MULTI-SCALE STACKED HOURGLASS NETWORK FOR HUMAN POSE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

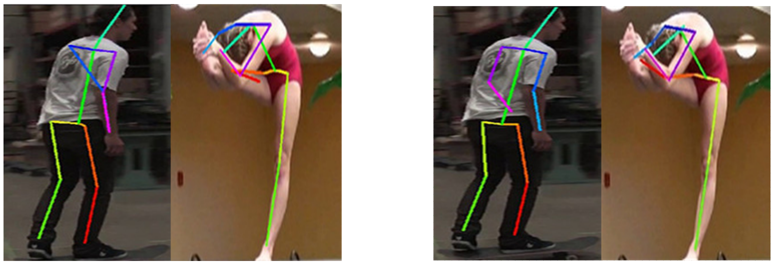
Stacked hourglass network has become an important model for Human pose estimation. The estimation of human body posture depends on the global information of the keypoints type and the local information of the keypoints location. The consistent processing of inputs and constraints makes it difficult to form differentiated and determined collaboration mechanisms for each stacked hourglass network. In this paper, we propose a Multi-Scale Stacked Hourglass (MSSH) network to highlight the differentiation capabilities of each Hourglass network for human pose estimation. The pre-processing network forms feature maps of different scales, and dispatch them to various locations of the stack hourglass network, where the small-scale features reach the front of stacked hourglass network, and large-scale features reach the rear of stacked hourglass network. And a new loss function is proposed for multi-scale stacked hourglass network. Different keypoints have different weight coefficients of loss function at different scales, and the keypoints weight coefficients are dynamically adjusted from the top-level hourglass network to the bottom-level hourglass network. Experimental results show that the proposed method is competitive with respect to the comparison algorithm on MPII and LSP datasets.

1 INTRODUCTION

Human pose estimation need locate the body keypoints (head, shoulder, elbow, wrist, knee, ankle, etc.) from the input image, and it is basic method for some advanced vision task Gong & Zhang (2016), such as human motion recognition, human-computer interaction, and human reidentification et al. We focus on single-person pose estimation problems in a single RGB image. Due to the high flexibility of the human body and limbs, diverse viewpoint, camera projection transformation and occlusion, it still is a difficult task to accurately determine body keypoints from a single image.

In recent years, the deep convolutional neural network (DCNN) has made significant progress in the human pose estimation; especially the stacked hourglass network Newell et al. (2016) has achieved good results and has attracted much attention. The human pose estimation involves two kinds of information: the type and location of body keypoints. The type of body keypoints needs to be determined in a larger receptive field, and the location of body keypoints needs to be based on the specific pixel position, which are respectively equivalent to global information and local information. The hourglass network uses a convolution layer and a deconvolution layer to form an hourglass structure, and establish crossover channels between convolution and deconvolution layers on different scales. Using the hourglass network for human pose estimation, the hourglass structure extracts global information through information compression, and the crossover channels compensates for local information loss in information compression. The stacked hourglass network continuously improves the human pose estimation by enhancing the context constraints among body keypoints through the stacked structure.

The stacked hourglass network theoretically increases the stacked depth to expand the receptive field and form a stronger context constraints among body keypoints. However, in practical applications, simply increasing the stacked depth is difficult to effectively improve the accuracy of human pose estimation. The main reason is that the consistent processing of inputs and constraints makes it difficult to form differentiated and determined collaboration mechanisms for each stacked hourglass



(a) (b)

Figure 1: (a) the stacked hourglass network may produce erroneous estimates due to occlusion and body distortion. (b) The MSSH network can better constrain the body keypoints under the condition of occlusion and body distortion.

network, to make up the information loss caused by the functional consistency of hourglass networks. Inspired by multi-scale information fusion from the single hourglass network, we propose a Multi-Scale Stacked Hourglass (MSSH) network. A pre-processing network consisting primarily of residual networks is designed to generate multi-scale features, and features of each scale are sent to different stacked hourglass networks. Each hourglass network has different inputs: the output of the pre-level hourglass network, and the received feature. Small-scale feature input makes the hourglass network tend to focus on global information, such as the type and context constraints of body keypoints, and large-scale feature input makes the hourglass network more likely to focus on local information, such as the location of body keypoints. The input of entire stacking hourglass network has changed from small-scale feature to large-scale feature, and it is a top-down route of body keypoints estimation. The iterative relationship between the loss functions of different scales of the MSSH network is established, so that the pre-level detection results affect the weight of the next keypoint loss function, and the optimization process of network model training is controlled by adaptive weighted loss function. The iterative relationship of the adjacent network loss function is established on the MSSH network, so that the weight of the loss function on pre-level hourglass network affects the weight of the loss function on the current hourglass network. The optimization process of the model training is controlled by the adaptive weighted loss function.

