BI-DIRECTIONAL MESSAGE PASSING BASED SCANET FOR HUMAN POSE ESTIMATION

Lu Zhou, Yingying Chen, Jinqiao Wang, Ming Tang and Hanqing Lu

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100190
University of Chinese Academy of Sciences, Beijing, China, 100049
{lu.zhou, yingying.chen, jqwang, tangm, luhq}@nlpr.ia.ac.cn

ABSTRACT

Articulated human pose estimation is one of the fundamental computer vision problems. In this paper, a Bi-directional Message Passing(BDMP) module is proposed to fuse convolutional features of different scales in the up-sampling process of the hourglass model for human pose estimation. Moreover, a novel module which integrates Spatial and Channel-wise Attention Network(SCANet) is proposed to refine the features obtained from the message passing stage. We design a Semantics-aware Channel-wise Attention(SACWA) module to reduce the feature redundancy and enrich the semantic information simultaneously. A Sharper Spatial Attention(SSA) module based on the Gumbel-Softmax sampling is proposed to exclude the interference from cluttered background and overcomes the gradient degradation induced by the softmax normalization. The proposed framework achieves leading position on MPII benchmark against the state-of-the-arts methods with much less parameters.

Index Terms— human pose estimation, message passing, SCANet

1. INTRODUCTION

Human pose estimation means to locate body parts(head, shoulders, elbows, writsts, knees, ankles, etc.) from RGB images. Human pose estimation has attracted growing interests in the past few years due to its extensive applications. It serves as a significant basis in the field of person ReID, activity recognition and human-computer interaction. However, the task is challenging due to the large variability of articulation of body limbs, self-occlusion, cluttered background, various clothes and foreshortening.

Traditional human pose estimation methods mainly relied on the hand-crafted features. Significant improvements have been made due to the development of the DCNNs. Some excellent works emerged in the last few years. [1, 2, 3] were devoted to model the relation among the human parts. The well-known stacked hourglass model proposed by [4] utilized the U-Net structure. However, it's still difficult for DNN-based methods to deal with the cases of scale changes and cluttered background.

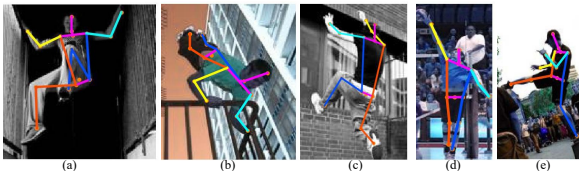


Fig. 1. Visualization of the estimated results on the LSP dataset. Our method achieves promising estimation results on the images which cover various joints scale changes(a,b,c) and cluttered background(d,e).

Except for the scale changes of the whole human skeleton, scale changes in the human parts have great influence on the final result. For example, feet may be larger than head in some images as shown in Figure 1 (a), while head may be larger than the feet in some other cases as shown in Figure 1 (b). The various changes of the joints scales make the prediction process much more difficult. Models such as hourglass tend to over-fit at a fixed scale by only using the last deconvolution feature for inference. Besides, cluttered background and crowded people(Figure 1 (d)(e)) make the estimation much more complicated.

Based on the analysis above, a bi-directional message passing based SCANet is put forward to solve the issues mentioned before. We propose to merge multi-scale convolutional features via message passing to solve the scales change problem. 1) We fuse the different scale features of the hourglass model after each up-sampling. 2) Bi-directional Message Passing module is proposed to fuse features of different scales more effectively. High level semantic information which encodes contextual relationship of the human joints can be delivered from the low resolution features to the high resolution features, while high resolution features convey more spatial details to the low resolution features. The aggregated features are complementary, and thus can handle the diverse scale changes on both the human skeleton and body joints.

Human pose estimation may be difficult under the occur-

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00184

rence of cluttered background with objects which are similar to body joints or human body appearance. In this case, exploiting visual context proves necessary to get a better understanding of the whole body appearance. The semantics encoder of the SACWA is designed to encode high-level visual context. The feature obtained from the semantics encoder not only exploits multi-level spatial context, but also increases the number of scales, which enhances the robustness of the whole system when facing various changes of the body parts scales. However, features of multi-level contexts are not of equal significance for the final prediction. To address this problem, gate function in the form of channel-wise attention is adopted to regulate the contributions from different scale features. In this case, useful features are conveyed to the SSA module and superfluous features are discarded.

