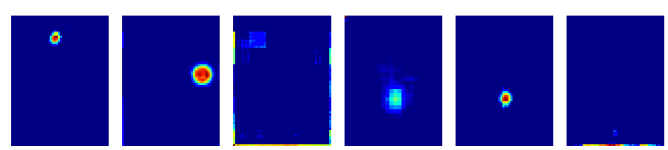


Fig. 9. Channel information of the pose feature. Six of thirty-two channels are taken as examples to illustrate that channels of the pose feature correspond to the type of keypoints, and the channel attention mechanism is able to make OKS-net pay more attention to visible keypoints.

ground truth OKS is computed by Eq. 1, given the annotation of the corresponding pose. The predicted keypoint locations are obtained from heatmap using the standard post-processing technique. The total loss is given as follows:

$$\mathcal{L} = \lambda_{heatmap}\mathcal{L}_{heatmap} + \lambda_{OKS}\mathcal{L}_{OKS} \tag{2}$$

where $\mathcal{L}_{heatmap}$ is heatmap loss, $\mathcal{L}_{OKS}$ is the loss for predicting the OKS, which we use l_1 loss. The loss weight $\lambda_{heatmap}$ is set to 2000, and λ_{OKS} is set to 0 or 1. We use λ_{OKS} to control the training process, and a detailed discussion is given in Section V-A.

We make $\mathcal{L}_{OKS}$ to be loss sensitive by giving more weights on hard examples. This is because the distribution of training samples to learn the OKS is long-tailed, as shown in Fig. 4(c). IoU-Net [12] and Mask Scoring RCNN [13] also learn the evaluation indicator (IoU) to get the score of detections, however, the training data is not so imbalanced in their problems (as shown in Fig. 5(b)). Therefore, we have to take special measures to tackle the long-tailed training data in learning the OKS. We enlarge the loss of the samples that get the largest 25 percent of the losses in each training batch as follows:

$$\mathcal{L}_{OKS} = e^{2\mathcal{L}_{OKS}} - 1 \tag{3}$$

Hence the hard samples can be fully explored during training, which may alleviate the imbalance in training data.

The inference procedure also conforms to the conventional pose estimation framework. For each input image, the network will produce K heatmaps of the pose and the corresponding prediction of the OKS. We use the heatmaps to choose the locations of K keypoints, and the predicted OKS is considered as the score depicting pose quality.

IV. EXPERIMENTS

The experiments are conducted thoroughly based on Hourglass [14], SimpleBaseline [15] and HRNet [9], various backbones under each framework and different image input sizes are utilized to present comprehensive experimental results. These experiments testify to the effectiveness of the proposed OKS-net.

A. Implementation Details

Dataset. We use the COCO dataset [10] for all the experiments. Following COCO 2017 settings, the train set includes $57K$ images and $150K$ person instances, the validation set and test-dev set contain $5K$ and $20K$ images respectively.

Evaluation metric. To judge how well the predicted OKS reflects pose quality, we use the correlation coefficient between the predicted OKS and their corresponding ground truth. The COCO keypoint evaluation metrics are utilized to demonstrate the proposed OKS-net can also improve the performance of human pose estimation. We report AP averaged over 10 OKS thresholds, AP^{50} and AP^{75} that the OKS threshold is 0.5 or 0.75, AP^{M} and AP^{L} for medium and large objects, and mean AR.

Training. The whole network is trained from scratch. Because OKS-net aims to predict the score of estimated poses, when the result of pose estimation (heatmap) is not stable, it will be meaningless to train OKS-net. We first set the loss weight of $\mathcal{L}_{OKS}$ as 0 in the beginning 200 epochs, thus a good backbone for pose estimation is able to be obtained. Then the loss weight of $\mathcal{L}_{OKS}$ is set as 1 in the following 50 epochs to train OKS-net. The whole training process goes for 250 epochs, the learning rate is decreased by a factor of 0.1 after 170 and 220 epochs respectively. For Hourglass, RMSprop with the base learning rate $2.5e-4$ is used as the optimizer. The Adam optimizer is applied in SimpleBaseline and HRNet with the base learning rate $1e-3$. The rest configurations just follow the standard settings [9].

TABLE I
THE CORRELATION COEFFICIENT BETWEEN PREDICTED SCORES AND THE GROUND TRUTH OKS. THE PREDICTION OF OKS-NET PERFORMS BETTER THAN THE SCORE OF HEATMAP POST-PROCESSING, NO MATTER WHAT BACKBONE NETWORK OR INPUT SIZE IS USED.

	HRNet-W32		HRNet-W48	
	256×192	384×288	256×192	384×288
Heatmap Post.	0.61	0.62	0.61	0.62
OKS-net	0.69	0.67	0.67	0.66

Testing. We use the common heatmap post-processing practice [9] to get keypoint locations. As flipping is utilized in test, we compute pose score by averaging the predicted OKS of original and flipped input.

B. Quantitative Results

How good is the predicted OKS. The comparison of correlation coefficients is given in Table. I. To do the comparison, we conduct experiments on COCO val2017 set using the ground truth bounding boxes. For the limit of space, here only the results based on HRNet are presented. Whether the backbone network is HRNet-W32 or HRNet-W48, OKS-net produces consistently reliable predictions that correlate very well with the ground truth. The correlation coefficient is around 0.67, regardless of the input size. The score obtained from heatmap post-processing does not make a poor correlation coefficient, however, it still underperforms the prediction of OKS-net by a clear margin. An illustration of the correlation between the predicted OKS and ground truth is given in Fig. 3(b). We can see that the predictions by OKS-net are very close to the ground truth, especially for the ones with the OKS larger than 0.85.

The improvement on COCO evaluation. Better pose score will bring noticeable improvement during COCO AP evaluation for the same provided detections. This section presents the improvement on AP, and demonstrate the superiority of the predicted OKS. Results on different frameworks including various backbones will be reported. And both of the widely-adopted input size 256×192 and 384×288 will be used in the experiment.

