



Multi-scale Attention Aided Multi-Resolution Network for Human Pose Estimation

Srinika Selvam^(✉) and Deepak Mishra^(✉)

Indian Institute of Space Science and Technology, Thiruvananthapuram, Kerala, India
srinika0812@gmail.com, deepak.mishra@iist.ac.in

Abstract. In this paper, we propose attention maps at various scales on multi-resolution feature extractor baseline network for human pose estimation. The baseline network captures information across various scales with the help of repeated bottom-up and top-down approach using successive pooling and up-sampling. We propose a network named Refinement Net for regressing the predicted heatmaps to 2D joint locations to remove ambiguities in predicted position. We experiment with three levels of attention schemes - global, heatmap and multi-resolution. Attention masks helps in generating basin of attraction that helps the network on deciding where to “look”. The proposed network performance is at par with the state-of-the-art two dimensional pose estimation methods on MPII dataset.

Keywords: Human pose estimation · Multi-resolution · Attention maps

1 Introduction

Pose estimation is a fundamental yet challenging problem in computer vision. Human pose estimation is task of identifying humans in an image and recovering their body poses. It is one of the longest serving problems in computer vision due to complex models involved in observing poses. People’s poses are extensively used for pedestrian detection, human-robot interaction, sign-language understanding, virtual reality, etc. Pose estimation can be solved very efficiently by posing it as a deep learning problem. Hence, it has seen a lot of advancements over the recent years.

Convolutional neural networks (CNN) have seen an explosive success in the field of deep learning. Most state-of-the art pose estimation networks use CNN architectures to achieve outstanding performance on publicly available datasets such as MPII, Human3.6M, Leeds Sports, COCO, etc. In this paper, we propose a CNN architecture, called multi-resolution network (MRN) that has conv-deconv structure by successive pooling and up-sampling. We stack multiple MRNs to allow repeated refinement of prediction at consecutive stages. The prediction

maps are then fed to Refinement Network for regressing accurate 2D joint locations. Since the joint locations do not depend on the background region (non-human), these regions need to be eliminated before feeding into the pose estimation network to prevent the network from learning false features. Hence we propose attention schemes at various scales that allows the network to learn the required attention basin. Attention is a self-learning module and does not require any labelled data while training. The output of the attention module is a probability map that boosts near-human regions and suppresses other regions that are not useful for identifying the joint locations. Detailed explanation of the work is described in Sect. 3.

This work makes the following contributions:

- **Multi-Resolution Network (MRN)** - allows repeated bottom-up and top-down inference across various scales. Output of MRN is a single heatmap for all 16 joint locations.
- **Refinement Net** - Output of MRN is fed to Refinement Net for accurate regression of 2D joint locations from output heatmap.
- **Multi-scale Attention Modules** - To help the network learn faster and efficiently, we propose attention modules to boost regions of interest and mask unimportant regions.

Baseline network without attention achieves highest accuracy of 83.2%. However, with global attention, we achieve 3.5% increase in accuracy whereas with multi-resolution attention, we get 0.9% increase on MPII dataset.

2 Related Work

CNNs have revolutionized the task of human pose estimation by incorporating complex feature extracting network with reduced the number of parameters as compared to fully connected networks. Although it has significantly enhanced the performance, learning the exact (x, y) positions of body joints from an image is a complex task. To solve this problem, researchers turned to the heatmap, which is made by placing Gaussian blobs at every joint location. Plenty of methods were designed to regress the heatmap instead of (x, y) coordinates, such as Tompson et al. [16], Newell et al. [13], Wei et al. [18].

Pose Estimation. Introduction of Stacked Hour-Glass structure by [13] has certainly steered pose estimation to a new direction by focusing on multi-scale feature extraction. It uses CNN as their base and produces heatmap output for each joint location. Other major contributions to pose estimation include [1, 9, 10]. Numerous variations to [13] were also experimented by [4, 6, 8]. Cascaded pyramid network by [3] came up with a module named RefineNet for fine tuning of the heatmaps obtained from previous network module. RefineNet concatenates all the pyramid features rather than simply using the up-sampled features at the end of hourglass module. Different from RefineNet, we propose a module

designed to eliminate any ambiguity that could arise due to spread spectrum of Gaussian blobs at each joint location or due to ordering of joint locations. This module regresses the heatmap output to corresponding joint locations.