Spatial visual attention has been proved effective in locating the regions of interest and excluding the influence of unrelated background. Widely used soft attention utilizes the softmax normalization to obtain the final attention map. However, directly applying the soft attention to our pose estimation model doesn't improve the result. Small number after the normalization which cause the small gradient values w.r.t the model parameters is the main cause of the degradation. Here we propose a Sharper Spatial Attention(SSA) module to make the attention map sharper and avoid the occurrence of extreme small numerical problem. We substitute the in-place Gumbel-Softmax sampling for the common Poisson sampling to make the whole process differentiable. Experiments show the effectiveness of our proposed spatial attention compared with the soft one.

We summarize our contributions as follows:

- A Bi-directional Message Passing mechanism is proposed to combine the convolutional features of different scales and facilitates the information flow among them.
- A Semantics-aware Channel-wise Attention module composed of semantics encoder and feature selector is proposed to encode richer contextual information and weights the feature of different scales dynamically.
- A novel spatial attention is proposed based on the Gumbel-Softmax sampling which makes the attention map sharper compared with the soft attention map and overcomes the degradation caused by the softmax normalization.

2. METHOD

The whole framework is shown in Figure 2. We embed the Bi-directional Message Passing(BDMP), Multi-scale Features Integration(MSFI), Semantics-aware Channel-wise Attention(SACWA) and Sharper Spatial Attention(SSA) modules into each stack of hourglass model sequentially. BDMP and MSFI are utilized to integrate multi-scale features more effectively. The features fetched from BDMP and MSFI are further decorated via the SCANet which is composed of SACWA and SSA. In the subsequent section, we will describe each component in detail.

2.1. Bi-directional Message Passing and Multi-scale Features Integration modules

We stack multiple hourglass modules continuously to perform repeated down-sampling and up-sampling operations. Besides, in the up-sampling process, the features are enhanced with features of the same spatial size from the down-sampling part by element-wise addition. The fused features of different layers are denoted as $F_{11}, F_{12}, F_{13}, F_{14}$, respectively. However, the large number of channels of the $F_{11}, F_{12}, F_{13}, F_{14}$ makes the message passing process consume too much parameters. Taking both the efficiency and effectiveness into consideration, we apply the 1×1 spatial convolution to reduce the channels of the fused features from 256 to 32. It can be seen as another way to reduce the feature redundancy. We denote the channel-reduced features as $F_{21}, F_{22}, F_{23}, F_{24}$. Then we up-sample the F_{22}, F_{23}, F_{24} to the same resolution with the F_{21} whose spatial resolution is 64×64. We denote the up-sampled features as $F_{31}, F_{32}, F_{33}, F_{34}$.

The Bi-directional Message Passing module contains two ways to conduct the message passing. Firstly, the information is propagated from the high resolution features to the low resolution features, which is denoted as top-down process. In this way, the rich spatial details in high resolution features can be transmitted to the low resolution features. The message passing that works in opposite directions are complementary. Therefore, we also build information flow from the low resolution features to the high resolution features, which is denoted as bottom-up process. In this process, global skeleton information is transmitted from the low resolution features to the high resolution features.

In the top-down pathway, we apply 3×3 convolution on the higher resolution features. The resulted features containing more spatial details are summed with the neighboring low resolution features. We keep the highest resolution features as it used to be. As follows, this process is iterated until the information is transmitted to the lowest resolution feature map:

$$\begin{aligned} F_{31}^{'} &= F_{31}, \\ F_{32}^{'} &= F_{32} + H(F_{31}^{'}), \\ F_{33}^{'} &= F_{33} + H(F_{32}^{'}), \\ F_{34}^{'} &= F_{34} + H(F_{33}^{'}), \end{aligned} \tag{1}$$

where H represents the 3×3 convolution.

In the bottom-up pathway, we apply 3×3 convolution on the lower resolution features. The convolved features depicting the global consistency are merged with the neighboring high resolution features by element-wise sum. We keep the