The main contributions of this paper can be summarized as follows:

- In a multi-scale stacked hourglass network, the pre-processing network generates features of different scales and dispatchs them to every hourglass network, where the small-scale features reaches the front of stacked hourglass network and the large-scale features reaches the rear of stacked hourglass network. From global information to local information, each hourglass network can form a differentiated function, which is conducive to the formation of collaborative processing.
- A new loss function of the MSSH network is proposed. The weighting coefficient of the loss function in hourglass network is defined. The weight coefficient and the loss function of the pre-level hourglass network are used to adjust the weight coefficient of the current hourglass network, and the convergence process of the model training is optimized by the adaptive weighted loss function.

The remainder of this paper is organized as follows. Section 2 briefly reviews recent work on human pose estimation. Section 3 details the structure of the MSSH network. Section 4 describes the new loss function for MSSH network networks. Section 5 describes the implementation details and experimental results. Section 6 summarizes our paper.

2 Related work

DeepPose Toshev & Szegedy (2014) firstly introduce CNN to solve pose estimation problem, which proposes a cascade of CNN to deal with pose estimation. Joint train approach Tompson et al. (2014) attempts to predict the keypoints of heatmaps of using CNN and graphical models. Most late work shows good performance by using a deep convolutional neural network to generate heatmaps of keypoints. The Convolutional Pose Machines Wei et al. (2016) uses a sequential convolutional architecture to express context relationships and uses multiple scales to process input feature maps. The hourglass network Newell et al. (2016) reuses bottom-up and top-down strategies and stacks up several hourglass modules to inference the keypoints of the human body.

On the basis of the hourglass network, In order to obtain better human body pose estimation results, researchers are more inclined to design more complex networks from multi-stage processing, the multi-scale feature and loss function. For multi-stage processing, Multi-Context Attention network Chu et al. (2017) uses the stacked hourglass network to generate attention maps with different resolution, and use CRF to enhancement the association of adjacent regions in the attention map. Pyramid structure Yang et al. (2017) increases the receptive field of the network through the complication of the building block, which enhance the deep convolutional neural networks using multi-scale subsampling to learn the features of different resolutions. For multi-scale feature, cascaded pyramid network Chen et al. (2017) suggests that GlobalNet predicts the keypoints on multi-scale feature map in the first-stage and makes RefineNet predict the online hard keypoints in the second phase. The multi-scale structure-aware network Ke et al. (2018) improves the detection result using multi-scale supervision and regression by matching features across all scales in building block. For loss function, deep consensus voting Lifshitz et al. (2016) adds voting constraints to the loss function. Recurrent human pose estimation Belagiannis & Zisserman (2017) uses multiple regression networks to generate multiple loss functions to optimize network. A novel bone-based part representation Tang et al. (2018) is proposed to avoid potentially large state spaces for higher-level parts through multi-scale loss function.

The proposed MSSH network draws on the concept of multi-scale, uses pre-processing networks to form feature maps of different scales, and assigns them to different locations in the stacked network, which is inspired by chain prediction Gkioxari et al. (2016). The small size features reach the front of stacked hourglass network, and the large-scale features reach the rear of stacked hourglass network. A new loss function is designed to dynamically adjust the keypoints weight coefficients from the top layer to the bottom layer, and it pay more attention to hard keypoints in multi-scale stacked hourglass networks.

3 Structure of multi-scale stacked hourglass network

In this paper, we propose a Multi-Scale Stacked Hourglass (MSSH) network to promote functional collaboration and relieve misleading caused by information loss for human pose estimation. An overall framework is illustrated in Figure 2. We adopt the stacked hourglass network as the basic structure of the MSSH network to process features across all scales and capture the various context relationships associated with the body. The pre-processing network generates feature maps of different scales, and dispatch them to each hourglass network, where the small-scale features reach the front of the MSSH network and the large-scale features reach the rear of the MSSH network. The input to each hourglass network consists of two parts, one is the dispatched feature of pre-processing network and another is the output of the pre-level hourglass network. In addition, to further enhance the performance of our network, we propose the inception-resnet as illustrated in Figure 4 to replace the original residual network as a building block for the hourglass network.

We first briefly review the structure of stacked hourglass network. Then a detailed introduction of our pre-processing network and network structure enhancement is presented.

