Anti-Confusing: Region-Aware Network for Human Pose Estimation

Xuan Cao Yanhao Ge Ying Tai Wei Zhang
Jian Li Chengjie Wang Jilin Li Feiyue Huang
Tencent Youtu Lab
{marscao, halege, yingtai, gavinwzhang}@tencent.com
{swordli, jasoncjwang, jerolinli, garyhuang}@tencent.com

Abstract

In this work, we propose a novel framework named Region-Aware Network (RANet) to achieve anti-confusing, including heavy occlusion, nearby person and symmetric appearance, for human pose estimation. Specifically, our proposed method addresses three key aspects for human pose estimation, *i.e.*, data augmentation, feature learning and prediction fusion. First, we propose Parsing-based Data Augmentation (PDA) to generate abundant data with confusing textures. Second, we not only propose a Feature Pyramid Stem (FPS) module to learn better low-level features in lower stage; but also incorporate an Effective Region Extraction (ERE) module to investigate better human body-specific features. Third, we introduce Cascade Voting Fusion (CVS) to explicitly leverage the visibility to exclude the deflected predictions and achieve final accurate pose estimation. Experimental results demonstrate the superiority of our method against the state of the arts with significant improvements on two popular benchmark datasets, including MPII and LSP.

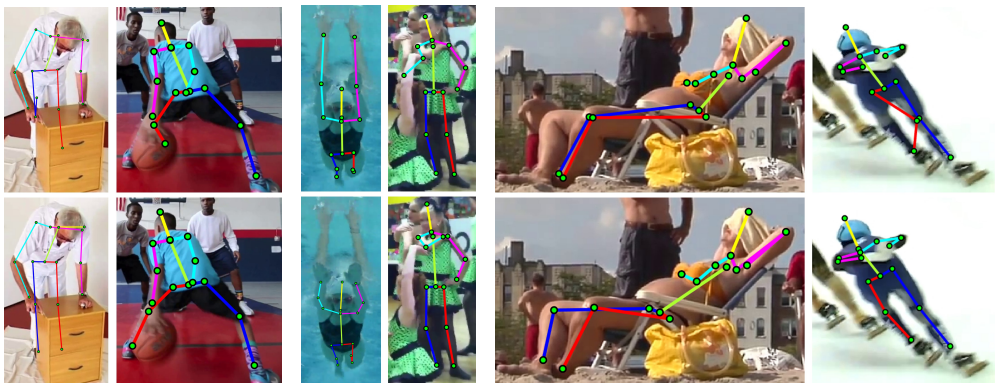
1 Introduction

Human Pose Estimation (HPE) localizes human anatomical keypoints (joints), which plays an important role in a variety of high-level vision tasks, such as action recognition [27], human tracking [29], human image synthesis [16], *etc*. The recent advances show that Deep Convolutional Neural Networks (DCNN) have achieved state-of-the-art performance [32, 22, 24, 13]. However, these networks can still be easily confused by three kinds of challenging cases: heavy occlusion, nearby person and symmetric appearance, which we term as *confusing texture*, as shown in Fig. 1.

Confusing texture in heavy occlusion and nearby person is self-explanatory, which is widely studied and addressed in [32, 13, 8, 24]. Regarding to the confusing texture in symmetric appearance, we know that human appearance is highly symmetric, such as the shoes, clothes, trousers and so on. We experimentally observe that the symmetrical human appearance can easily confuse the network, especially in case of crossed arms and legs. To our best knowledge, it is the first work to identify the concept of symmetric appearance and discuss its influence on human pose estimation.

How to resolve the confusing texture is a core problem in human pose estimation. Typically, the previous works focus on three aspects to achieve anti-confusing: First, various kinds of *data augmentation* schemes are adopted, including scaling, rotating, flipping, pose synthesis [8], keypoint masking [13], *etc*. Second, effective *feature learning* is pivotal. For example, the popular hourglass model [17] and its variants [3, 29, 32, 30, 25, 22] stack different kinds of *high-low-high* sub-networks, and the repeated bottom-up and top-down processing effectively learn the high-level features for heatmap prediction. Third, adaptive *prediction fusion* on heatmaps or coordinates at different

Preprint. Work in progress.



