# LEARNING RECURRENT STRUCTURE-GUIDED ATTENTION NETWORK FOR MULTI-PERSON POSE ESTIMATION

*Zhongwei Qiu[1*], Kai Qiu[2], Jianlong Fu[2], Dongmei Fu[1]*

[1]University of Science and Technology Beijing, Beijing, China

[2]Microsoft Research, Beijing, China

[1]qiuzhongwei@xs.ustb.edu.cn, fdm_ustb@ustb.edu.cn, [2]{kaqiu, jianf}@microsoft.com

## ABSTRACT

Multi-person pose estimation aims to localize tens of human joints (e.g., elbow, wrist, etc.) from multiple human bodies in an image. Existing approaches mainly adopt a two-stage pipeline, which usually consists of a human detector (i.e., generating a bounding box for each person) and a single person pose estimator (i.e., generating human joints from each bounding box). However, these approaches neglect the challenges of large pose variations and heavy occlusions in each bounding box, which often results in imprecise human joint localization. In this paper, we propose a structure-guided attention network (SGAN) for multi-person pose estimation. Specifically, a structured pose representation is encoded by learning a joint confidence map and a joint association map, which can be further refined by a structure-guided attention network (SGAN) in a recurrent way. Note that SGAN enables a deep neural network to take initial pose estimation as references, and to discover multi-scale pose features as completion, and thus the learning of pose structures can be reinforced. Extensive experiments show the best single-model results against the state-of-the-art approaches, with a relative $3.5\%$ mAP gain in the challenging COCO Keypoint dataset.

***Index Terms***— Pose estimation, Attention model

## 1. INTRODUCTION

Multi-person pose estimation aims to recognize and localize tens of human joints for each of persons appearing in a given image. This task has drawn extensive attention in recent years, which not only poses a fundamental challenge in research field, but can benefit a broad range of applications in industry field. For example, pose-guided human-computer interaction, video analysis via action recognition, and so on.

Promising progresses have been made on this topic, along with the advancement of Convolutional Neural Networks (CNNs). Recent state-of-the-art approaches can be usually divided into two categories: bottom-up approaches [1, 2, 3] and top-down approaches [4, 5, 6]. Bottom-up methods learn joint heatmaps through a CNN for a whole input image, and further group joints to constitute multiple human poses. Top-down methods propose to detect persons with bounding boxes, and conduct single-person pose estimation in each bounding box. Although most promising results have been achieved by top-down methods, the performance of human pose estimation is still limited by human detectors. For example, some bounding boxes can include multiple individuals, due to the large pose variations of a person and the fixed rectangle shape of a bounding box (shown in Fig. 1). These detection results often bring great challenges for the subsequent single-person pose estimator to recognize and localize precise human joints for a target person. The joints from different two persons as shown in Fig. 1 can be even detected and grouped together, which leads to an abnormal human pose estimation result.

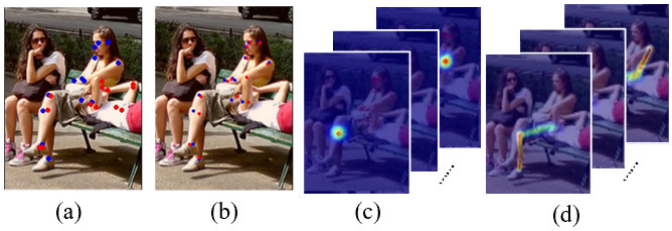


**Fig. 1**. Comparison of multi-person pose estimation results between a state-of-the-art baseline model [5] in (a) and the proposed SGAN model in (b). We can observe that by considering a joint confidence map in (c) and a joint association map in (d), SGAN generates more precise results for the target person in center than the baseline model. Blue/red points in (a) and (b) indicate the ground truth and the predicted 17 human joints. [Best viewed in color]

In this paper, we propose to integrate a recurrent structure-guided attention network (SGAN) into the top-down methods for multi-person pose estimation. The proposed network is composed of three closely-related modules. First, we propose to learn a deep structured feature representation for human pose by encoding human joints with their relations. This