Results on the validation set. To avoid the influence of human detection on AP evaluation, we use the ground truth bounding boxes on COCO val2017 set. Table. II reports the results based on Hourglass [14]. Experiments are conducted on all the three backbones including 2, 4 and 8-stage Hourglass. The input size 384×288 is not used simply because it is not compatible with the original network setting. Compared to the conventional pose score, OKS-net improves AP by 0.9 points on 2-stage Hourgalss, 0.7 points on 4-stage Hourglass, and 0.6 points on 8-stage Hourglass. The results based on SimpleBaseline [15] are given in Table. III. Using the input size 256×192, OKS-net makes impressive gain by about 1.0 points consistently on backbones ResNet-50/101/152. The improvement under the input size 384×288 differs

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3005522, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II
COMPARISON OF AP EVALUATION RESULTS BASED ON HOURGLASS. THE INPUT SIZE IS 256×192.

Backbone	OKS-net	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
HG 2-stage		73.7	91.6	81.4	71.1	77.9	76.6
	√	**74.6**	**92.6**	**81.9**	**71.9**	**78.6**	**76.8**
HG 4-stage		75.8	92.6	82.7	73.3	79.9	78.6
	√	**76.5**	**92.7**	**84.0**	**74.0**	**80.8**	**78.7**
HG 8-stage		76.9	93.6	83.8	74.1	81.4	79.6
	√	**77.5**	**93.7**	**84.1**	**74.6**	**82.3**	79.6

TABLE III
COMPARISON OF AP EVALUATION RESULTS BASED ON SIMPLEBASELINE.

Backbone	Input size	OKS-net	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
ResN-50	256×192		72.4	91.5	80.4	69.7	76.5
		√	**73.5**	**92.6**	**80.9**	**70.4**	**77.9**
	384×288		74.1	92.6	80.5	70.5	79.6
		√	**75.1**	**92.7**	**81.9**	**71.3**	**80.8**
ResN-101	256×192		73.4	92.6	81.4	70.7	77.7
		√	**74.3**	92.6	**81.9**	**71.4**	**79.0**
	384×288		75.5	92.5	82.6	72.4	80.8
		√	**76.2**	**92.6**	**83.0**	**73.0**	**81.6**
ResN-152	256×192		74.3	92.6	82.5	71.6	78.7
		√	**75.0**	92.6	**82.9**	**71.8**	**79.8**
	384×288		76.6	92.6	83.6	73.7	81.3
		√	**77.1**	92.6	**83.9**	**73.8**	**82.1**

according to the backbone, however, OKS-net still achieves tangible improvements: 1.0 and 0.7 points respectively for the backbone ResNet-50/101. HRNet [9] is the state-of-the-art framework. But OKS-net is still able to boost the performance as shown in Table IV. For the backbone HRNet-W32, OKS-net brings more than 0.8 points gain regardless of the input size. AP increases less on the backbone HRNet-W48 by 0.6 points. It is interesting to see that OKS-net can produce even larger improvements than using a stronger backbone. Taking the input size 256×192, OKS-net achieves 77.5 AP based on the backbone HRNet-W32, outperforming 77.1 AP obtained by upgrading the backbone to HRNet-W48. Considering the computation cost of OKS-net is negligible in the whole network, this is a cost-effective way to improve performance on AP evaluation.

Results on the test-dev set. The detected human bounding boxes provided by SimpleBaseline [15] for test-dev set are utilized. We report improvements on AP evaluation achieved by using the OKS-net prediction as pose score in Table V. Compared to the results on the validation set, the gain may be relatively smaller because the score of bounding boxes also has an influence on AP evaluation. Nevertheless, OKS-net still produces stable improvements, no matter which backbone network or input size is used.

TABLE IV
COMPARISON OF AP EVALUATION RESULTS BASED ON HRNET.

Backbone	Input size	OKS-net	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
HR-W32	256×192		76.5	93.5	83.7	73.9	80.8
		√	**77.5**	**93.7**	**85.0**	**74.7**	**82.0**
	384×288		77.7	93.6	84.7	74.8	82.5
		√	**78.5**	93.6	**85.1**	**75.3**	**83.8**
HR-W48	256×192		77.1	93.6	84.7	74.1	81.9
		√	**77.7**	**93.7**	**85.0**	**74.8**	**82.9**
	384×288		78.1	93.6	84.9	75.3	83.1
		√	**78.7**	93.6	84.9	75.3	**84.1**

TABLE VI
COMPARISON OF AP EVALUATION RESULTS USING OKS-NET PREDICTING SCORES WITH OTHER SIMPLE RE-SCORING METHODS.

Backbone	Re-scoring	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Hourglass-4	NMS	73.5	89.5	80.9	70.2	80.0
	SoftNMS	73.5	89.0	80.7	70.2	79.9
	OKS-net	**74.1**	**89.9**	**81.6**	**70.6**	**80.6**
ResNet-101	NMS	71.4	89.3	79.3	68.1	77.1
	SoftNMS	71.5	88.9	79.2	68.1	78.1
	OKS-net	**71.9**	**89.5**	**79.6**	**68.4**	**79.0**
HRNet-W32	NMS	74.4	90.5	81.9	70.8	81.0
	SoftNMS	74.3	90.0	81.6	70.7	80.9
	OKS-net	**75.1**	**90.7**	**82.2**	**71.4**	**82.0**

C. Comparison with Other Simple Re-scoring Methods

Re-scoring detections to get higher AP in evaluation is widely used in tasks such as object detection and segmentation. Here we apply these simple re-scoring methods to pose estimation, and do the comparison with the proposed OKS-net. Specifically, for every two detected poses, we calculate their OKS to make the overlap. The conventional NMS (Non-maximum Suppression) is employed to re-score the detections, and this serves as the baseline method. Then we use soft-NMS [39] to replace the conventional NMS for re-scoring the detections. The proposed OKS-net is compared with these two simple re-scoring methods to demonstrate its superiority. For the comparison, we utilize the detected human bounding boxes provided in SimpleBaseline [15] on COCO val2017 set. Table VI gives the comparison results using the input size 256×192. Whether the backbone network is Hourglass, ResNet or HRNet, using the OKS-net predicting scores consistently produces higher APs. SoftNMS just performs on par with the baseline NMS method, which shows simple re-scoring methods are not as capable as OKS-net to refine the scores of estimated poses.