3.1 Stacked hourglass network

The hourglass network is motivated by capturing information contained in the images at different scales. First, the convolution and pooling process are performed, and multiple downsampling operations are performed to obtain some features with lower resolution, thereby reducing computational complexity. In order to increase the resolution of the image features, multiple upsampling is performed. The upsampling operation increases the resolution of the image and is more capable of predicting the exact position of the object. Through such a process, the network structure can obtain more context information by increasing the operation of the receptive field compared to other networks. With intermediate supervision at the end of the hourglass network, the the type and location information of the body keypoints are integrated into the output feature maps. The stacked hour-

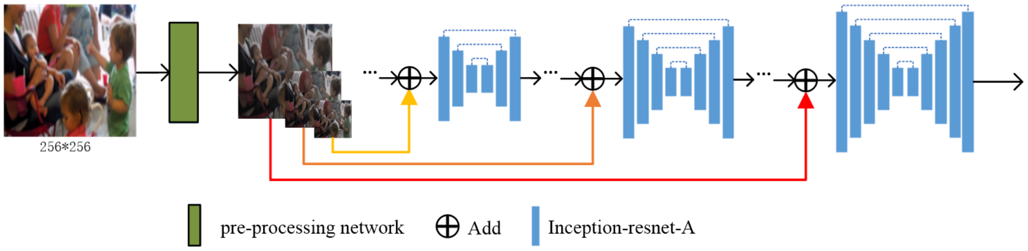


Figure 2: Overview of MSSH network

glass network serially connect multiple hourglass networks. So the subsampling and upsampling are repeated for several times to construct the stacked hourglass network. This means that the entire network has multiple bottom-up and top-down processes to capture information at different scales. And the information of the body keypoints are continuously enhanced as the stack number increases. Stacked hourglass network can fine-tune keypoints gradually.

3.2 Multi-scale pre-processing network

The goal of the multi-scale pre-processing network is to generate different-scale feature maps. The multi-scale pre-processing network help the stacked hourglass networks to form differentiate the determined collaboration mechanisms, and avoid information loss caused by the functional consistency of hourglass networks. As shown in Figure 2, the pre-processing network is construct as a feature preprocessing module before stacked hourglass network. The pre-processing network generates different scales feature maps. The smallest scale feature map is sent to the first-level hourglass network, and the largest scale feature map is sent to the last-level hourglass network, and other features from the small to the large are successively dispatch to the hourglass network from front to rear.

To generate these multi-scale feature maps, it consists of multiple branches with different depths to form feature maps as shown in Figure 3. The convolution layers on each branch are used to extract the features, and the max pooling layers are used to change the resolution of the input as well as to expanding the receptive field.

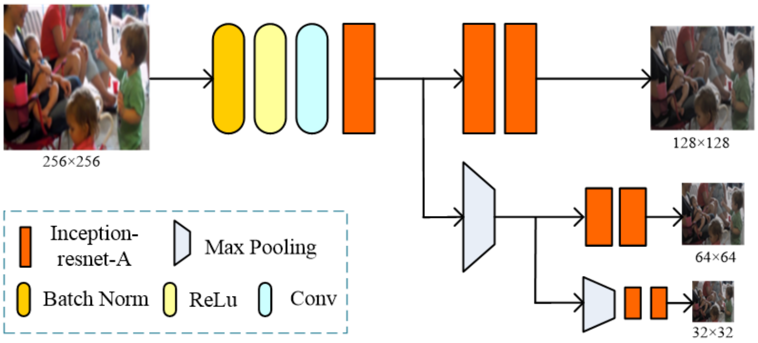


Figure 3: Pre-processing network

3.3 Hourglass network enhancement

To enhance the performance of the hourglass network, the inception-resnet is used as the basic building block in each hourglass network. As shown in Figure 4(a), inception-resnet-A consists of convolutional layers, batch norm layers and Relu units, with channel-wise concatenation and pixel-wise additions. The concatenation of two branches maintains different level of information, but the concatenated features across different channels need to be transformed and normalized by the subsequent convolutional layers. The benefit of the convolutional layers with 1*1 kernels is that the input and output have the same resolution while the depth of channels can be flexible. In addition, inception-resnet-A increases the receptive field of the unit structure by adding a small number of convolution layers, effectively learning the context relationships and improving the implicit space model without the gradient disappearing. Compared with inception-resnet-A, inception-resnet-B uses a subsampling layer and upsampling layer to deepen the building block and further extract different levels of information on the feature maps, as shown in Figure 4(b).