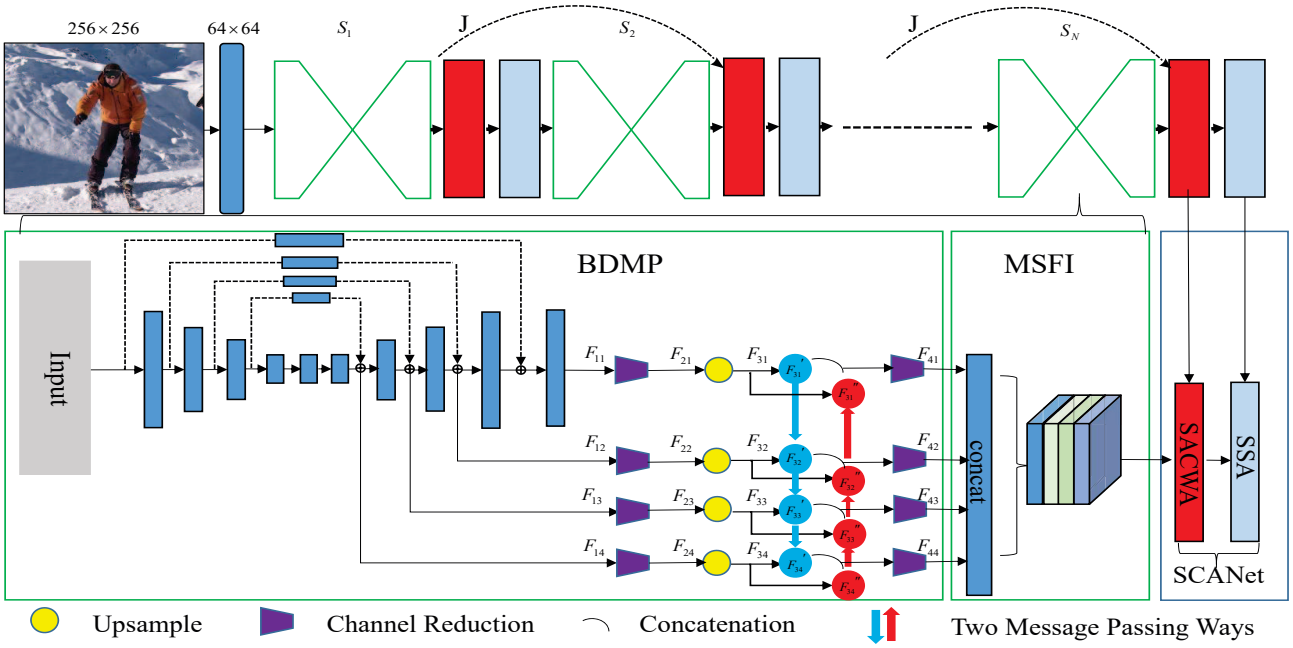


Fig. 2. The proposed Bi-directional Message Passing based SCANet. Our approach contains several important components: Bi-directional Message Passing network(BDMP) module, Multi-scale Features Integration(MSFI) module, Semantics-aware Channel-wise Attention(SACWA) module, and Sharper Spatial Attention(SSA) module.

lowest resolution features as it used to be. The information propagation in the bottom-up pathway is shown as follows:

$$\begin{aligned} F_{34}^{''} &= F_{34}, \\ F_{33}^{''} &= F_{33} + G(F_{34}^{''}), \\ F_{32}^{''} &= F_{32} + G(F_{33}^{''}), \\ F_{31}^{''} &= F_{31} + G(F_{32}^{''}), \end{aligned} \tag{2}$$

where G represents the 3×3 convolution.

The fusion of the spatial details and the global consistency can effectively combine different scales features and generate more discriminative features.

We concatenate the features from the two passing ways and reduce the channels of the concatenated features from 64 to 32 by 1×1 convolution.

$$\begin{aligned} F_{41} &= R(F_{cat}(F_{31}^{'}, F_{31}^{''})), \\ F_{42} &= R(F_{cat}(F_{32}^{'}, F_{32}^{''})), \\ F_{43} &= R(F_{cat}(F_{33}^{'}, F_{33}^{''})), \\ F_{44} &= R(F_{cat}(F_{34}^{'}, F_{34}^{''})), \end{aligned} \tag{3}$$

where R and F_{cat} represent the channel reduction function and concatenation function, respectively.

Then we apply the MSFI by concatenating $F_{41}, F_{42}, F_{43}, F_{44}$ along the channel dimension.

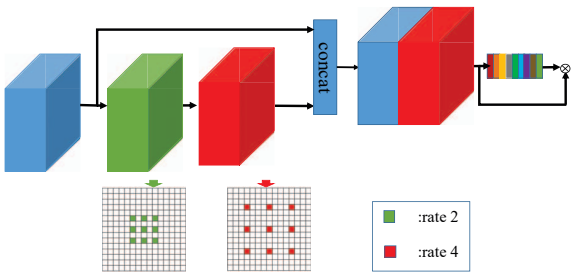


Fig. 3. The proposed Semantics-aware Channel-wise Attention module.