D. How the Predicted Scores Improve AP?

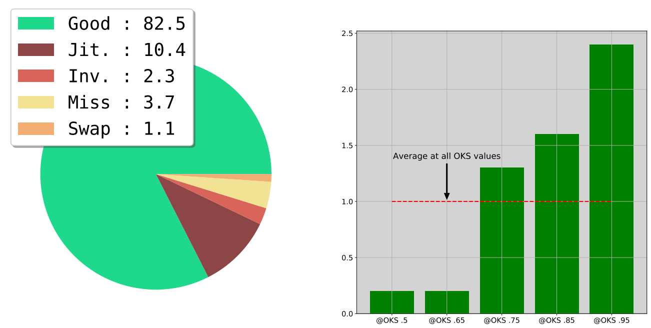
Ronchi et al. [43] provide a tool for error diagnosis in multi-instance pose estimation, which can analyze localization errors and scoring errors made by an algorithm. We use this tool

TABLE V
COMPARISON OF AP EVALUATION RESULTS ON THE COCO TEST-DEV SET.

Method	Backbone	Input size	OKS-net	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{M}	AR^{L}
Hourglass [14]	4-stage Hourglass	256 × 192		72.8	91.4	80.8	69.7	78.5	78.2	74.2	83.6
			√	**73.2**	**91.9**	**81.3**	**70.0**	**79.1**	**78.4**	**74.4**	**83.9**
SimpleBaseline [15]	ResNet-50	256 × 192		70.0	90.9	77.9	66.8	75.8	75.6	71.5	81.3
			√	**70.2**	90.9	**78.5**	**67.0**	**76.5**	**75.7**	71.5	**81.4**
		384 × 288		71.5	91.1	78.7	67.8	78.0	76.9	72.3	83.2
			√	**71.8**	91.1	**79.1**	**67.9**	**78.8**	**77.1**	**72.4**	**83.6**
HRNet [9]	HRNet-W32	256 × 192		73.5	92.2	81.9	70.2	79.2	79.0	75.0	84.5
			√	**73.9**	**92.4**	**82.3**	**70.7**	**79.7**	**79.3**	**75.4**	**84.7**
		384 × 288		74.9	92.5	82.6	71.3	80.9	80.1	75.9	85.8
			√	**75.2**	**92.7**	**83.0**	**71.7**	**81.2**	**80.4**	**76.3**	**86.0**

to see which errors are alleviated by the proposed OKS-net, and how the predicted score improves the performance in AP evaluation. The COCO 2017 validation set and ground truth bounding boxes are utilized for analysis. Fig. 10(a) shows the localization errors contained in the detections of HRNet-W32 under the input size 256×192. Using both the predicted scores by OKS-net and the original heatmap post-processing scores give the same localization errors. It is reasonable since the proposed OKS-net just corrects the score and does not aim to refine the heatmap of keypoints. Based on the same detection of keypoints, we exhibit the AP improvement obtained when using the scores predicted by OKS-net at five OKS thresholds in Fig. 10(b). The improvement can be seen at all OKS thresholds, and the advantage of the OKS-net predicting scores is larger at the higher OKS thresholds. Fig. 11 gives the scoring errors analysis provided by [43]. We compare the histogram of OKS-net predicting scores with both the ones of heatmap post-processing scores and the optimal scores. The red bars in each figure highlight how many detections have high OKS and low score or vice versa. Clearly, the OKS-net predicting scores manifest similar qualities with the optimal scores. From the histogram of the heatmap post-processing scores, we can see numerous detections with high OKS get low score. These mistakes are largely rectified in the OKS-net predicting scores, which brings the improvement in AP evaluation. There are more detections with high score and low OKS in the histogram of the OKS-net predicting scores than the optimal scores, this is caused by some typical mistakes made by OKS-net, which we will discuss in Section V-D.

Heatmap post-processing scores lack the ability to deal with the occlusion of poses as shown in Fig 6, Fig. 12 presents how the OKS-net predicting scores improve the PR curves on the four occlusion benchmarks. The OKS-net predicting scores reduce a lot of the scoring errors on the two mostly occluded benchmarks, and certainly the refinement is not obvious on the other two benchmarks with little occlusion, as the heatmap post-processing scores can do well in AP evaluation when keypoints are mostly visible. Thus combining the scoring errors analysis provided in Fig. 11, we can conclude that OKS-net is able to produce right scores for good detections of occluded poses, which are assigned with low confidence by the heatmap post-processing scores, and the correction increases the AP in evaluation.