Self Visual Attention. Visual attention concept is computationally efficient and effective for object detection [7,22], image recognition [2,17], caption generation [11,19,21], etc. In most cases, attention masks are usually learned from extra bounding box manual annotations. Learning maps from annotated data not only is time-consuming but also does not bring any progress to the area of deep learning. However, very few work has been done incorporating attention modules for human pose estimation, they include [5,15]. Attention aided pose estimation helps the network learn the regions of interest in the image and lead to faster convergence.

3 Network Architecture

In this section we elaborate the important constituents of the proposed human pose estimation architecture i.e. MRN, Refinement Net, and attention modules.

3.1 Multi-Resolution Network

The proposed network architecture is as described in Fig. 1. The architecture is inspired from Stacked Hourglass structure [13]. It was aimed at collecting information from various scales from the image. In the original implementation of the hourglass network, they make use of residual blocks that have skip connections within itself. We believe that the second tier of skip connection is redundant and does not help in learning better features required for pose estimation. Our network omits the implementation of residual blocks. The network starts with 7×7 convolutional layer with stride 2, followed by a residual module and a round of max pooling to bring the resolution down from 256 to 64. This is implemented in the Front Module. The network branches off to skip connections at every resolution to apply more convolutions at the pre-convolved layers as shown in Fig. 1. After reaching the final resolution of 4×4 , the network does nearest neighbour interpolation with few more convolutional layers to bring back the network output to 64×64 resolution. At the lowest resolution, two consecutive rounds of 1×1 convolutional are performed. At each level of resolution, there are 256 feature vectors. At every up-sampled resolution, the feature vector is added with the skip connections coming from the corresponding resolution of bottom part of the hourglass. Local evidence helps in identifying faces, hands, etc whereas the consolidated features from various scales help in understanding the body pose. This feature is very critical to the network's performance over other networks. We extend our network architecture by stacking multiple MRNs back-to-back, feeding the output of one as input to the next. This provides the network a mechanism for repeated bottom-up, top-down inference allowing for reevaluation of initial estimates and features across the whole image. Multiple

MRNs help in refining the output prediction of previous MRNs. The output of this network is a single heatmap for all joint locations in the image. The resulting prediction heatmap value at a position gives the probability of any joint occurring at that location.

3.2 Refinement Network

Predictions from MRN does not provide any information about correspondence of prediction to joint locations. That is, we do not know which Gaussian peak corresponds to which joint in the human body. In order to regress body joint locations from the heatmap, we propose a module named ‘‘Refinement Net’’. Output of stacked MRNs is fed to Refinement Net. Refinement net helps in learning probability map to joint correspondence. The order of output of the Refinement network encodes the information about the sequence of body joint locations. Also, the heatmaps are obtained by placing Gaussian blobs at corresponding true values of the 2D pixel locations in the image. So, when going back from heatmap to joint locations, due to the spread spectrum of Gaussian blob, multiple peaks for a single joint location may occur. To avoid this complication, we redundantly predict 2D joint locations from the predicted heatmap from previous network. This module refines the prediction from MRN. Moreover, while training with heatmaps, the network (Front Module + MRN) gets stuck at a local minima corresponding to a heatmap with all joint prediction values 0.5. Refinement Network also helps in overcoming this saddle point and allows faster convergence of the model. Refinement network architecture is as described in Fig. 2 and the overall network architecture is as described in Fig. 3.

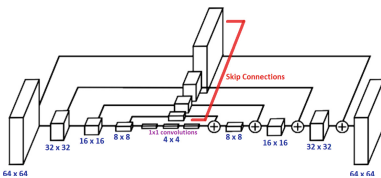


Fig. 1. Multi-resolution network is inspired from the popular Stacked hourglass network - aimed at capturing features at various resolution. Each block in MRN is a convolutional layer with 256 feature vectors - unlike stacked hourglass in which each block is a residual module.

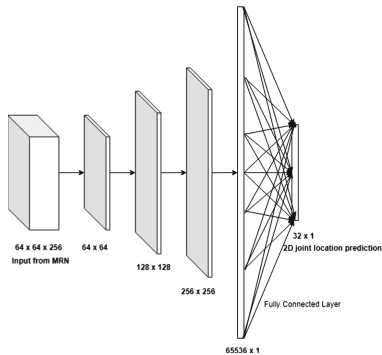


Fig. 2. Architecture of Refinement Net. Refinement Net helps in regressing joint positions from heatmaps that lacks information about correspondence between predictions and joint locations.