(a) Heavy occlusion (b) Nearby person (c) Human symmetric similar appearance

Figure 1: Illustration of confusing texture in human pose estimation. **Top**: Pose predictions from DLCM [24], which is confused by various confusing textures. **Bottom**: Our predictions that successfully resolve the confusing texture. In case of heavy occlusion, DLCM gets error prediction on ankles, while ours successfully captures the details of basketball shoe and white trouser to give correct prediction. For nearby person, DLCM is confused by the surrounding arms, while ours learns to resist the interference. Moreover, due to human's symmetric similar appearance, DLCM becomes confused between the autologous knees, while ours correctly distinguishes the symmetric joints.

stacks [30, 32] is a key step to improve accuracy of both easy and hard joints. Despite the great success achieved by the above methods, none of these methods address the three aspects simultaneously.

To address the above issues, we propose a novel framework, namely Region-Aware Network (RANet), which comprehensively address the three aspects in human pose estimation at the same time. First, to deal with various kinds of confusing textures, we propose a *Parsing based Data Augmentation* (PDA) scheme that builds a semantic body part pool, in which the body parts from unknown persons may be mounted on top of the current person. Second, different from the previous methods that focus on learning effective high-level features, we emphasize the important role of the low-level features, and thus propose a *Feature Pyramid Stem* (FPS) module to make full use of the input images under different resolutions. Third, after getting the prediction at different stages, we propose *Cascade Voting Fusion* (CVF) to explicitly utilize visibility cues to exclude the deflected predictions and merge the rest predictions in a weighted manner for final accurate pose estimation. To make full use of the effective human body region in the input image, we further propose an Effective Region Extraction (ERE) module to crop the effective region (as shown in Fig. 2), according to the estimated bounding box calculated from the intermediate prediction, from the original large image so as to complement more useful details of the human body. The proposed ERE module is complementary to the other three modules since it provides more compact human body information, and thus reduces the effect of the background to extract target-specific features, which are also helpful for the final fusion strategy.

In summary, the main contributions of this work are four-folds:

• We identify different kinds of confusing textures, and then propose a novel parsing based augmentation scheme, which achieves significant improvement on difficult joints, like elbow, wrist, knee and ankle.

• We propose a feature pyramid stem module to learn better low-level features with very small increase of complexity.

• We design a cascade voting fusion module that explicitly leverages the visibility as confidence, which is effective, no additional burden and can be easily applied on other network structures.

• We further introduce the effective region extraction module to investigate more useful details of the target, and achieve state-of-the-art performance on two representative benchmark datasets, *i.e.*, MPII human pose dataset (MPII) [1] and extended Leeds Sports Poses (LSP) [12].

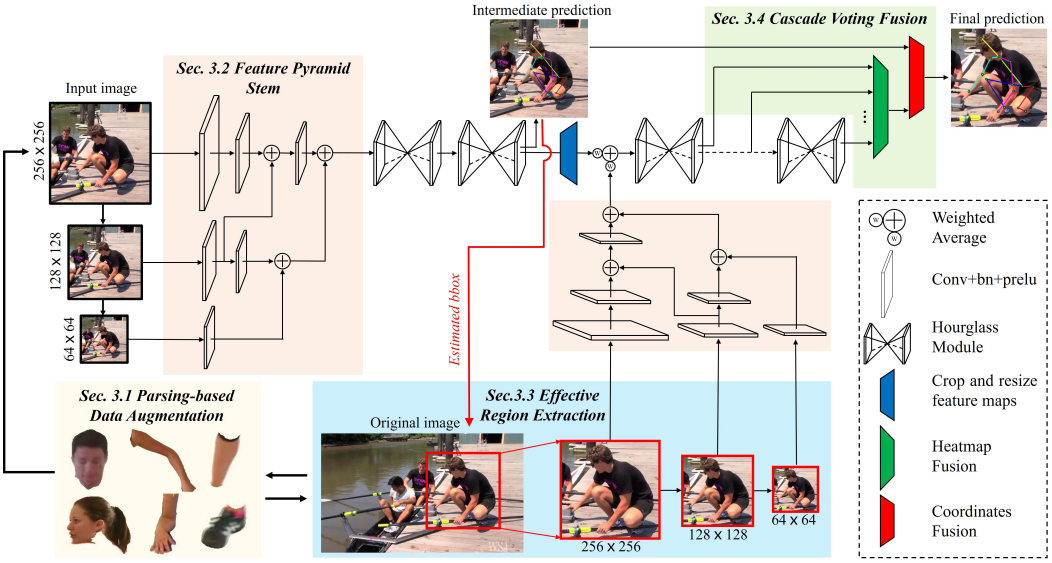


Figure 2: Framework of the proposed Region-Aware Network.