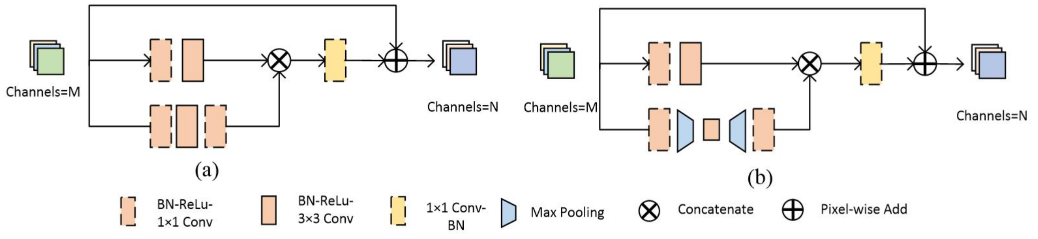


Figure 4: Inception-resnet-A and inception-resnet-B

4 Loss function for multi-scale stacked hourglass network

The consistent processing of constraints makes it difficult to form differentiated collaboration mechanisms for each stacked hourglass network. Motivated by recurrent human pose estimation that uses multiple regression networks to generate multiple loss functions, new loss function with adaptive weight coefficients is designed to pay more attention to hard keypoints, where different keypoints have different weight coefficients of loss function at different scales, and the key weight coefficients are dynamically adjusted from the small scale to the large scale. If the loss function value of a keypoint in the pre-level hourglass network is large, the weight of loss functionin is increased corresponding keypoint on the current hourglass network. If the loss function value of a keypoint in the pre-level hourglass network is small, the weight of loss functionin is decreased corresponding keypoint on the current hourglass network. By dynamically adjusting the weight coefficient, the network gradually increases the focus on the hard keypoints.

4.1 Loss function definition

In this paper, the method of Tompson et al. (2014) is used to generate a 2-D Gaussian heatmap centered on the position of keypoints. The 2-D Gaussian heatmap of the ith keypoint at the jth level is generated as

$$I_{j,i}\left(m,n\right) = a\exp\left(-\left(\frac{\left(m-\mu_m\right)^2}{2\sigma_m^2}+\frac{\left(n-\mu_n\right)^2}{2\sigma_n^2}\right)\right) \tag{1}$$

Where a is the amplitude of the Gaussian funciton, which is set to be +1, if the landmark is non-occluded or set to be -1, if the landmark is occluded. μ represents mean and σ represents standard deviation of the Gaussian function.

The MSE loss function is used on the heatmap of each hourglass network to obtain the loss function of the ith keypoint at the jth level, which is expressed as

$$Loss_{j,i} = \sum_{m,n}\left[I_{j,i}(m,n) - I'{}_{j,i}(m,n)\right]^2 \tag{2}$$

Where $I_{j,i}(m,n)$ represents the predicted heatmap of the ith keypoint at the jth level, $I'_{j,i}(m,n)$ represents the ground truth heatmap of the ith keypoint at the jth level.

According to the structure of MSSH network, it is unreasonable to add the same weight to the loss function of the keypoints directly at all scales, because the standard deviation of the heatmap on the small scale is large, and the standard deviation of the heatmap on the large scale is small, so we need to weight on the basis of the loss function. The weighted loss function is expressed as

$$Loss_j = \sum_i w_{j,i} Loss_{j,i} \tag{3}$$

Where $w_{j,i}$ represents the weight coefficient of ith keypoint at jth level.

4.2 Loss function adjustment

First, initialize the weight coefficient to $w_{1,i} = \frac{1}{N}, i \in [1, N]$, where N represents the total number of keypoints, and according to formula 3, the loss function of the keypoints at the first level is calculated.

Based on the adaboost regression algorithm, the greater the error in this iteration, the greater the weight given by the classifier in the next iteration. We use the following formula to assign the weight coefficients of the loss function for each keypoint at the (j+1)th level.

$$w_{j+1,i} = \frac{w_{j,i}\alpha_j^{(1-Loss_{j,i})}}{Z_j} \tag{4}$$

Where the layer coefficient α_j is calculated as

$$\alpha_j = \frac{Loss_j}{1 - Loss_j} \tag{5}$$

Where Z_j in formula 4 is the normalization factor and is expressed as

$$Z_j{=}\sum_i w_{j,i}\alpha_j^{(1-Loss_{j,i})} \tag{6}$$

It can be seen from formula 4 : if the detection effect of the ith keypoint is smaller, $Loss_{j,i}$ is smaller, then the weight $w_{j+1,i}$ is bigger, and the corresponding learner will focus on this keypoint. Otherwise, the weight of the loss function is bigger, and $w_{j+1,i}$ is smaller and learner will reduce its interference with keypoints that worse for detection. The overall calculation process of the weight coefficient is as follows:

Algorithm 1 Adaptive weight adjustment

```
Input: Initialization: $w_{1,i} = \frac{1}{N}, i \in [1, N], Loss_1 = 0$ ;
Output: Weight collection: $\{w_{J,1}, ..., w_{J,N}\}$ ;
for $j = 1, 2, ..., J - 1$ do
    $Loss_j{=}\sum_i w_{j,i} Loss_{j,i}, s.t. \sum_i w_{j,i} = 1$ ;
    $\alpha_j{=}\frac{Loss_j}{1-Loss_j}$ ;
    $Z_j{=}\sum_i w_{j,i}\alpha_j^{(1-Loss_{j,i})}$ ;
    $w_{j+1,i}{=}\frac{w_{j,i}\alpha_j^{(1-Loss_{j,i})}}{Z_j}$ ;
end for
```

5 Experiment

Our overall structure uses a multi-scale stacked hourglass network with adaptive weight coefficients of loss function. In the experimental section, the database, criteria and implementation details are first introduced. Then, the influence of adaptive adjustment of weight coefficient on the convergence of the model training, the impact of pre-processing network on performance, and the comparison of hourglass network structure are discussed in detail. At last, quantitative assessments are performed on baseline datasets, and their performance is analyzed and discussed

5.1 Experimental setup

5.1.1 Datasets

MPII dataset is a state of the art benchmark for evaluation of articulated human pose estimation. The dataset includes around 25K images containing over 40K people with annotated body joints. The images were systematically collected using an established taxonomy of every day human activities. Overall the dataset covers 410 human activities and each image is provided with an activity label. Each image was extracted from a YouTube video and provided with preceding and following unannotated frames. In addition, for the test set we obtained richer annotations including body part occlusions and 3D torso and head orientations.

LSP dataset contains 2000 pose annotated images of mostly sports people gathered from Flickr using the tags shown above. The images have been scaled such that the most prominent person is roughly 150 pixels in length. Each image has been annotated with 14 joint locations. Left and right joints are consistently labelled from a person-centric viewpoint. Attributions and Flickr URLs for the original images can be found in the JPEG comment field of each image file.

5.1.2 Criteria

PCP is a widely-used criterion for human pose estimation, which evaluates the localization accuracy of body parts. It requires the estimated part end points must be within half of the part length from the ground truth part end points. Some early work requires only the average of the endpoints of a part to be correct, rather than both endpoints. Moreover, the early PCP implementation selects the best matched output without penalizing false positives.

PCK is similar to the Percentage of Detected Joints (PDJ) criterion which is to measure the detection rate of body joints, where a joint is considered to be detected if the distance between the detected joint and the true joint is less than a fraction of the torso diameter. The only difference is that the torso diameter is replaced with the maximum side length of the external rectangle of ground truth body joints. For full body images with extreme pose (especially when the torso becomes very small), the PCK may be more suitable to evaluate the accuracy of body part localization. PCK can be calculated by equation 4.

$$\frac{\left\|y_i - \tilde{y}_i\right\|_2}{\left\|y_{lhip} - \tilde{y}_{rsho}\right\|_2} \leq \gamma \tag{7}$$

Where y_i represents the Ground-Truth position of the ith keypoint and $\tilde{y}_i$ represents the predicted position of the ith keypoint. y_{lhip} denotes the position of the Ground-Truth of the keypoint of the shin, and y_{rsho} denotes the position of the Ground-Truth of the keypoint of the shoulder. The value of γ is between $0 \sim 1$.

PCKh is the modified PCK measure that uses the matching threshold as 50% of the head segment length.

5.1.3 Training details

The input image is 256*256 cropped from a resized image according to the annotated body position and scale. We randomly rotate and flip the images, perform random rescaling and color jittering to make the model more robust to scale and illumination changes. Training data are augmented by scaling, rotation, flipping, and adding color noise. All the models are trained using pyTorch. We use RMSProp to optimize the network on a 1080 GPU with a batchsize of 4 for 220 epochs. The learning rate is initialized as 2.5*10^{-4} and is dropped by 150 at the 175th and the 200th epoch. The Mean-Squared Error (MSE loss) was used in the experiment to compare predicted scoremaps with Ground-Truth scoremaps consisting of 2D Gaussians centered around the human joint position.

5.2 Ablation experiment

In this subsection, we validate the effectiveness of our network from various aspects: the adaptive adjustment of weight coefficient, the performance of pre-processing network and the comparison of hourglass network structure. Since the testing annotations for MPII are not available to the public, the train is on a subset of training images and the evaluation is on a held-out validation set of around

3000 samples. According to the experimental results, the most appropriate method is used to build our network.

