

Human Pose Estimation based on HRNet and Feature Pyramids

Fangjie Dong^{1, †}, Miaowei Yang^{2, †, *} and Pengshen Yang^{3, †}

¹ School of Computer Science and Engineering, Shandong University of Science and Technology, Shandong, China

² School of Electronic and Information Engineering, Hebei University of Technology, Tianjin, China

³ School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

* Corresponding author email: 193841@stu.hebut.edu.cn

[†]These authors contributed equally

Abstract. A popular area of research in the realm of computer vision is human pose estimation, which is the process of recovering human joint points from the given images or videos. Convolutional neural networks, which have great feature representation capabilities, have become a fundamental component of human pose estimation algorithms as a result of the deep learning field's quick development. To enhance feature quality and raise the precision of human pose estimate, in this paper, we propose a new model based on HRNet neural network and feature pyramid to more accurately capture each main part of the human body. The model uses HRNet as the backbone network, taking advantage of its ability to combine low resolution with high resolution, and afterwards adds characteristic pyramids to the HRNet network in order to enhance the capability of detecting small objects by utilizing its ability to solve multi-scale problems. The model can estimate a human's pose accurately, according to experimental results, and on the MPII Human Pose dataset, the addition of the feature pyramid enhances the model's detection performance by 0.05. In order to achieve smoothing after adding feature mapping with different resolutions, the detection performance improves by 0.17 after further improvement by adding 3×3 convolution. For the elbow, it improves by almost 0.5. The model's ability to increase the precision of human pose estimate is supported by all of the outcomes.

Keywords: Pose Estimation; HRNet; Image Inspection; FPN Feature Pyramid.

1. Introduction

Pose estimation is the basis for many high-level semantic tasks and downstream application scenarios. The aim is to estimate the position of the "object of observation", so estimation is based on the "human body", with key points being the head, left hand, right foot, etc. The Human Pose Estimation task generally refers to a single-purpose estimation task, that is, to predict the position of each key point only through a 2D RGB image. In addition, image-based pose estimation will be further extended to video-based pose estimation if the human pose estimation considers the temporal flow information. At present, Human pose estimation has been widely used in film and animation, virtual reality, human-computer interaction, video surveillance, medical rehabilitation, unmanned vehicles, motion capture analysis and other fields [1].

Thanks to the rapid development of convolutional neural networks, the human body pose estimation method based on deep learning have become the mainstream framework of research. Deep learning-based methods are more robust, and usually use globally observed features without recognizing the local features of objects. The inability to obtain dense sampling and the difficulty in ensuring the accuracy and continuity of estimates are fundamental difficulties that cannot be overcome by learning-based pose estimation methods. According to the difference in algorithm design ideas, it can be mainly divided into two types: top-down [2], [3] and bottom-up [4]. The top-up method first detects the human body and then estimates the human body pose. The bottom-up approach is to first detect the position of each key point of each person, and then use the human body model to combine these key points, or use other algorithms to group them. The larger the number of

people in multi-person pose estimation, the greater the calculation amount of the top-down method, while the calculation amount of the bottom-up method will remain relatively stable. However, when the human body in the image has a large area of overlap, the robustness of the bottom-up method will be reduced.

With the goal of improving detection performance, in this paper, we choose a more advantageous network structure, and then improve and refine it from the aspect of feature fusion to improve detection performance. For the representative network such as Alexnet [5], VggNet [6], GoogleNet [7], ResNet [8], the obtained feature representations are low resolution. In contrast, we choose the more advanced and advantageous High-Resolution Net (HRNet) [9]. The HRNet principle is to change the series connection between different resolutions to a parallel connection, always keeping the high-resolution features. Since the feature fusion of different resolution feature maps in HRNet is done by direct adding, we introduce the feature pyramid to fuse features adaptively to improve the accuracy. The basic idea of Feature Pyramid Networks (FPN) [10] is to train and test a model by constructing a series of images or feature maps at different scales, with the aim of improving the robustness of the detection algorithm for detection targets of different sizes. However, if FPNs are computed directly from the original definition, they incur a large computational overhead. To reduce the computational effort, FPN uses a multi-scale feature fusion approach that can significantly improve the scale robustness of the feature representation without significantly increasing the computational effort. So that the semantic information of the same path of low resolution is fused into the output of high resolution, thus improving the accuracy of the final result.

Focusing on the above issue, the assignment of this paper is as following. In Section 2, we will introduce the method in detail. We then report and analyze the experiment results in Section 3. We also discuss the existing research problems in Section 4.

2. Method

We propose to choose the HRNet as the network structure and further added the feature pyramid network (FPN) to improve the feature quality as well as the detection performance. The key modules of our method will be introduced in the following subsection in detail.