2 Related Work

Previous deep learning based human pose estimation works can be roughly divided into two categories. The first group directly regresses the location coordinates of joints [26, 31], called regression-based methods; while the second group predicts heatmaps followed by estimating joint locations according to the peak or integration response of heatmap [28, 17], termed heatmap-based methods. Our work is closely related to the second group while differing from three perspectives.

Data augmentation Conventional data augmentation methods on human pose estimation task [17, 3, 32] mainly performs scaling, rotating and flipping, *etc* on the training images. Recently, PoseRefiner [8] synthesizes possible input poses and improve the network to identify the erroneous body joint predictions and to refine them. MSR-net [13] introduces keypoint masking to simulate the hard training samples. Different from the previous data augmentation strategies, we propose a novel parsing-based data augmentation scheme that takes advantage of the semantic segmentation mask for limb-region duplication.

Feature Learning Most previous networks [17, 29, 22, 32, 3] focus on learning effective high-level features for heatmap prediction, which incorporates the stacked hourglass structure [17] that includes down-sampling for feature encoding and up-sampling for heatmap decoding. Before entering the hourglass structure, a rough stem module that converts the input image to smaller feature maps is adopted to reduce the complexity. For instance, stacked hourglass [17] accepts input in resolution of 256×256 while generates heatmap of 64×64. Simple-baseline [29] and HRNet [22] directly resize the high-resolution input images to low-resolution ones. However, the rough stem module may not make full use of the effective pixel-level information from the raw input images. In contrast, we on one hand propose a multi-scale Feature Pyramid Stem module for better low-level feature learning; on the other hand propose an Effective Region Extraction module for better target-specific features.

Prediction Fusion Prediction fusion strategy is a common solution to improve the hard cases like heavy occlusion, complex body pose or cluttered background. Zhang *et. al* [32] design a Cascade Prediction Fusion (CPF) network that takes all prediction maps in different semantic levels into considerations for final prediction. Yang *et. al* [30] concatenate coarse output heatmap with raw input for further point refinement. Compared with these methods, our work for the first time imports the visibility information of each key point as the basis to influence not only training label generation but also final prediction fusion.

Figure 3: Parsing based data augmentation. Firstly we apply human parsing on the training images. Then a data pool filled with segmented body parts texture is built, as the red arrow shown. Finally the body parts are mounted on the training data, as the blue arrow illustrated. The body parts texture may occlude the joints or semantically repeat around the joints.

3 Method

3.1 Parsing based Data Augmentation

Semantic Parsing For common human pose estimation methods [17, 29, 24, 22, 32], data augmentation on scale, rotation, flipping are applied. These augmentation methods are not strong enough to provide robustness against confusing texture resulted from heavey occlusion, nearby person and symmetric appearance. Here, we propose a novel augmentation method by firstly parsing the human body [14, 9] and then semantically add the part texture on the training data, as shown in Fig. 3. The part texture from parsing could occlude the current person' human body, and thus act as the confusing parts around the original joints.

Comparison to Previous Augmentation Pose-Refiner [30] synthesizes hard joints by only switch keypoints location. Instead, our method synthesizes the texture by simulating the challenging cases. Compared to the cropped patch in Keypoint-Masking [13], our proposed PDA provides segmented body parts without background texture. More importantly, our method segments all training images and build a data pool filled with various semantic body parts. Therefore, the body parts from various persons may be mounted on the current person. In addition, we mount the part texture with semantics. For example, the shank texture could only be mounted on or around keens/ankles. As the result, our training data consist of much stronger confusing texture which forces our network to learn ability of anti-confusing.