5.2.1 Multi-scale stacked hourglass network

We first trained the hourglass network as a baseline model with a PCKh score of 88.78% on the validation set. Through the pre-processing network, dispatch the feature maps of different resolutions (32, 64, 128) into each hourglass network, and use the adaptive weighting function to adjust the proportional weight of each keypoint loss function to optimize the progressive relationship between the loss functions of different scales. As the input size of the images increases, more location details of human keypoints are fed into the network resulting in a large performance improvement. Additionally, the adaptive weighting function works better when the input size of the images is enlarged in 8 stages. The experimental results show that the PCKh score of the model trained by the new method reaches 89.25%, which is 0.47% better than the original structure.

5.2.2 Inception-resnet-A, inception-resnet-B and residual network for building block

In order to enhance the hourglass network for more information, we propose unit structures inception-resnet-A, and inception-resnet-B compared with the residual network as the baseline building block. The structure of inception-resnet-A and inception-resnet-B has been elaborated in Section 3.3 . In this section we mainly evaluate the effects of three structures on the pose estimation. Under the same conditions, the PCkh scores of inception-resnet-A, inception-resnet-B and residual network are respectively 89.91%, 82.95% and 89.19%. By comparing inception-resnet-A with the baseline building block, it is found that in the absence of a gradient disappearing, the addition of a small number of convolution layers increases the receptive field of the unit structure, effectively learns the context relationships, and improves the implicit space model. Comparing inception-resnet-A with inception-resnet-B, it is found that the use of subsampling in the cell structure destroys the structural consistency of the key feature map. Therefore, inception-resnet-C does not apply to the estimation of human pose.

5.2.3 The design of pre-processing network

We need to design a pre-processing network to provide multi-scale feature maps for MSSH network. Because the quality of the feature map output by the pre-processing network directly affects the subsequent detection results, the design of the pre-processing network is crucial. Borrowed the idea of the FPN network and gradually reduced the feature map through the inception-restnet building block. As a comparative test, we used the addition of horizontal connections as a criterion to study the optimal pre-processing network that fits MSSH network and two networks is shown in Figure 3 and 5 . By testing on the MPII verification set, the results of the structure with horizontal connection and the structure without horizontal connection are 88.66% and 89.91% respectively. Therefore, the feature map output by the pre-processing network needs to possess more local information instead of more advanced semantic information.

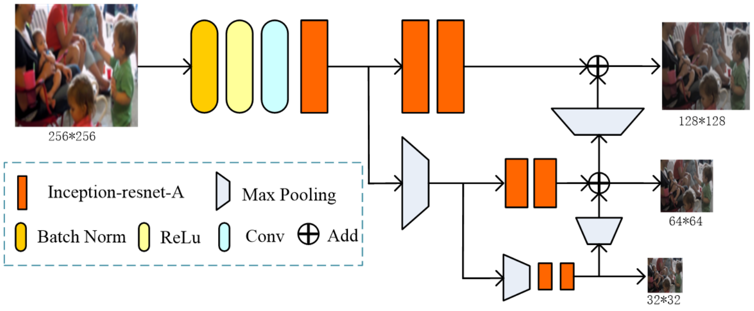


Figure 5: pre-processing network with the addition of horizontal connections

5.2.4 Pixel-wise add and concatenate for information fusion in front of each hourglass network

As shown in Figure 2, in stacked hourglass network, the input of the hourglass network stack with the dispatched feature of the pre-processing network and the output of the pre-level hourglass network, so each hourglass networks is able to access new information. Therefore, the way of information fusion of these two parts is particularly important. In the benchmark hourglass network, we used pixel-wise add as the way of information fusion. Using concatenation as a comparison test to combine features generated from two pipelines, which is similar to inception models. Results show that pixel-wise addition has the better performance with an accuracy improvement of 0.66%, which the pixel-wise add method is 89.91% and the concatenate method is 89.36%. Therefore, we ended up using pixel-wise add.

5.3 Results on MPII and LSP

We use PCKh@0.5 on the MPII test set, use PCK@0.2 and PCP@0.5 on the LSP dataset. The comparisons of our method and state-of-the-art methods are shown in the Tabel 1,2,3. Specifically, on the MPII test set, our method achieves 0.6% and 0.8% improvements on elbow and ankle, where ankle is considered as one of the most challenging parts to be detected.