(a) Localization errors analysis (b) Overall AP improvement

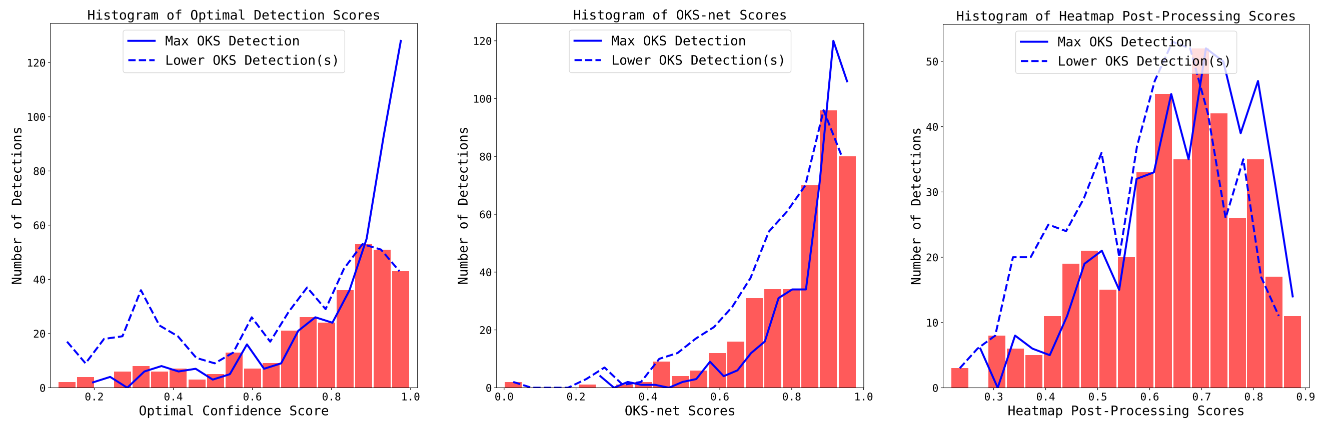
Fig. 10. Error analysis by [43]. (a) Distribution of localization errors containing jitter, inversion, miss, and swap. (b) Comparing to the heatmap post-processing scores, the AP improvement can be obtained when using the OKS-net predicting scores at five OKS evaluation thresholds.

E. Ablation Study

Different strategies of utilizing pose feature. In Section III-C, we apply the channel attention mechanism to the pose feature, which makes OKS-net deal with the occlusion of keypoints more effectively. Here, several other strategies of utilizing the pose feature are tested for comparison. Fig. 13 illustrates the employed strategies, and the details are explained as follows:

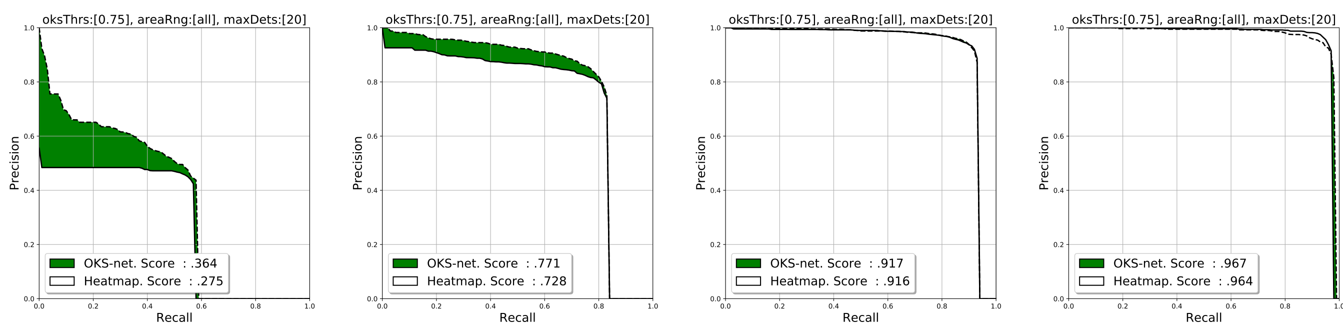
1) Pose feature only: OKS-net directly takes the pose feature to predict the OKS.
2) Channel attention: Pose feature is put into a channel attention layer, then the generated weight is multiplied with the pose feature. OKS-net takes the weighted feature as input.
3) Spatial attention: Pose feature is put into a spatial attention layer, then the generated weight is multiplied with the pose feature. The weighted feature is used as the input of OKS-net.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3005522, IEEE Transactions on Circuits and Systems for Video Technology



(a) The ground truth OKS (b) OKS-net predicting scores (c) Heatmap post-processing scores

Fig. 11. Scoring errors analysis by [43]. Each sub-figure separately plots the detections achieving the maximum OKS with a given ground-truth instance (continuous line) and the other detections achieving OKS at least .1 (dashed line). The red bars highlight how many detections have high OKS and low score or vice versa; a smaller count indicates an overall better score.



(a) 0-5 visible keypoints (b) 6-10 visible keypoints (c) 11-15 visible keypoints (d) 16-17 visible keypoints

Fig. 12. Comparison of PR curves at .75 OKS on four occlusion benchmarks. The COCO val2017 set is divided into four benchmarks according to the number of visible keypoints, which are 0-5, 6-10, 11-15 and 16-17, according to [43]. The white color depicts the AP obtained by heatmap post-processing scores, and the green color gives the improvement made by the OKS-net predicting scores.

TABLE VII
COMPARISON OF AP EVALUATION RESULTS USING DIFFERENT STRATEGIES OF UTILIZING THE POSE FEATURE.

Backbone	Strategies	AP	AP^{50}	AP^{75}
HRNet-W32	(1) Pose Feature Only	77.2	93.7	84.1
	(2) Channel Attention	77.5	93.7	85.0
	(3) Spatial Attention	77.2	93.7	84.0

TABLE VIII
COMPARISON OF AP EVALUATION RESULTS USING THE MODEL TRAINED WITH HARD SAMPLE MINING OR NOT.

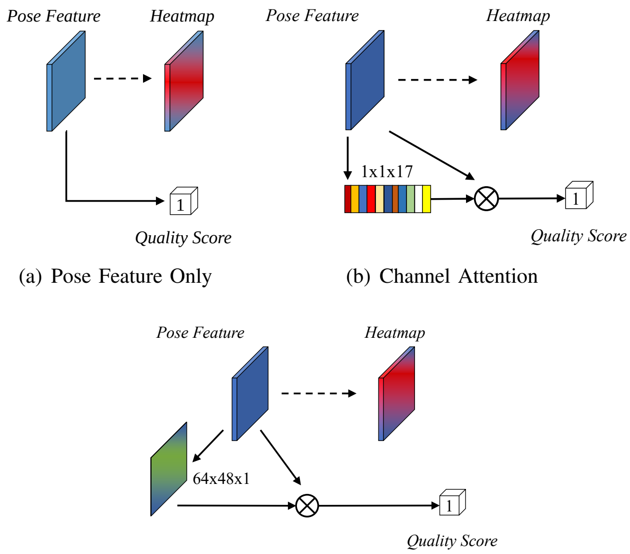
Backbone	Hard Sample Mining	AP	AP^{50}	AP^{75}
HRNet-W32		77.3	93.7	84.9
	√	77.5	93.7	85.0

Table VII shows the results of the three design choices. As can be seen, the employed channel attention mechanism produces the best performance. The spatial attention mechanism just produces the same result as only using the pose feature. This clearly demonstrates that taking care of the occlusion is useful, and the channel attention mechanism is capable of perceiving the occlusion.