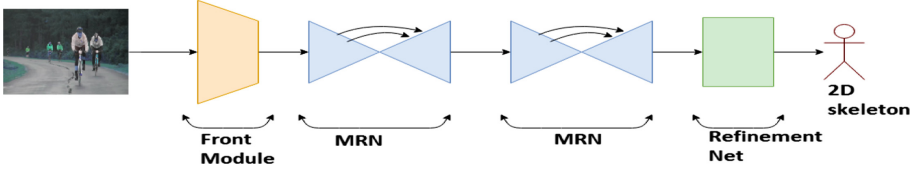


Fig. 3. Baseline network with MRN and Refinement Network

3.3 Attention Module

Attention schemes help the network in deciding where to look. The output of the attention module is a relevance map that boosts near-human regions and suppresses other regions that might not be useful for identifying the joint locations. Attention can be of two types: 1. Hard Attention and 2. Soft Attention. In hard attention, the attention mask has binary values - $\{0, 1\}$. 0 at a location indicates that information content at that position is irrelevant to the task of joint prediction whereas 1 indicates otherwise. In soft attention, each pixel position has values in the range $[0, 1]$ - indicative of the relevance of the information at that position for output prediction. In our implementation, we extensively use soft attention masking since hard attention masks are difficult to back propagate. In this work, we propose attention schemes at three levels: **Global**, **Heatmap** and **Multi-Scale** Attention Schemes. Moreover, we do not require any manual annotations for attention masks.

Global Attention. Global attention is implemented at the input resolution. Every image is passed through global attention module and then to MRN and Refinement Net. The output of the module is an attention mask, same size as that of the image. The attention mask is element-wise multiplied (denoted by \odot) with all 3 channels of the image. This emphasizes the regions of interest and de-emphasizes unimportant regions in the original image. The output is then fed to MRN followed by Refinement Net. Global attention module also has hourglass structure with skip connections. It is aimed at capturing masking information from various scales across the image. Figure 4 shows the architecture of Global attention scheme.

Attention mask on the image is generated by passing the image through attention convolutional layers described in Eq. 1 where f denotes the attention transformation, I denotes the input image, W_{att} is the weights of the attention module, M is the attention mask and X is the input to stacked MRNs.

$$M = f_{att}(I; W_{att}) \quad (1)$$

$$X = I \odot M \quad (2)$$

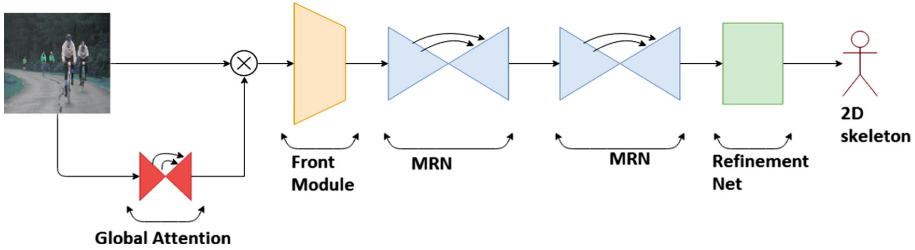


Fig. 4. Global Attention Scheme - Attention basins are generated on the input image. Unimportant regions are masked before passing through the network. Attention map dimension is as that of the input image.

Heat Map Attention. Heat map attention masks are implemented on the output of stacked MRNs. Attention module for heatmap predictions is an redundant layer that helps in filtering out wrong predictions before passing the predictions onto subsequent stages. This is important because, in our implementation of stacked MRN network, the subsequent stages learn from the previous stage’s prediction. Applying attention onto each of the predicted output heatmaps will help the network learn faster and efficiently. Attention mask in each layer is element-wise multiplied with the predicted heatmap and passed to the next stage of the hourglass. Figure 5 shows the architecture of heatmap attention scheme.