Table 1: Comparison of PCKh@0.5 score of on the MPII test set

Methods	Head	Sho.	Elb.	Wri.	Hip.	Knee	Ank.	Total
Tompson et al. (2014)	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Lifshitz et al. (2016)	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxari et al. (2016)	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi et al. (2016)	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Wei et al. (2016)	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Newell et al. (2016)	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Our method	98.0	96.4	91.8	87.5	90.4	87.8	84.2	91.2

Table 2: Comparison of PCK@0.2 score of on the LSP dataset

Methods	Head	Sho.	Elb.	Wri.	Hip.	Knee	Ank.	Total
Belagiannis & Zisserman (2017)	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2
Lifshitz et al. (2016)	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Wei et al. (2016)	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Our method	96.9	92.5	87.7	85.2	92.7	93.3	91.7	91.4

Table 3: Comparison of PCP@0.5 score of on the LSP dataset

Methods	Torso	U.Leg	L.Leg	U.Arm	Forearm	Head	Total
Lifshitz et al. (2016)	97.3	88.8	84.4	80.6	71.4	94.8	84.3
Yu et al. (2016)	98.0	93.1	88.1	82.9	72.6	83.0	85.4
Wei et al. (2016)	98.0	82.2	89.1	85.8	77.9	95.0	88.3
Our method	97.9	94.3	91.4	86.6	78.5	95.7	89.5

6 Conlusion

In this work, we have proposed to dispatch the multi-scale feature maps from pre-processing network to each stacked hourglass network, which can be potentially used to aid other deep neural network in training tasks. With adaptive weight loss function, it increases the weight coefficient value of the hard keypoints and optimize the convergence performance, which can be applied to similar multi-stage training loss function for optimization convergence. The effectiveness of the proposed structure and loss function is evaluated on two widely used benchmarks. Later, we hope to explore the extended test of the method under the condition of complex loss function constraints.

REFERENCES

V. Belagiannis and A. Zisserman. Recurrent human pose estimation. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE*, pp. 468–475, 2017.

Y. Chen, Z. Wang, and Y. Peng. Cascaded pyramid network for multi-person pose estimation. *arXiv preprint arXiv:1711.07319*, 2017.

X. Chu, W. Yang, and W. Ouyang. Multi-context attention for human pose estimation. *arXiv preprint arXiv:1702.07432*, 1(2), 2017.

G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. *European Conference on Computer Vision. Springer, Cham*, 18:728–743, 2016.

W. Gong and X. Zhang. Human pose estimation from monocular images: a comprehensive survey. *Sensors*, 16.12, 2016.

L. Ke, M. C. Chang, and H. Qi. Multi-scale structure-aware network for human pose estimation. *arXiv preprint, arXiv:1803.09894*, 2018.

I. Lifshitz, E. Fetaya, and S. Ullman. Human pose estimation using deep consensus voting. *European Conference on Computer Vision. Springer, Cham*, pp. 246–260, 2016.

A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *European Conference on Computer Vision. Springer, Cham*, pp. 483–499, 2016.

U. Rafi, B. Leibe, and J. Gall. An efficient convolutional network for human pose estimation. *British Machine Vision Conference*, 109:1–11, 2016.

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

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

A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. *Computer Vision and Pattern Recognition IEEE*, pp. 1653–1660, 2014.

S. E. Wei, V. Ramakrishna, and T. Kanade. Convolutional pose machines. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, 2016.

W. Yang, S. Li, and W. Ouyang. Learning feature pyramids for human pose estimation. *The IEEE International Conference on Computer Vision (ICCV)*, 2(7), 2017.

X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. *European Conference on Computer Vision. Springer, Cham*, pp. 52–70, 2016.

Appendix

A The training curves of two networks

The figure below shows the training curve for the hourglass network and the MSSH network. As can be seen from the figure, the MSSH network is easier to converge at the beginning and has a higher accuracy in the final stage.

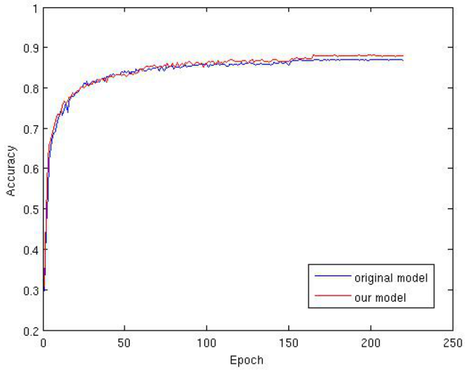


Figure 6: The training curves of stacked hourglass and MSSH network

B Qualitative results

Figure 7 shows the detection results on the MPII validation set and the LSP dataset when the body joints are not twisted and the keypoints are not occluded. Figure 8 shows the detection results on the MPII validation set and the LSP dataset when the body joints are severely twisted and he keypoints are occluded. Figure 9 shows the detection results on the MPII validation set when the human body is occluded or the body joints are twisted.