2.2. Semantics-aware Channel-wise Attention

In this subsection, we will introduce the Semantics-aware Channel-wise Attention(SACWA) module, which consists of semantics encoder and feature selector.

The detailed structure of semantics encoder is shown in Figure 3. Two consecutive dilated convolution operations with dilated rate 2 and dilated rate 4 which enlarge the respective field of the network are conducted in the first branch. The second branch retains the original input to preserve the context extracted from the conv-deconv process of the hourglass module. The outputs of the two branches are concatenated as follows:

$$U = F_{cat}(D(O), O), \tag{4}$$

where O is the output of the second branch and $D(O)$ is the output of the first branch, F_{cat} means the concatenation. The output of the semantics encoder may be redundant for the final prediction. Therefore we utilize channel-wise attention[5] which serves as feature selector to weight each feature channel. The feature selector plays the same role as gate function to control the information flow.

2.3. Sharper Spatial Attention

In this part, spatial attention is proposed to refine the output features of SACWA.

The widely used soft attention can be found in [6]. However, directly applying the soft attention to the hourglass module doesn't work as expected. The resolution of the output feature maps of the final hourglass module reaches up to 64×64. In this case, the usage of spatial softmax operation will lead to too small attention weights and severe degradation of both useful and useless information. The small attention values lead to small attended feature values, and thus lead to small gradient values of the convolutional parameters. To address this dilemma, we propose a novel sampling method to overcome the defect resulted from the small feature values and the gradients with respect to the parameters of the last convolution layer whose input is the attended feature are 10-100x larger than the soft one from our experiments. The advantage of the proposed Sharper Spatial Attention mechanism is summarized as follows:

1. The attention map obtained from the Gumbel-Softmax sampling method avoids excessively small numeric values. Attention values within the interesting region approach 1, while attention values of the unrelated background approach 0.
2. The attention map obtained from the Gumbel-Softmax sampling is sharper than the original soft attention map. The sharper map can distinguish the most discriminative visual part from unrelated background more aggressively and reduces the ambiguity.

Next, we will make a detailed description of the Gumbel-Softmax sampling method, and then introduce how to employ this sampling method on the human pose task.

In the first step, we employ the traditional soft attention mechanism to obtain the initial degraded attention map f. The soft attention mask is then input into the sharper attention generator. As is usually set, we normalize the values of the soft attention mask f.

$$f^n_{i,j} = \frac{f_{i,j} - f_{min}}{f_{max} - f_{min}}, \tag{5}$$

where f_{max} means the maximum value of the whole attention map, while f_{min} means the minimum value of the whole attention map. We treat the normalized value of each position in the attention map as the probability we choose it. The larger the value is, the higher probability it can be chosen with. Poisson sampling is an ideal method to implement the idea. However, if Poisson sampling is implemented on the attention map, the resulted attention map is not differentiable to the f^n. Gumbel-Softmax sampling is an alternate choice to replace the Poisson sampling and makes the whole process differentiable. Each position of the resulted attention map f^n is subject to the Bernoulli distribution,

$$p^1_{i,j} \triangleq f^n_{i,j}, p^0_{i,j} \triangleq 1 - p^1_{i,j}. \tag{6}$$

Gumbel-Softmax sampling technique comes from Gumbel-Max sampling[7]:

$$f^{sharp}_{i,j} = \underset{k\in\{0,1\}}{argmax}(g^k_{i,j} + log(p^k_{i,j})). \tag{7}$$

The Gumbel distribution is as follows:

$$Gumbel(x;\mu,\beta) = e^{-z-e^{-z}}, z = \frac{x-\mu}{\beta}. \tag{8}$$

Here, $g^0_{i,j}, g^1_{i,j}$ are i.i.d samples drawn from Gumbel(0,1).

However, the argmax operation of the Gumbel-Max sampling is also not differentiable. Gumbel-Softmax sampling approximates the Gumbel-Max sampling by the softmax trick, which can be seen as a way of reparametrization. Through the Gumbel-Softmax sampling, the whole process is differentiable. The approximation is as follows:

$$f^{sharp}_{i,j} = \frac{exp((g^1_{i,j} + log(p^1_{i,j}))/\tau)}{\sum_{k\in 0,1} exp((g^k_{i,j} + log(p^k_{i,j}))/\tau)}. \tag{9}$$

Temperature parameter τ controls the degree of the approximation. When τ approximates 0, the distribution generated by the Gumbel-Softmax sampling looks more like one-hot[8]. However, The larger the τ is, the smoother the distribution generated by the Gumbel-Softmax sampling is. When τ becomes large enough, the distribution generated by the Gumbel-Softmax sampling is completely a uniform distribution.