3.2 Feature Pyramid Stem

Stronger Feature in Lower Level Due to the limited model capacity and computation source, typically a feature map in much lower resolution is passed through the heatmap prediction subnetwork, such as 64×64 in stacked hourglass [17] and 64×48 in simple baseline [29]. It's inevitable to squeeze the information during converting the input image to the lower resolution feature map. We introduce a feature pyramid stem module to learn stronger low-level features. Based on an original downsample network, our feature pyramid stem repeatedly performs the downsample process on different input resolutions until reaching the target feature resolution. Specifically, with the input image of size $W \times H \times 3$, the proposed FPS consists of N stems to extract feature. The input of i^{th} stem is in resolution of $1/(2^{i-1}) * [W, H]$, $i \in [1, N]$. We further add ectopic paths over different stems to enhance the communication between features with the same resolution. For example, based on the stem network in stacked-hourglass [17] that accepts input resolution of 256×256 and output heatmap in resolution of 64×64. The original downsample process consist of stride $\times 2$ convolution and max pooling. We set $N = 3$ and the 1^{th} stem keeps the same with the original downsample process. The 2^{th} stem accepts input resolution 128×128 followed by the similar downsample process except the max pooling. The input resolution of the 3^{th} stem is 64×64 and stride $\times 1$ is

applied. Fig. 2 shows the illustration as the orange-masked module. More than this example, the proposed FPS could be widely applied on various feature extraction network.

Relationship to Previous Methods Most existing HPE network architectures consist of *high-low-high* sub-networks that first decline and then raise the feature resolution, like stacked hourglass [17] and simple baseline [29]. Previous works focus on redesigning the *high-low-high* sub-network, such as pyramid residual module [30] and densely connected U-nets [25]. These modified *high-low-high* sub-networks indeed produce better features in higher stage, but fail to pay attention to the feature learning in lower stage. In fact, the features in lower stage play an important role in the following heatmap prediction. The proposed FPS actually reinforce the lower-level feature learning.

3.3 Effective Region Extraction

Due to the highly degree of freedom in human pose configure, like squatting, up-right, lie-low, the limbs could reach to different scopes. Given an input image of 256×256, the full body may occupy different region sizes. In some cases, the pixel region of human appearance could be very small, then much details fail to be fed into the network. Lost of details is one of the main reason that results in confusing. For top-down human pose estimation [20, 11], an extra detector is applied at first. Previous work [7] has shown that the subtle deviation of bounding box could result in serious errors in pose estimation. We found the heatmaps in lower stages could sever as a free detector to support predictions in the following stages. Given the heatmaps from lower stage, we can easily infer the bounding box of human body. Sequentially the minimal region that contains whole body is cropped from the original image. Obviously the cropped image contains uttermost region of human body. As the blue-masked module shown in Fig. 2, the cropped image pass through another feature pyramid stem. The feature maps from two FPSs are then merged as the input for the following stacked hourglasses. It should be noted that the two FPSs has exactly the same network structure but don't share parameters.

3.4 Cascade Voting Fusion

During Inference Stacked hourglass outputs heatmap at each stack. Previous works simply chooses the heatmap of last stack as final prediction [22, 29, 23, 4]. In our experiments, we found the heatmaps in different stages carry varied information. How to adaptively fuse these information is a key problem for improving accuracy of both easy and hard joints. We propose Cascade Voting Fusion module (CVF) to make use of the auxiliary heatmaps. Compared to the fusion method in [32], the proposed CVF explicitly utilize visibility cues to exclude the deflected prediction and weighted merge the rest prediction for final accurate pose estimation. It should be noted that our fusion strategy can be easily built on most multistage pose estimation frameworks.

During Training To support the voting fusion, we add supervision with visibility during training. The heatmap is generated according Gaussian-like distribution as $e^{-\frac{(x-u)^2*(y-v)^2}{2\sigma^2}}$, where σ is the standard deviation and the main energy (about 99.73%) is distributed in [-3σ, +3σ]. Obviously, larger σ results in the larger radius of main energy. In case of larger radius, the network can more easily learn heatmap but in lower accuracy. We classify visibility into three status: visible, occluded and outer. For visible joints, we set the heatmap with $\sigma = 1$ as supervision. In order to easily learn occluded joints, we set twice value for the $\sigma = 2$. And heatmap of all zeros is provided as supervision for outer joints.

4 Experiments

Datasets We evaluate our method on two representative benchmark datasets including MPII human pose dataset (MPII) [1] and extended Leeds Sports Poses (LSP) [12]. MPII consists of 25,000 images with over 40,000 annotated poses. We split training and validation sets following [17]. The LSP dataset consists of 11,000, training images and 1,000 testing ones from sport activities.