Let X^i denote the output (heatmap) of i^{th} MRN. To generate attention maps on X_i , we pass it through attention module (Eq. 3). This mask is element-wise multiplied with heatmap prediction and added with original prediction to enhance the regions of interest (Eq. 4). This is done in order to overcome the problem of vanishing gradient. Attention mask M^i is a relevance map that consists of values in the range $[0, 1]$. 1 implies high relevance and 0 implies low relevance. When this mask is multiplied with the predicted heatmap which is a probability map whose range is also $[0, 1]$, the resulting values are infinitesimal that leads to vanishing gradient when propagated through multiple stacks of MRN. Hence, we add identity mapping to avoid this complication. Y^i is the input to $(i+1)^{th}$ MRN.

$$M^i = f_{att}(X^i; W_{att}^i) \tag{3}$$

$$Y^i = X^i \odot M^i + X^i \tag{4}$$

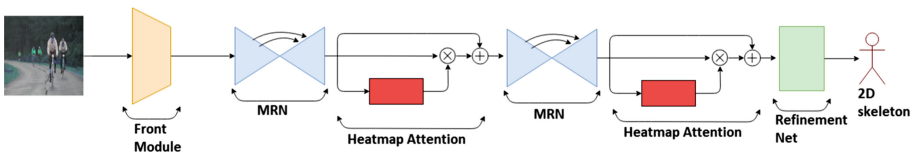


Fig. 5. Heatmap attention scheme - Attention basins are generated for each heatmap. Unwanted regions of the heatmap are masked before passing through subsequent stacks of MRN. Each attention map dimension is 64×64 - as that of MRN input/output.

Multi-Resolution Attention. Instead of building an explicit attention module, we can incorporate attention into MRNs itself. This attention scheme learns attention masks on all resolutions present in MRN. The attention masks run parallel to the skip connections in MRN. Each resolution attention mask is element-wise multiplied with corresponding resolution layer output and propagated to the subsequent layers in each hourglass and eventually to subsequent MRN stacks and Refinement net. Multi-resolution attention emphasize relevant information present at each scale. This improves the quality of each resolution layer output as well as predicted 2D joint positions. Figure 6 shows the architecture of multi-resolution attention scheme.

Let X_t^i denote the i^{th} layer in the top half of the MRN and X_b^i denote its corresponding layer in the bottom half in a MRN. Attention maps are generated only on bottom half of MRN. Attention maps are learnt from X_b^i . X_b^i is passed through the attention module. This mask is element-wise multiplied with the corresponding feature vector for upsampled layer in the network and added with the original feature vector to enhance the regions of importance to overcome the problem of vanishing gradient as described in Subsect. 3.3. Y^i is added with the skip connections coming from the corresponding top section of the MRN. X_s^i denotes the skip connection coming from i^{th} top layer in MRN. Z^i is the input to $(i+1)^{th}$ layer in MRN. This is described in Eqs. 5, 6, and 7.

$$M^i = f_{att}^i(X_b^i; W_{att}^i) \quad (5)$$

$$Y^i = X_t^i \odot M^i + X_t^i \quad (6)$$

$$Z^i = Y^i + X_s^i \quad (7)$$

3.4 Training Details

Entire network was trained on MPII dataset with single person annotations with all visible body joints. At every convolutional layer, kernel and activity regularization are added to avoid over-fitting. We add batch normalization at every convolutional layers to speed up the training process, except in the 1×1 convolutional layers. Leaky ReLU activation is imposed for all convolutional layers, except the attention layers. We use sigmoid activation to generate the attention masks. Kernels of size (3, 3) and stride (1, 1) are used throughout MRNs. The network weights are all initialized with tensors of all ones. Adam optimizer is used to optimize the network on a single Titan X GPU. The network is trained on mini-batches of size 30 and shuffled between epochs. Network is trained using MSE as loss function between 16 predicted (x, y) joint positions

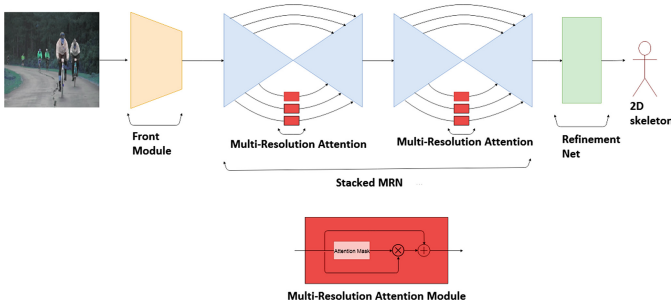


Fig. 6. Multi-resolution attention scheme - attention basins are generated on every scale in MRN. Unimportant regions are filtered out before passing through subsequent resolution levels in MRN. Each attention map has dimensions as that of each scale in MRN - 64×64 , 32×32 , 16×16 , 8×8 and 4×4 .

and ground truth as described in Eq. 8. For images with multiple people in it, output is predicted on a single person with all visible body joints.

$$Loss = Mse(P_{x,y}; G_{x,y}) = \sum_{i=1}^{16} (G_{x,y}^i - P_{x,y}^i)^2 \tag{8}$$

Tasked Learning. The training method is derived from [15]. We represent body joint learning as task A and attention learning as task B. The result of task B has great influence on task A. For example, if task B filters out body regions, task A would never converge to a better solution, because the input feature maps do not contain much important information. So, firstly, we freeze the parameters of task B, and initialize all attention masks with all ones tensors and train task A. Then we freeze the parameters for task A, and train task B. Finally, we jointly train the two tasks. Training algorithm is as described in Algorithm 1. The learning rate is dropped by a factor 10 at the completion of each tasked training epoch.