2.1 HRnet

The initial layer is obtained by first passing two successive down-sampled convolutional layers, which are down-sampled by a total of four times, and then HRNet is applied to a series of Transition structures and Stage structures, with each Transition structure adding a new scale branch. Each Transition is based on the output of its corresponding Layer by combining two convolutional layers with a convolutional kernel size of 3×3 , resulting in two different scale branches. A new scale with a higher downsampling multiplier is added to the original two scale branches in the next Transition, with the top one remaining at high resolution. The top two channels are directly connected to the next layer, and a third channel with lower resolution is added to gradually introduce the lower resolution features into the higher resolution channel.

If the Transition structure of the HRNet network is considered a multiplication, then the Stage structure is the summation of the multiplied results and is undoubtedly the most important structure. In Transition 2, a new downsampled 16x scale is added to the original two scale branches (the top one is kept at high resolution, the original top two channels are directly connected to the next layer, and a third channel at low resolution is added to gradually introduce low resolution features into the high-resolution channel). Here the downsampled 16x scale is obtained directly from the downsampled 8x scale by a convolutional layer with a convolutional kernel size of 3×3 in step 2. The Transition2 is given by fusing the feature layers at different scales (the red rectangular box in the figure below). However, according to the source code implementation, there is only one 3×3 convolutional layer without fusion of different scales, including other code repository implementations of HRNet. The

information at different scales is fused according to the scales obtained from the Transition, i.e., for each scale branch the output is obtained by fusing the output from all branches.

2.2 Feature Pyramid Network

When HRNet is simply applied to real life, some drawbacks and shortcomings become apparent, such as the inability to avoid a certain probability of obscuring important parts of the object of study or incomplete feature extraction during motion. Improvements to the former require data augmentation, such as on-the-fly rotation, random scaling, random horizontal flipping, half-body, and other methods. Some cropping of the target to avoid full masking of important parts, GridMask is at most partially masked and will almost certainly be partially masked. The approach used is to align square regions for masking; for the latter the improved approach requires feature fusion of the originally captured features, adding FPN to allow the semantic information from the low-resolution pathway to be fused to the high-resolution output. The latter, i.e., feature fusion using the addition of an FPN feature pyramid, was ultimately chosen as the next experimental approach.

In the design of FPN, for deep convolutional networks, convolution from one feature layer to another, regardless of whether the step size is 1 or 2 or more, the convolution kernels are spread across the whole picture for convolution. Large targets take up more pixel points than small targets, so large targets are passed through the convolution kernels far more often than small targets, so in the next feature layer, there will be more response to the characteristics of large targets. In particular, in the case where the step size is greater than or equal to 2, features of targets that occupy a larger proportion of the frame are more likely to be retained, while feature points of targets that occupy a smaller proportion are easily overlooked.

3. Experiments

3.1 Dataset

MPII human pose dataset is used for evaluating our methods about articular human pose estimation. The data set included approximately 25K of images, including more than 40K of people annotating body joints. For all of the 40K subjects, 12K is for the test and the rest is for the training set.

We use data augmentation includes random scale, random rotation and random flipping. Adam optimizer is adopted over all process. The learning rate is set as $5e-3$ for both two experiments. The use of pre-trained model greatly improves our training speed and the training process is terminated within 30 epochs.

3.2 Evaluation Metric

PCK is used to measure the accuracy of single-person pose estimation. PCKh@0.5 indicates that, using the head length as a reference, if the normalized distance exceeds the threshold value 0.5, the prediction is considered correct.

3.3 Model Inference

Instead of detecting the person instance by detectors, the provided person boxes are used as the first stage results of our two-stage top-down paradigm, and then predict detection key points using the heatmap representation. The heatmap-based method has been widely used for years due to its excellent performance. The heatmap is a 2D Gaussian distribution generated at the center of each key point. Each key point on the heatmap represents its probability as the ground truth point. We compute the heatmaps of two kinds of images, the original and the flipped, and then average them as the new heatmap. For comparing the predicted heatmaps and the ground-truth heatmaps, we use the MSE loss. After that, the direction of the highest heat-value point from the highest score to the second highest score is adjusted with an offset of 0.25 to predict the location of each key point. False positives can

be suppressed in this way and the training process can be smoothed by assigning probability values to each position.