Loss-sensitive hard sample mining. To tackle the long-tailed training data in learning pose score, $\mathcal{L}_{OKS}$ is set to be loss-sensitive by giving more weights to hard samples. We compare the results of training OKS-net using hard sample mining or not in Table VIII. Though the measure taken to mine hard examples is simple, it offers valuable aid in the training process. And better performance is able to be seen in testing by using the model trained with hard sample mining.

The value of the loss weight $\lambda_{heatmap}$. In Eq. 2 we set the loss weight $\lambda_{heatmap} = 2000$, because the value of the loss $\mathcal{L}_{heatmap}$ is less than a thousand of the loss $\mathcal{L}_{OKS}$, and a balance should be achieved between regressing heatmap and predicting the OKS. Getting accurate predictions of heatmap is the base of human pose estimation, predicting the OKS must be achieved without affecting the regression of heatmap. If the loss of heatmap $\mathcal{L}_{heatmap}$ is much smaller than $\mathcal{L}_{OKS}$, the training process will mainly focus on optimizing the sub-network of learning the OKS, and good estimation of heatmap cannot be guaranteed. We have tested setting different values

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3005522, IEEE Transactions on Circuits and Systems for Video Technology



(a) Pose Feature Only (b) Channel Attention

(c) Spatial Attention

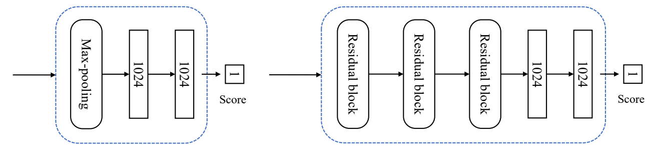
Fig. 13. Different design choices of OKS-net, which explores three strategies of utilizing pose feature.

TABLE IX
COMPARISON OF AP EVALUATION RESULTS USING DIFFERENT VALUES OF THE LOSS WEIGHT λ.

Backbone	$\lambda_{heatmap}$	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
HRNet-W32	1	75.3	93.6	82.9	72.5	80.2
	20	76.0	93.6	83.1	73.3	80.7
	200	76.9	93.6	84.0	73.8	81.7
	2000	77.5	93.7	85.0	74.7	82.0
	20000	77.4	93.7	85.0	74.7	82.1

for $\lambda_{heatmap}$ from 1 to 20000, poor performance is delivered by setting $\lambda_{heatmap}$ less than 200, and $\lambda_{heatmap} = 20000$ gets similar results with $\lambda_{heatmap} = 2000$. Table IX gives the detailed results for comparison. The experimental results demonstrate the importance of striking a balance between $\mathcal{L}_{heatmap}$ and $\mathcal{L}_{OKS}$, and the model can be robust to different values of $\lambda_{heatmap}$ by guarantying $\mathcal{L}_{heatmap}$ a little larger than $\mathcal{L}_{OKS}$.

Different structures of OKS-net. The proposed OKS-net utilizes three convolutional layers and two fully connected layers to predict the score, where the convolutional layers are used to down-sample the feature size. Here, we test whether a simpler structure with only one max-pooling layer for down-sampling can perform as well, or a more complex structure replacing the convolutional layers with residual blocks [34] may produce better results. These two compared structures are illustrated in Fig. 14. Table X gives the comparison results. The simpler structure just performs a little worse than the proposed OKS-net, which manifests the input pose feature contains enough high-level semantic information for predicting score. But using max-pooling drops some useful information, thus the performance is affected. Making the structure more complex does not bring better results, since there is no need to conduct deeper convolution on the pose feature. In addition, a more complex structure can be harder for optimization, which makes its performance worse than the proposed OKS-net.



(a) Simpler OKS-net (b) More complex OKS-net

Fig. 14. Two different structures from the proposed OKS-net (as shown in Fig. 8). We do the comparison to analyze the effect of different structures.

TABLE X
COMPARISON OF AP EVALUATION RESULTS USING DIFFERENT STRUCTURES OF OKS-NET.

Backbone	Structures of OKS-net	AP	AP^{50}	AP^{75}
HRNet-W32	Simpler (Max-pooling)	77.3	93.7	84.0
	The proposed OKS-net	77.5	93.7	85.0
	More Complex (Res. blocks)	77.3	93.7	84.1

V. DISCUSSION

In this section, we discuss some properties of the proposed OKS-net. Experiments are done based on the network HRNet-W32 using the input size 256×192, unless otherwise specified. And all results are acquired on COCO val2017 set using the ground truth bounding boxes.