Data augmentation During training, the input image is cropped according to the approximate human center and scale, and warped to size 256×256. We augment the dataset by randomly scaling in [0.75, 1.25], rotation in [-60, +60] degree, horizontal flipping and color adjustment. As mentioned

in section 3.1, we further mount the part texture from body parsing on the input image. For the wrong parsing results, we simply remove the misshapen parts which takes several hours by one person. During testing, we crop the image with the given rough body center and scale. For LSP dataset, the size and center of the image are utilized as rough scale and center.

Evaluation Criteria For fair comparison, we follow the common evaluation criteria. On the LSP dataset, we use the Percentage Correct Keypoints (PCK) to evaluate results as done in [32, 30]. For MPII, we use PCKh[1], a modified version PCK, which normalize the distance errors with respect to the size of head.

Implementation Details The network is implemented on PyTorch with optimizer RMSProp. We train the network in 250 epochs and batch size is 20. The learning rate starts at 0.0005 and decrease by 2 times at 20^{th}, 50^{th}, 100^{th}, 150^{th}, 200^{th} epoch. Following [6, 30, 24, 22, 32], a six different scale testing procedure is performed with horizontal flipping.

4.1 Results on MPII

Accuracy Tab. 1 lists our results on MPII test set. Our method yields result 92.9% PCKh@0.5, which is the highest on this dataset at the time of paper submission. It is noteworthy that, for the joints on symmetric body parts, including elbow, wrist, knee and ankle, our method achieve significant improvement compared to the state-of-the-art. It proves the effectiveness on resolving confusing texture resulted from symmetric appearance.

Table 1: Performance comparisons on the MPII test set (PCKh@0.5)

Method (%)	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Wei *et al.* [28] (CVPR'16)	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat *et al.* [2] (ECCV'16)	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell *et al.* [17] (ECCV'16)	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Ning *et al.* [19] (TMM'17)	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Luvizon *et al.* [15] (arXiv'17)	98.1	96.6	92.0	87.5	90.6	88.0	82.7	91.2
Chu *et al.* [6] (CVPR'17)	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou *et al.* [5] (arXiv'17)	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Chen *et al.* [3] (ICCV'17)	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Yang *et al.* [30] (ICCV'17)	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Xiao *et al.* [29] (ECCV'18)	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
Ke *et al.* [13] (ECCV'18)	98.5	96.8	92.7	88.4	90.6	89.4	86.3	92.1
Tang *et al.* [24] (ECCV'18)	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
Nie *et al.* [18] (CVPR'18)	98.6	96.9	93.0	89.1	91.7	89.0	86.2	92.4
Sun *et al.* [22] (CVPR'19)	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
Zhang *et al.* [32] (arXiv'19)	**98.6**	97.0	92.8	88.8	91.7	89.8	86.6	92.5
RANet (Ours)	98.5	**97.0**	**93.4**	**89.8**	**92.0**	**90.3**	**87.6**	**92.9**

Visualize In Fig. 4, we visualize pose estimation of baseline [1] [24] and our model. The baseline model is confused by heavy occlusion, nearby person and symmetric appearance. Our model successfully resolve the confusing texture in the three challenging situations. As shown in *col.* 1 and 2 of Fig. 4, the wrists are occluded and very limited hand / arm textures are exposed. The proposed Effective Region Extraction module help the network to see more details, and make it possible for following modules to learned the looming cues. Induced by the nearby person, as shown in *col.* 3 and 4 of Fig. 4, the baseline model predict ankles on other person incorrectly. Our parsing-based data augmentation force the network to learn the ability of anti-confusing for nearby body parts. In addition, the proposed Cascade Voting Fusion module excludes the incorrect prediction to some extent and improve the accuracy of final prediction. When the symmetric similarity meets image degradation, in *col.* 5 and 6 of Fig. 4, it's more difficult to distinguish the symmetric joints. In such cases, our Feature Pyramid Stem module learned stronger low-level feature from the input image and provide more useful information for final heatmap prediction. More results are available in supplementary material.

[1] `http://www.ece.northwestern.edu/~wtt450/project/ECCV18_DLCM/`

Figure 4: Qualitative results in the MPII dataset. Top row is the pose predictions obtained from DLCM [24] and bottom row is ours

4.2 Results on LSP

Tab. 1 shows our results on LSP test set. Following previous methods [6, 30], we add the MPII training set to the extended LSP training set. Our method outperforms the state-of-the-art and still maintain our competitive edge on the symmetric human joints, elbow, wrist, knee and ankle.