3.4 Results

3.4.1 HRNet twith FPN-like Structure

Indeed, HRNet itself is constantly fusing features of different resolutions, and the authors also believe that in this process, the high-resolution features have merged the information of the low-resolution features, so final direct output is the high-resolution branch. However, we believe that the output features with relatively low-resolution also contain certain semantic information, which may affect the subsequent prediction, so they should also be reflected in the output. In the previous HRNet structure, the fusion method was up-sampled n times by interpolation method and then added directly, which may lead to insufficient feature fusion. Here, we use an FPN (feature pyramid network)-like structure to fuse features layer by layer, so as to obtain a more fully integrated feature map. Therefore, on the premise of avoiding small targets being ignored, it also improves detection performance. Our improved HRNet achieves a 90.38 score, improving the detection performance by 0.05. More detailed results for different positions can be seen in Table 1.

Table 1. PCKh@0.5 results of the HRNet with FPN-like structure

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
HRNet	97.10	95.94	90.34	86.45	89.10	87.08	83.28	90.33
Ours(lr=1e-4)	97.10	96.01	90.39	86.45	89.04	87.02	83.47	90.36
Ours(lr=5e-5)	97.16	96.15	90.46	86.50	89.06	86.94	83.35	90.38

3.4.2 Effects of the Features in FPN

Due to the limited improvement of the direct application of the feature pyramid structure, after comparing the network structure, we find that the direct addition of features with different resolutions leads to inadequate feature fusion, which may be the key to limited semantic information fusion. For the purpose of small target detection, the feature with the highest resolution is finally selected for prediction. Therefore, the above FPN-like structure is improved, and 3×3 convolution is added to achieve the purpose of smoothing after the addition of the feature maps with different resolutions, that is, before each up-sampling. The results are as follows in Table 2 and there's a lot more improvement than just applying the feature pyramid structure directly.

Table 2. Results on the test set. Table shows the PCKh@0.5 results after adding the 3×3 convolution, which is a huge improvement over what it was before.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Ours(lr=1e-4)	97.10	96.01	90.39	86.45	89.04	87.02	83.47	90.36
Ours(lr=5e-5)	97.17	96.15	90.46	86.50	89.06	86.94	83.35	90.38
Ours(lr=1e-4)	97.20	96.15	90.64	86.50	89.27	86.94	83.47	90.46
Ours(lr=5e-5)	97.31	96.21	90.81	86.55	89.20	87.08	83.32	90.50

Our improved HRNet achieves a 90.48 score, improving the detection performance by 0.17 in total. For the part of the elbow, it improves almost by 0.5. In the course of our experiments, we found that, compared to the head, shoulder, and other parts, the score of the wrist and hip improved less. We found the following image, which may prove that the wrist position is more difficult to predict than the head position. As shown in Figure 1, people often show their heads, but wrist joint identification is often wrong if they are standing with others, in a particular pose, or if they have covered with indistinguishable clothing. The missing part of the human body will make the algorithm unable to detect the occluded human body node, and thus unable to give a correct prediction of the occluded node. Because the body's joints are interrelated, the loss of some joints will also have an impact on the prediction of open joints. Therefore, it is not only required that the algorithm can correctly detect

the unobstructed nodes, but also required that the algorithm has the ability to extract more favorable features and make reasonable predictions of the unobstructed nodes.

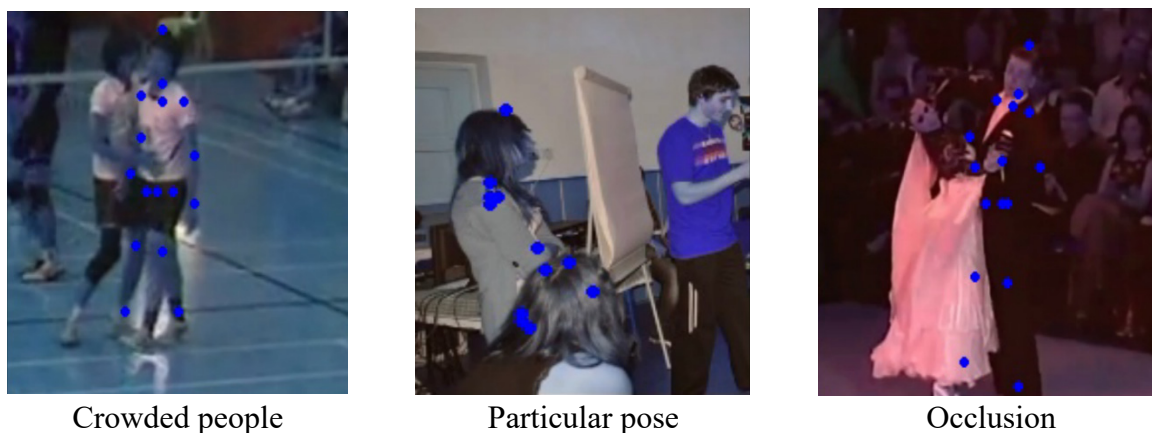


Fig 1. Detection results for different scenes

That's why our improved HRNet gets good performance. Simple targets can be identified by high-resolution features, while complex targets can be identified by deep features. For small targets, although the small-scale feature maps contain more semantic information, they cannot provide more resolution information. If more large-scale feature maps are fused, we can obtain more useful resolution information which can help us detect small targets and understand complex targets more. That's why our algorithm can detect complex scenes, special poses, and occlusions of limbs below in Figure 2.