A. The Training Strategy

There are several different strategies to train OKS-net, here we discuss the proper way to conduct the training. The whole network for heatmap regression and score prediction can be trained end to end, the following three strategies are studied for comparison. (1) The whole network is trained from scratch, both heatmap regression and score prediction take part in the training from the beginning. (2) Score prediction is not trained before getting a good backbone of heatmap regression (setting the loss weight of $\mathcal{L}_{OKS}$ as $\lambda_{OKS} = 0$), then it participates in the training process until the end. (3) The whole network is fine-tuned from a well pre-trained state-of-the-art human pose estimation model. The detailed settings of training using the first and second strategy are similar, which are given in Section IV-A. The third strategy is conducted by training for 50 epochs, and the base learning rate is set as $1e-4$, which is decreased by a factor of 0.1 after 20 epochs. Table XI presents the results obtained by utilizing the above three strategies. Clearly, the first strategy does not produce as good performance as the other two. This demonstrates the importance of getting a good backbone of heatmap regression before training OKS-net, because OKS-net needs stable pose estimation results to predict the score. Training the network from scratch utilizing the second strategy is able to deliver the same result as finetune, hence we employ this strategy in the paper.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3005522, IEEE Transactions on Circuits and Systems for Video Technology

TABLE XI
COMPARISON OF AP EVALUATION RESULTS USING DIFFERENT TRAINING STRATEGIES. BOTH STRATEGY (1) AND (2) TRAIN THE MODEL FROM SCRATCH, WE USE λ_{OKS} TO REPRESENT THE LOSS WEIGHT OF $\mathcal{L}_{OKS}$.

Backbone	Training Strategies	AP	AP^{50}	AP^{75}
HRNet-W32	(1) $\lambda_{OKS} = 1$ (Epoch 0)	76.7	93.6	84.0
	(2) $\lambda_{OKS} = 1$ (Epoch 200)	77.5	93.7	85.0
	(3) Fine-tune	77.5	93.7	85.0

TABLE XII
APS OBTAINED WITHOUT USING THE OKS-NET PREDICTING SCORES ON SIX OKS THRESHOLDS ON BOTH THE LARGE INPUT SIZE 384×288 AND THE SMALL INPUT SIZE 256×192.

Input size	AP^{50}	AP^{60}	AP^{70}	AP^{75}	AP^{85}	AP^{95}
256×192	93.5	91.4	88.0	83.7	69.8	22.0
384×288	93.6	91.5	88.1	84.8	71.0	28.1

B. The Discrepancy of Improvement on Different Input Sizes

From Table III and IV, we can see that the improvement made by using the OKS-net predicting scores on the large input size 384×288 may be a little smaller than on the small input size 256×192, the discrepancy in AP is about 0.2. Taking the model of HRNet-W32 for example, we carry out a detailed study into the discrepancy and find the reason may be as follows. Table XII gives the AP obtained without using the OKS-net predicting scores on six OKS thresholds on both input sizes. We can see the improvement made by enlarging the input size is mainly on the OKS thresholds bigger than 0.75. Thus on every OKS threshold bigger than 0.75, comparing to the small input size, there are more detections meeting the threshold on the large input size. Using the OKS-net predicting scores, the improvement obtained on both input sizes is presented in Table XIII. As can be seen, the improvement made by re-scoring on both input sizes is also made mainly on the OKS thresholds bigger than 0.75. Combining the analysis given in Section IV-D, it is convenient to conclude that for the detections with OKS bigger than 0.75 but getting low score, OKS-net corrects their scores and improves the AP. And the number of these wrongly scored detections will not be changed by enlarging the input size from 256×192 to 384×288, since as exhibited in Table XII enlarging the input size just makes the localization of already good detections (OKS $\geq .75$) more precise. However, the number of the detections meeting the OKS thresholds bigger than 0.75 is increased on the large input size. Hence according to the evaluation metric of AP, the improvement made using the OKS-net predicting scores will be a little smaller on the input size 384×288 than 256×192.

C. The Upper Bound Performance of OKS-net

Using the ground truth bounding boxes, we are able to calculate the ground truth OKS of each pose detection. Then, the upper bound AP can be got by replacing the OKS-net predictions with the ground truth OKS. Table XIV gives the results. We make the comparison both based on a strong model, HRNet-W48 under the input size 384×288, and a relatively weaker model, SimpleBaseline under the backbone ResNet-50 and the input size 256×192. Regardless of which model is used, OKS-net reduces the performance gap between the score obtained by heatmap post-processing and the ground truth OKS. Nonetheless, the predictions of OKS-net still have room to be improved, which are 1.9% AP and 1.8% AP for the weak and strong model respectively.

TABLE XIII
THE IMPROVEMENT OBTAINED ON APS BY USING THE OKS-NET PREDICTING SCORES ON BOTH THE LARGE INPUT SIZE 384×288 AND THE SMALL INPUT SIZE 256×192.

Input size	AP^{50}	AP^{60}	AP^{70}	AP^{75}	AP^{85}	AP^{95}
256×192	0.2	0.2	0.3	1.3	1.6	2.4
384×288	0	0.1	0.2	0.3	2.6	1.5

TABLE XIV
COMPARISON OF AP EVALUATION RESULTS USING OKS-NET PREDICTIONS AND THE GROUND TRUTH OKS AS POSE SCORE.

Method	Input size	Pose score	AP
SimpleBaseline ResNet-50	256×192	Heatmap Post	72.4
		OKS-net prediction	73.4
		Ground truth OKS	75.3
HRNet-W48	384×288	Heatmap Post	78.1
		OKS-net prediction	78.6
		Ground truth OKS	80.4

D. Failure Case Analysis

We investigate how the wrong predictions are made by OKS-net in this section. Fig. 15 presents some typical failure cases. The first two lines exhibit some examples that get high OKS predictions, but the ground truth OKS is actually very low. We can see that inversion (confusing between semantically similar parts belonging to the same person) and swap (confusion between semantically similar parts of different persons) are the main causes of this kind of error. It is hard for OKS-net to produce correct predictions when the model confuses the left and right parts of the body or targets on a wrong person (see the appendix for the cause of this error). Since there are quite a lot of inversion and swap errors as shown in Fig. 10(a), this explains why the OKS-net predicting scores produce more detections with high score and low OKS than the optimal scores, which is presented in the scoring errors analysis of Fig. 11. There are also some cases that are given low OKS predictions, but the ground truth OKS is high, as shown in the last line of Fig. 15. This mainly happens on input images with low quality. The model may not feel very confident with its estimation, however, the chosen location of keypoint is somehow close to the ground truth. These mistakes are fairly inevitable in complex situations, e.g. crowded with people, few visible keypoints or images of low quality.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3005522, IEEE Transactions on Circuits and Systems for Video Technology