3. EXPERIMENT

3.1. Experiments Settings

The proposed model is evaluated on the widely used benchmark MPII Human Pose. Our implementation follows [4]. The model is trained with Torch7 toolbox with the initial learning rate 2.5×10^{-4}. We drop the learning rate by 10 at the 150th, 170th, and 200th epoch. We make use of RMSprop algorithm to optimize the parameters of the model. Each person is cropped from the image according to the center and scale of the human body and resized to the size of 256×256. We

Table 1. Evaluation results using PCKh@0.5 as measurement on the MPII dataset

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Tompson et al. [2]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Hu et al. [9]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Gkioxary et al. [10]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi et al. [11]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Wei et al. [12]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat et al [13]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell et al. [4]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Ning et al. [14]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Chu et al. [6]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
ours	98.4	96.3	92.0	87.9	90.5	**88.6**	**85.5**	**91.6**

augment the data by implementing rotation, scaling, and color jittering. Six-scale image pyramids combined with flipping are employed during testing. 4-stack hourglass model is used in our final experiments and finetuned with the assistance of the hard point mining loss.

3.2. Comparison with the State-of-the-Art

Our approach achieves competitive results compared with the previous state-of-the-art methods in the leading board of MPII dataset. The results on the MPII dataset are summarized in Table 1. Our method surpasses [6], which aggregated multi-scale features via multi-scale spatial attention. The parameters number of our model approximates 15.6M, while 8-stack U-Nets[4] approximates 25.5M and [6] approximates 58.1M.

3.3. Ablation Study

The ablation study is implemented on the validation set of the MPII dataset.

3.3.1. The Effectiveness of the MSFI

The baseline in this part is the pure 2-stack hourglass model if not specified. We first compare our Multi-scale Features Integration mechanism with the baseline and summarize the results in Table 2. The MSFI aggregates semantic information covering 4 levels as described in section 2.1. The MSFI improves the mean result of baseline by 0.3% in that the multi-scale information which enhances the robustness of the model when facing various scale changes is included in the aggregated features.

All the experiments below incorporate the individual submodule based on the MSFI framework. The comparison of different submodules can also be found in Table 2.

3.3.2. The Effectiveness of the BDMP

Message passing among different level features promotes the communication of the semantic information among different scales. BDMP gets 88.66% PCKh score finally and exceeds the MSFI by 0.27% as features which contain high-level context and spatial details are obtained via BDMP for final prediction.

Table 2. Comparison of different submodules. (1) SCA: The integration of SACWA and SSA. (2) ALL: The integration of BDMP, SCA and hard point mining.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Baseline	96.38	95.09	89.02	83.48	87.81	83.90	80.16	88.09
MSFI	96.45	95.30	89.16	84.48	88.26	83.76	80.61	88.39
MSFI+BDMP	96.49	95.38	89.33	84.72	88.11	84.54	81.22	88.66
MSFI+SACWA	96.42	95.28	89.42	84.84	88.52	84.73	80.56	88.64
MSFI+SSA	96.32	95.14	89.21	84.50	88.75	84.67	81.79	88.74
MSFI+BDMP+SCA	96.38	95.67	89.43	84.68	88.16	85.01	82.03	88.87
MSFI+ALL	96.35	95.21	89.57	84.82	87.92	85.41	82.12	88.90

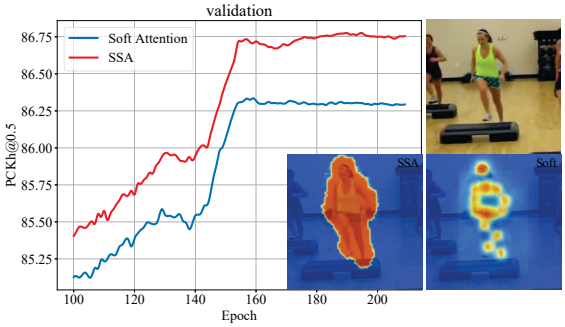


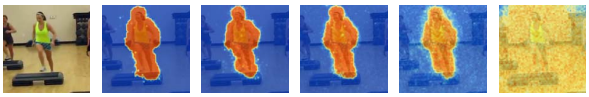
Fig. 4. Comparison of the sharper attention and soft attention. The curve of the figure shows the PCKh@0.5 scores on the MPII validation set over the training process.