4 Results

We compare our results with leading state-of-the-art results in Table 1. The metric used for comparison is average accuracy (in percent) over all body joints. Example results can be found in Fig. 7 for baseline network and in Fig. 9 for attention aided architecture. We also analyze the performance of the baseline network and attention aided network only various categories such as walking, dancing, exercising, standing, etc. Table 2 stages the results. Intermediate results of the baseline network are staged in Fig. 8.

Algorithm 1. Tasked Training Algorithm

```

1: procedure TRAIN(model, input, max_epochs)
2:   for i in range(max_epochs) do
3:     prediction  $\leftarrow$  model.predict(input)
4:     error  $\leftarrow$  (ground_truth - prediction)2
5:     Back propagate error
6: Initialize model.attention_weights to ones
7: while convergence not attained do
8:   model.attention_weights.trainable  $\leftarrow$  False            $\triangleright$  Freeze attention weights
9:   TRAIN(model, input, max_epochs)
10:  model.attention_weights.trainable  $\leftarrow$  True            $\triangleright$  Release attention weights
11:  model.pose_network_weights.trainable  $\leftarrow$  False  $\triangleright$  Freeze baseline network weights
12:  TRAIN(model, input, max_epochs)
13:  model.pose_network_weights.trainable  $\leftarrow$  True        $\triangleright$  Release baseline weights to
    make the network end-to-end trainable
14:  TRAIN(model, input, max_epochs)
15:  Reduce learning_rate

```

\triangleright while loop continues

Table 1. Comparison of results on MPII human pose dataset

Method	Average accuracy	Method	Average accuracy	Method	Average accuracy
[16]	82.0	[14]	82.4	[12]	85.0
[18]	88.5	[13]	90.9	[5]	91.5
[20]	92.0	[15]	92.0	Ours - Fig. 3	83.2
Ours - Fig. 4	86.7	Ours - Fig. 5	78.6	Ours - Fig. 6	84.1

**Fig. 7.** Example results of our baseline (MRN + Refinement Net) on MPII dataset

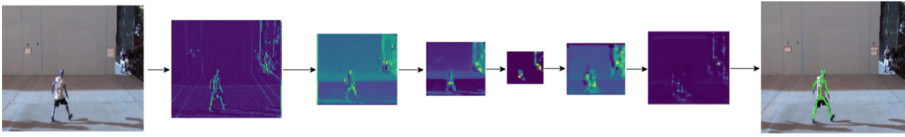


Fig. 8. Intermediate features extracted from the network. Bottom half of MRN does human localization whereas top half of the network derives higher level features from localized.

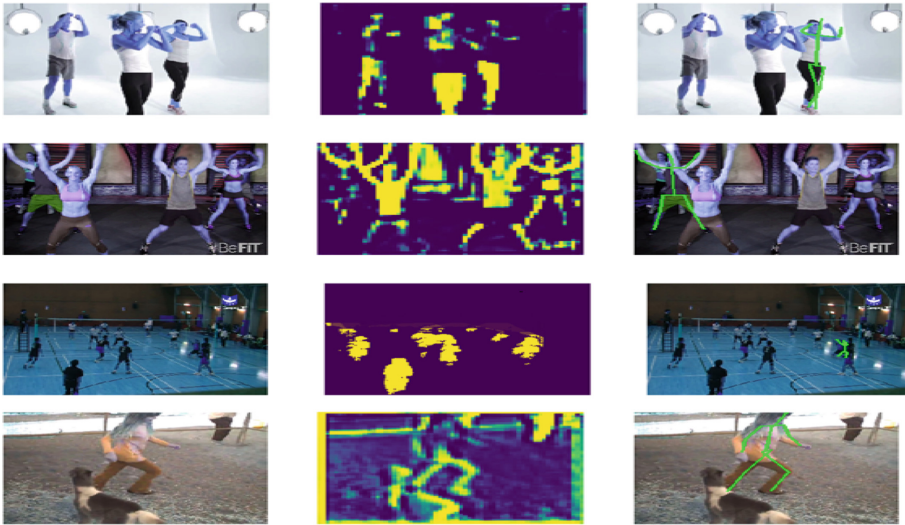


Fig. 9. Example results of our attention aided baseline network on MPII dataset. First column is the original input images. Second column is the corresponding attention masks generated for the input image. Third column is the predicted 2D joint locations of the network.