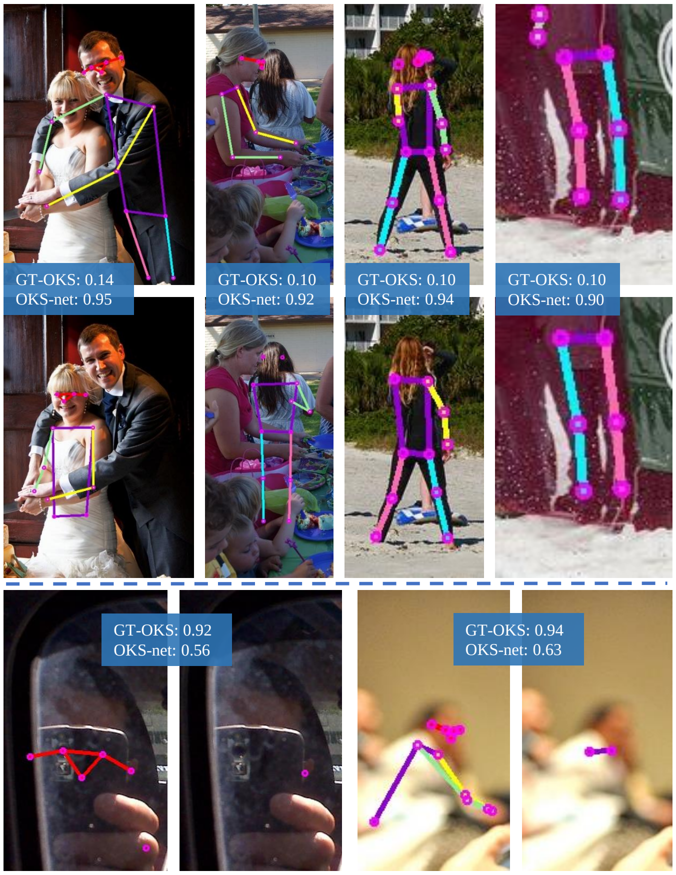


Fig. 15. Examples of failure predictions by OKS-net. In the figure, different colors represent different body parts. The first two lines show cases that get wrong predictions because of targeting on a wrong person (see the appendix for the cause of this error) or inverting the left and right body parts. The last line gives examples that obtain high OKS, however, OKS-net produces low scores because the image quality is poor. The ground truth poses are presented in the second line and the right part of the third line.

E. Model Size

The model size of OKS-net is fixed regardless of the backbone. It contains 2.7M and 4.7M parameters respectively under the input size 256×192 and 384×288. The increased GFLOPs of using OKS-net is 0.1 or 0.2 depending on the input size. Comparing to the model size and GFLOPs of the backbone, for example HRNet-W48 contains 63.6M parameters and has 32.9 GFLOPs under the input size 384×288, the burden of using OKS-net for learning pose quality is very limited.

VI. Conclusion

In this paper, we investigate how to make a pose estimation network build direct awareness of its detection quality. Our answer is to learn the quality from pose features in an end-to-end way, and a novel network block termed OKS-net is proposed, which regresses the object keypoint similarity between detections and their corresponding ground truth keypoints. Integrating OKS-net into the existing pose estimation models gains two real advantages: (i) the obtained quality score directly reflects the localization accuracy of estimated poses, making selecting usable poses simple; and (ii) Noticeable gain can be achieved by using the predictions of OKS-net as pose score during COCO AP evaluation. The effectiveness of OKS-net is demonstrated through extensive experimental results. We hope our work can facilitate the application of human pose estimation in high-level computer vision tasks and help future research.

Fig. 16. Wrong annotations in the COCO dataset. In each figure, we show the annotations containing bounding box, mask, and pose. As can be seen, the pose of these two samples is wrongly annotated onto another person.

Appendix A
Cause of Human Pose Detections Occurring on the Wrong Person

In Fig. 15 the first two samples show pose estimation results occurring on the wrong person. For example, in the first sample the ground-truth pose is on the women, but the pose detection result is on the man. The pose estimation method used is HRNet-W32 [9], all samples in Fig. 15 are from COCO 2017 validation set, and the ground truth bounding boxes are utilized. Thus there may be confusion how this error can happen. We have carefully checked the detection results and the ground truth poses of these two samples. We find that actually the error is not made by the pose estimation model, and the annotations of the samples are wrong. For the first sample, the bounding box and mask are all annotated on the man, however, the pose is annotated on the women. The same annotation error happens on the second sample. Since COCO is a very large dataset, some of such wrong annotations are inevitable. The annotations of these two samples given in the COCO dataset are shown in Fig. 16.

Acknowledgment

This work was supported by NSF of China (Grant Nos. 61802189, U1713208), NSF of Jiangsu Province (Grant Nos. BK20180464), the Fundamental Research Funds for the Central Universities (Grant Nos. 30918014107, 30919011280, 30920032202), the Young Elite Scientists Sponsorship Program by Jiangsu Province, and the Young Elite Scientists Sponsorship Program by CAST (Grant No. 2018QNRC001).

References

[1] H. Wang and L. Wang, "Cross-agent action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2908–2919, 2017.

[2] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[3] S. A. Ahmed, D. P. Dogra, S. Kar, and P. P. Roy, "Trajectory-based surveillance analysis: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 7, pp. 1985–1997, 2018.

[4] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 996–12 004.

[5] J. Lei, L. Niu, H. Fu, B. Peng, Q. Huang, and C. Hou, "Person re-identification by semantic region representation and topology constraint," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2453–2466, 2018.

[6] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2347–2356.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2011, pp. 1297–1304.

[9] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[10] 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 *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[11] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[12] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 784–799.

[13] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6409–6418.

[14] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[15] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 466–481.