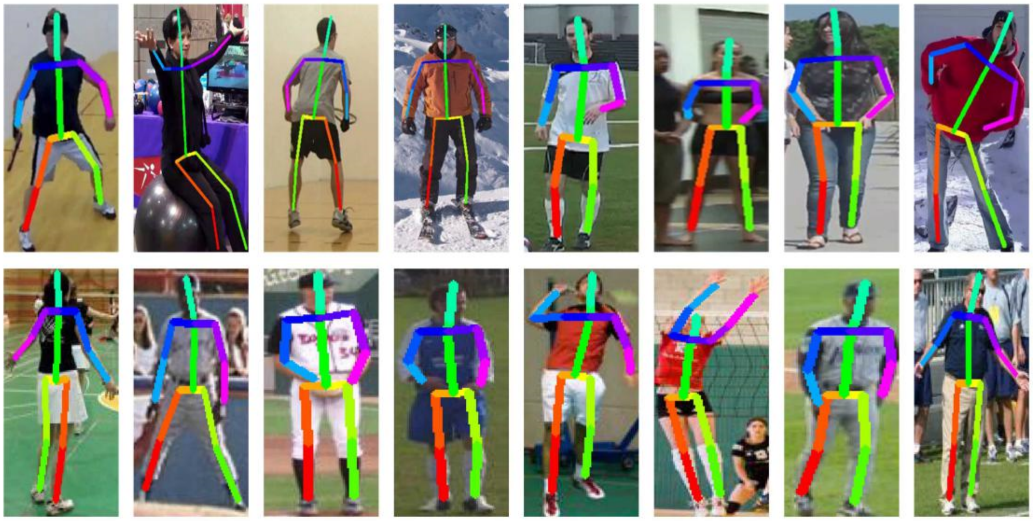


Figure 7: Results on the MPII validation set (top) and the LSP dataset (bottom), when the body joints are not twisted and the keypoints are not occluded.

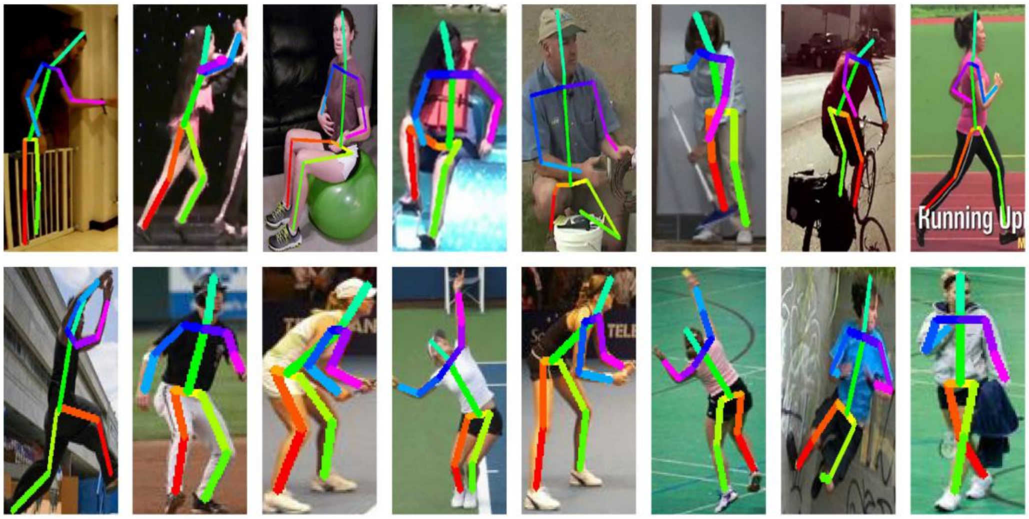


Figure 8: Results on the MPII validation set (top) and the LSP dataset (bottom), when the body joints are twisted and the keypoints are not occluded.

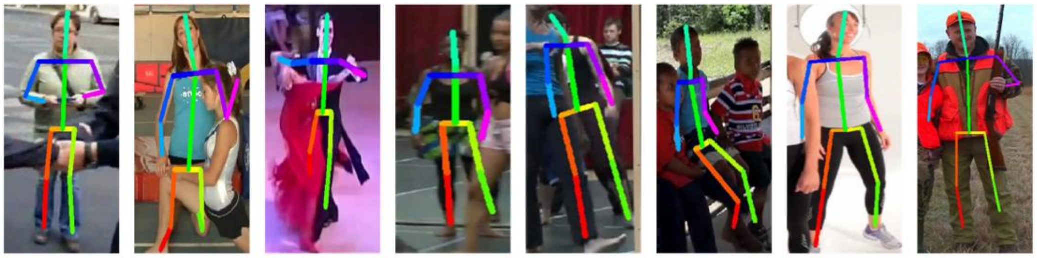


Figure 9: Results on the MPII validation set, when the body joints are twisted or the keypoints are occluded.