Table 2. Category-wise assessment of average accuracy (in percentage) on MPII dataset

Our method	Standing	Walking	Dancing	Skiing	Exercising	Bicycling	Mountain biking	Volleyball	Basketball
Fig. 3	81.93	93.62	76.67	84.29	82.82	83.12	80.00	60.2	78.42
Fig. 4	83.86	92.98	76.67	81.43	87.00	82.69	82.38	62.46	80.18
Fig. 5	81.93	89.36	76.67	80.00	79.90	79.89	73.31	60.82	78.42
Fig. 6	85.44	89.36	76.67	84.29	87.00	83.12	80.00	67.38	81.93

5 Conclusion

Our work is a significant contribution to attention-based human pose estimation since seldom researches have explored this field. Baseline network performs at an average accuracy of 83.2%. The reduced accuracy compared to [13] is due to omission of residual connections at the benefit of lesser number of parameters. With the use of attention schemes, the performance of the network (in terms of accuracy and time to convergence) has significantly improved. Global attention schemes gives an additional 3.5% increase in accuracy whereas multi-resolution attention gives an additional 0.9% improvement. However, heatmap attention modules decrease the performance of the baseline network by a significant 4.6%. Heatmaps have very concise and important information about the body joints. Imposing attention on heatmaps leads to disposal of important regions from the heatmap. Moreover when multiple MRNs are stacked, this leads to successive filtration of the heatmaps by the attention modules which leads to significant loss of information, reflecting in the reduced accuracy of the network.

References

1. Alp Güler, R., Neverova, N., Kokkinos, I.: DensePose: dense human pose estimation in the wild. In: CVPR (2018)
2. Chen, Y., Zhao, D., Lv, L., Li, C.: A visual attention based convolutional neural network for image classification. In: 2016 WCICA (2016)
3. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR (2018)
4. Chou, C.J., Chien, J.T., Chen, H.T.: Self adversarial training for human pose estimation. In: 2018 APSIPA ASC (2018)
5. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: CVPR (2017)
6. Guo, C., Du, W., Ying, N.: Multi-scale stacked hourglass network for human pose estimation (2018)
7. Hara, K., Liu, M.Y., Tuzel, O., Farahmand, A.M.: Attentional network for visual object detection. arXiv preprint [arXiv:1702.01478](https://arxiv.org/abs/1702.01478) (2017)
8. Huang, F., Zeng, A., Liu, M., Qin, J., Xu, Q.: Structure-aware 3D hourglass network for hand pose estimation from single depth image. arXiv preprint [arXiv:1812.10320](https://arxiv.org/abs/1812.10320) (2018)
9. Insafutdinov, E., et al.: Arttrack: articulated multi-person tracking in the wild. In: CVPR (2017)
10. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 34–50. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_3
11. Li, L., Tang, S., Deng, L., Zhang, Y., Tian, Q.: Image caption with global-local attention. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
12. Lifshitz, I., Fetaya, E., Ullman, S.: Human pose estimation using deep consensus voting. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 246–260. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_16

13. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
14. Pishchulin, L., et al.: DeepCut: joint subset partition and labeling for multi person pose estimation. In: CVPR (2016)
15. Sun, G., Ye, C., Wang, K.: Focus on what's important: self-attention model for human pose estimation. arXiv preprint [arXiv:1809.08371](https://arxiv.org/abs/1809.08371) (2018)
16. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation (2014)
17. Wang, W., Shen, J.: Deep visual attention prediction. *IEEE Trans. Image Process.* **27**, 2368–2378 (2018)
18. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016)
19. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: ICML (2015)
20. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: ICCV (2017)
21. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: CVPR (2016)
22. Zhang, D.Z., Liu, C.C.: A visual attention based object detection model beyond top-down and bottom-up mechanism. In: ITM Web of Conferences (2017)