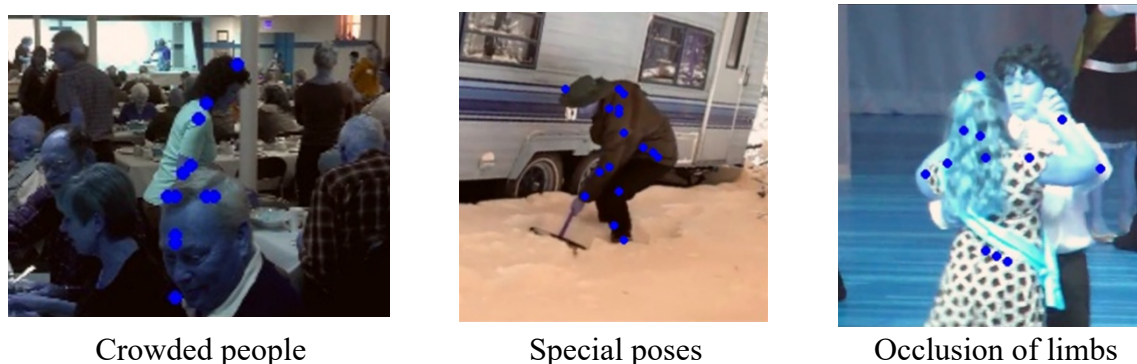


Fig 2. Detection results for more challenging scenes

4. Discussion

By reproducing the HRNet, a relatively high similarity can already be obtained. Then for the improvement of HRNet, and combined with the situation that will occur in real-life applications, two deficiencies of HRNet are considered: the situation where important parts of the study object are covered or the situation where feature extraction is not comprehensive. For the former improvement, data augmentation is required, and for the latter improvement, the option of adding an FPN feature pyramid is required to allow low-resolution semantic information to be fused to the high-resolution output, rather than adding it directly as in HRnet, thus avoiding the loss or omission of small targets.

No matter whether the step size is 1 or 2 or more, the convolution kernel must be distributed over the entire image in a generic network convolution from one feature layer to another. The convolution kernel is passed through the large targets much more times than the tiny targets since the large targets occupy more pixel points than the small ones, so in the next feature layer, there will be more responsive to the characteristics of the large targets. The characteristics of large targets are more likely to be kept whereas the feature points of small targets are simply bypassed, especially when the step

size is higher than or equal to 2. And combining HRNet with FPN can handle multi-scale objects and extract multi-scale features, which can avoid small targets from being ignored and also improve the detection performance.

5. Conclusion

An enhanced solution based on HRNet model is suggested for the issue of improving the accuracy of human posture estimation. The module of the feature pyramid is added after the HRNet network neural, and the single-resolution feature map is replaced by the multi-resolution feature of FPN to improve the detection accuracy. The experimental findings on the MPII Human Pose dataset finally demonstrate that the strategy presented in this paper significantly increases the accuracy of human pose estimation. After adding the use of FPN feature pyramids to the HRNet model, the detection performance of the model has a relatively obvious improvement, and we can have a 0.03 detection performance improvement for each feature part after improvement. And for the purpose of small target detection, after adding the 3×3 convolution module before upsampling, the experimental results have more improvement compared to the direct use of the feature pyramid, with an overall improvement of 0.17, while some parts improve by 0.5, which proves that the method has better accuracy improvement for human pose estimation detection, and the accuracy of the detection results from multiple experiments also proves that the method has better stability.

References

- [1] Li Yanping, Liu Ruyi, Wang Xiangyang, Wang Rui. Human pose estimation based on lightweight basicblock [J]. Machine Vision and Applications, 2022, 34(1).
- [2] Zhang Xiaona, Wu Qingtao. Top-down human pose estimation algorithm based on deep learning[J]. Electronic Measurement Technology, 2021, 044(009).
- [3] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In ECCV, pages 472–487, 2018.
- [4] Li Jia. Research on bottom-up multi-person attitude estimation Method[D]. University of Science and Technology of China, 2021.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1106–1114, 2012.
- [6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
- [7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In CVPR, pages 1–9, 2015.
- [8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.
- [9] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In CVPR, 2019.
- [10] T. Lin, P. Dollar, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In CVPR, pages 936–944, 2017.