[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[17] L. Zhao, X. Gao, D. Tao, and X. Li, "Tracking human pose using max-margin markov models," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5274–5287, 2015.

[18] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 588–595.

[19] L. Zhao, X. Gao, D. Tao, and X. Li, "A deep structure for human pose estimation," *Signal Processing*, vol. 108, pp. 36–45, 2015.

[20] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[21] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proceedings of the Advances in neural information processing systems*, 2014, pp. 1799–1807.

[22] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Proceedings of the Advances in neural information processing systems*, 2014, pp. 1736–1744.

[23] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 190–206.

[24] U. Iqbal, A. Milan, and J. Gall, "Posetrack: Joint multi-person pose estimation and tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2011–2020.

[25] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2334–2343.

[26] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4929–4937.

[27] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Advances in Neural Information Processing Systems*, 2017, pp. 2277–2287.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[29] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[30] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1246–1259, 2018.

[31] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1831–1840.

[32] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1281–1290.

[33] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1212–1221.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7103–7112.

[36] K. Su, D. Yu, Z. Xu, X. Geng, and C. Wang, "Multi-person pose estimation with enhanced channel-wise and spatial information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5674–5682.

[37] S. Zhang, G. Qi, X. Cao, Z. Song, and J. Zhou, "Human parsing with pyramidical gather-excite context," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[38] K. Lin, L. Wang, K. Luo, Y. Chen, Z. Liu, and M. Sun, "Cross-domain complementary learning using pose for multi-person part segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[39] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5561–5569.

[40] L. Tychsen-Smith and L. Petersson, "Improving object localization with fitness nms and bounded iou loss," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6877–6885.

[41] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, "Revisiting rcnn: On awakening the classification power of faster rcnn," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 453–468.

[42] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[43] M. Ruggero Ronchi and P. Perona, "Benchmarking and error diagnosis in multi-instance pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 369–378.

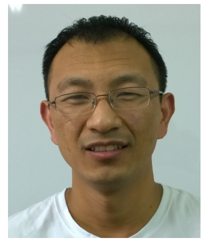
[44] J. Hu, S. Li, and S. Gang, "Squeeze-and-excitation networks," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

Lin Zhao received his B.E. degree from Xidian University in 2010, and the PhD degree in pattern recognition and intelligent systems form Xidian University in 2017. He is currently a lecture with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include human related computer vision tasks, such as human pose estimation and tracking, 3D human shape reconstruction, and anomaly detection.

Jie Xu received his bachelor's degree from Nanjing University of Science and Technology(NJUST), on the subject of network engineering in 2017. He is currently working toward the Ph.D degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology(NJUST). His research interests mainly include pattern recognition, image processing, and computer vision.

Chen Gong received his B.E. degree from East China University of Science and Technology (ECUST) in 2010, and dual doctoral degree from Shanghai Jiao Tong University (SJTU) and University of Technology Sydney (UTS) in 2016 and 2017, respectively. Currently, he is a full professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests mainly include machine learning, data mining, and learning-based vision problems. He has published more than 90 technical papers at prominent journals and conferences such as IEEE T-PAMI, IEEE T-NNLS, IEEE T-IP, ICML, NeurIPS, CVPR, AAAI, IJCAI, etc. He also serves as the reviewer for more than 20 international journals such as AIJ, IEEE T-PAMI, IEEE T-NNLS, IEEE T-IP, and also the SPC/PC member of several top-tier conferences such as ICML, NeurIPS, CVPR, AAAI, IJCAI, ICDM, AISTATS, etc. He received the "Excellent Doctorial Dissertation" awarded by Shanghai Jiao Tong University (SJTU) and Chinese Association for Artificial Intelligence (CAAI). He was enrolled by the "Young Elite Scientists Sponsorship Program" of Jiangsu Province and China Association for Science and Technology. He was also the recipient of "Wu Wen-Jun AI Excellent Youth Scholar Award".

Jian Yang received the PhD degree from Nanjing University of Science and Technology (NUST), on the subject of pattern recognition and intelligence systems in 2002. In 2003, he was a Postdoctoral researcher at the University of Zaragoza. From 2004 to 2006, he was a Postdoctoral Fellow at Biometrics Centre of Hong Kong Polytechnic University. From 2006 to 2007, he was a Postdoctoral Fellow at Department of Computer Science of New Jersey Institute of Technology. Now, he is a Chang-Jiang professor in the School of Computer Science and Technology of NUST. He is the author of more than 200 scientific papers in pattern recognition and computer vision. His papers have been cited more than 6000 times in the Web of Science, and 15000 times in the Scholar Google. His research interests include pattern recognition, computer vision and machine learning. Currently, he is/was an associate editor of Pattern Recognition, Pattern Recognition Letters, IEEE Trans. Neural Networks and Learning Systems, and Neurocomputing. He is a Fellow of IAPR.



Wangmeng Zuo (M'09-SM'14) received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, China, in 2007. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology. He has published over 100 papers in top-tier academic journals and conferences. His current research interests include image enhancement and restoration, image generation and editing, visual tracking, object detection, and image classification. He has served as a Tutorial Organizer in ECCV 2016, an Associate Editor of the IET Biometrics, The Visual Computers, Journal of Electronic Imaging, and the Guest Editor of Neurocomputing, Pattern Recognition, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Transactions on Neural Networks and Learning Systems.

Xinbo Gao (M'02-SM'07) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a research fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a post-doctoral research fellow at the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been at the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education of P. R. China, a Professor of Pattern Recognition and Intelligent System, and the President of Chongqing University of Posts and Telecommunications. His current research interests include Image processing, computer vision, multimedia analysis, machine learning and pattern recognition. He has published six books and around 300 technical articles in refereed journals and proceedings. Prof. Gao is on the Editorial Boards of several journals, including Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is a Fellow of the Institute of Engineering and Technology and a Fellow of the Chinese Institute of Electronics.