3.3.3. The Effectiveness of the SACWA and SSA

By adding SACWA module, the PCKh score reaches 88.64% and surpasses MSFI by 0.25%.

The SSA module achieves promising result by generating sharper attention map compared with the traditional soft attention mechanism. Figure 4 displays the comparison of the sharper attention map and soft attention map. The sharper attention module surpasses the soft attention module by a large margin as shown in Figure 4. By adding SSA module, our algorithm obtains 88.74% PCKh score and improves the result of the MSFI framework by 0.35%.

The attention map generated from the SSA module is much sharper than the soft attention map and covers more regions of the whole body. The soft attention map mainly concentrates on the separate body parts instead of the whole body and thus loses the global skeleton information and appearance cues. We experiment with different values of τ to observe how temperature impacts on the generation of the attention map. The parameter τ's influence on the generation of the attention map can be found in Figure 5. We can see that the attention map with the larger τ looks more irregular in that the sampling approximates uniform sampling in this case, while the attention map with the smaller τ looks much sharper. To avoid numeric instability caused by the smaller $\tau(\tau < 1.0)$ when implementing the algorithm on Torch7, we adopt τ as 1.0 in the end.



(a) Input (b) $\tau = 1$ (c) $\tau = 2$ (d) $\tau = 3$ (e) $\tau = 5$ (f) $\tau = 10$

Fig. 5. Sharper attention map generated by setting different values of temperature.

3.3.4. The Effectiveness of Whole Framework

The SCANet which combines SACWA and SSA proves effective compared with the separate single module and gets 88.76% PCKh score in the end. If we combine the BDMP and SCANet together, the final PCKh score can reach 88.87% and performs better than all the module mentioned above. We achieve 88.90% PCKh score by training the overall framework combined with hard points mining technique.

4. CONCLUSION

A Bi-directional Message Passing based SCANet is proposed in this paper. We tackle the scale variance and cluttered background problem by combing the Bi-directional Message Passing mechanism, SACWA module and SSA module together. With the design of Bi-directional Message Passing module and Multi-scale Features Integration subnetwork, we integrate different resolution features together and get a comprehensive understanding of different scale features. The feature concatenated across different levels is further refined via the proposed SACWA module. On the one hand, the proposed SACWA module enriches semantic information by the semantics encoder we design. On the other hand, feature is filtered with the assistance of the channel-wise attention. Finally, SSA module which generates sharper attention map to weaken the influence of the unrelated cluttered background further boosts the performance of the whole network.

5. ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China 61772527, 61806200.

6. REFERENCES

[1] Arjun Jain, Jonathan Tompson, Mykhaylo Andriluka, Graham W Taylor, and Christoph Bregler, "Learning human pose estimation features with convolutional networks," *arXiv preprint arXiv:1312.7302*, 2013.

[2] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.

[3] Xiao Chu, Wanli Ouyang, Xiaogang Wang, et al., "Crf-cnn: Modeling structured information in human pose estimation," in *Advances in Neural Information Processing Systems*, 2016, pp. 316–324.

[4] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[5] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, vol. 7, 2017.

[6] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang, "Multi-context attention for human pose estimation," *arXiv preprint arXiv:1702.07432*, vol. 1, no. 2, 2017.

[7] Chris J Maddison, Daniel Tarlow, and Tom Minka, "A* sampling," in *Advances in Neural Information Processing Systems*, 2014, pp. 3086–3094.

[8] Eric Jang, Shixiang Gu, and Ben Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[9] Peiyun Hu and Deva Ramanan, "Bottom-up and top-down reasoning with hierarchical rectified gaussians," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5600–5609.

[10] Georgia Gkioxari, Alexander Toshev, and Navdeep Jaitly, "Chained predictions using convolutional neural networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 728–743.

[11] Umer Rafi, Bastian Leibe, Juergen Gall, and Ilya Kostrikov, "An efficient convolutional network for human pose estimation.," in *BMVC*, 2016, vol. 1, p. 2.

[12] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[13] Adrian Bulat and Georgios Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *European Conference on Computer Vision*. Springer, 2016, pp. 717–732.

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