Table 2: Performance comparisons on the LSP test set (PCK@0.2)

Method (%)	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Insafutdinov *et al.* [10] (ECCV'16)	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Wei *et al.* [28] (CVPR'16)	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bulat *et al.* [2] (ECCV'16)	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Chu *et al.* [6] (CVPR'17)	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Chen *et al.* [3] (ICCV'17)	98.5	94.0	89.8	87.5	93.9	94.1	93.0	93.1
Yang *et al.* [30] (ICCV'17)	98.3	94.5	92.2	88.9	94.4	95.0	93.7	93.9
Peng *et al.* [21] (CVPR'18)	**98.6**	95.3	92.8	90.0	94.8	**95.3**	94.5	94.5
Zhang *et al.* [32] (arXiv'19)	98.4	94.8	92.0	89.4	94.4	94.8	93.8	94.0
RANet (Ours)	98.5	**95.5**	**93.8**	**90.5**	**95.1**	95.2	**94.5**	**94.7**

5 Conclusion

We presented Region-Aware Network (RANet) to effectively resolve the confusing texture in human pose estimation. Experimental results have demonstrated the effectiveness of our approach. The success stems from parsing-based data augmentation and three novel modules, *i.e.*, Feature Pyramid Stem (FPS), Effective Region Extraction (ERE) and Cascade Voting Fusion (CVF). FPS reinforce feature learning in lower stage which provides more useful information for the following heatmap prediction. ERE detect human body for free and extract the uttermost pixel region which help the network to see more details. CVF exclude the deflected predictions and adaptively fuse the rest prediction for more accurate pose estimation.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

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

[3] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1212–1221, 2017.

[4] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. FSRNet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Chia-Jung Chou, Jui-Ting Chien, and Hwann-Tzong Chen. Self adversarial training for human pose estimation. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 17–30. IEEE, 2018.

[6] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2017.

[7] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017.

[8] Mihai Fieraru, Anna Khoreva, Leonid Pishchulin, and Bernt Schiele. Learning to refine human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 205–214, 2018.

[9] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017.

[10] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[11] Sheng Jin, Xujie Ma, Zhipeng Han, Yue Wu, Wei Yang, Wentao Liu, Chen Qian, and Wanli Ouyang. Towards multi-person pose tracking: Bottom-up and top-down methods. In *ICCV PoseTrack Workshop*, volume 2, page 7, 2017.

[12] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, volume 2, page 5, 2010.

[13] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 713–728, 2018.

[14] Ting Liu, Tao Ruan, Zilong Huang, Yunchao Wei, Shikui Wei, Yao Zhao, and Thomas Huang. Devil in the details: Towards accurate single and multiple human parsing. *arXiv preprint arXiv:1809.05996*, 2018.

[15] Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *arXiv preprint arXiv:1710.02322*, 2017.

[16] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017.

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

[18] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan. Human pose estimation with parsing induced learner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[19] Guanghan Ning, Zhi Zhang, and Zhiquan He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, 20(5):1246–1259, 2018.

[20] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017.

[21] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S. Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[22] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019.

[23] Ying Tai, Yicong Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyue Huang, and Yu Chen. Towards highly accurate and stable face alignment for high-resolution videos. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[24] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 190–206, 2018.

[25] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, and Dimitris Metaxas. Quantized densely connected u-nets for efficient landmark localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 339–354, 2018.

[26] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[27] Chunyu Wang, Yizhou Wang, and Alan L Yuille. An approach to pose-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2013.

[28] 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*, pages 4724–4732, 2016.

[29] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baseline for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 466–481, 2018.

[30] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1281–1290, 2017.

[31] Xiang Yu, Feng Zhou, and Manmohan Chandraker. Deep deformation network for object landmark localization. In *European Conference on Computer Vision*, pages 52–70. Springer, 2016.

[32] Hong Zhang, Hao Ouyang, Shu Liu, Xiaojuan Qi, Xiaoyong Shen, Ruigang Yang, and Jiaya Jia. Human pose estimation with spatial contextual information. *arXiv preprint arXiv:1901.01760*, 2019.