*This work was done when Zhongwei Qiu conducted internship at Microsoft Research.

representation is specifically optimized by both a joint confidence map (JCM) and a joint association map (JAM). JCM models the joints of a person, while JAM models the limbs of a person. Second, as joints of different persons can be mixed from deep features, we further integrate all levels of features from a backbone encoder by a proposed structure-guided attention network (SGAN). Such a design can help discover and complete the most related human joint features to a target person, and thus imprecise location on joints can be relieved. Third, we propose to stack the SGAN in a recurrent way to progressively refine the learned structured human representation, which enables a multi-step completion even for challenging human instances.

Extensive experiments show the superior results of the proposed SGAN model on two widely-used multi-person pose estimation datasets. In particular, we have obtained the best single-model results against the state-of-the-art approaches, with a relative $3.5\%$ mAP gain in the challenging COCO Keypoint dataset.

## 2. RELATED WORK

Multi-person pose estimation is an active research field in recent years. Graph models or other models rely on hand-craft features [7, 8, 9, 10] are adopt in classical approaches. Superior performance has been improved by using deep convolution neural network [11, 12, 13] in recent years. Our human pose estimation method is also based on convolution neural network. Bottom-up [1, 2, 3] approaches and top-down [14, 15, 6, 5] approaches are two main methods to estimate human pose.

Bottom-up approaches firstly predict all joints in an image, and then allocate them to person based on structure information. Zhe Cao *et al*. [1] propose part affinity fields (PAFs) to map both orientation and location information between two parts, and then assemble them into different people. This structure constraint is helpful to predict more accurate joints in complex situations.

There are two steps for top-down methods. Human bounding boxes are firstly detected by a human detector. Then, joints are predicted by a human pose estimator in the image area of bounding box. Yilun Chen *et al*. [6] propose cascaded pyramid network to integrate multi-scale information. Heatmaps and offsets are predicted together in [14] and offsets are used to refine heatmaps. Bin Xiao *et al*. [5] add a deconvolution head network after ResNet , and obtain good performance as a baseline for human pose estimation. These top-down methods are benefited from accurate bounding box and multi-scale information.

Our work also adopt top-down method due to its excellent performance in many scenarios. Inspired by the previous great works, we propose a novel structure-aware loss JAM and a novel recurrent structure-guided attention network for multi-person pose estimation.

## 3. METHODS

Overview of our architecture is shown in Fig. 2. The architecture consists of a feed-forward module (including a ResNet to extract features and a decoder to generate heatmaps) and 3 recurrent structure-guided attention blocks. The feed-forward module is designed to generate JCMs and JAMs. JCM models the joints information while JAM models the limbs information. In order to extract more useful structural features which could help to refine the decoder feature, we design a recurrent structure-guided attention mechanism. The decoder feature $(D^{(k)})$ is used to calculate attention with the combination of all levels of ResNet and generate refined decoder feature $(D^{(k+1)})$ in a recurrent way. This recurrent structure-guided attention mechanism could continuously extract specific features to improve the accuracy of results.

As input, a color image of size $3 \times w \times h$ is cropped from origin image according to human bounding box $(w \times h)$. As output, a set of JCMs $(H)$ and JAMs $(L)$ are generated by a feed-forward network. The set $H = (H_1, H_2, ..., H_J)$ contains $J$ confidence maps for each person, where $H_j \in \mathbb{R}^{w \times h}, j \in \{1...J\}$. The set $L = (L_1, L_2, ..., L_J)$ contains $J$ vector fields for each person, where $L_j \in \mathbb{R}^{w \times h \times 2}, j \in \{1...J\}$. Finally, the JCMs are used to generate 2D joints coordinates by $argmax$ function.

### 3.1. Joint Confidence Maps and Joint Association Maps

The joint confidence maps(JCMs) and joint association maps(JAMs) are shown in Fig. 1 (c) and (d). JCMs are commonly used in pose estimation task because regressing coordinate directly is difficult. For one joint in each image, the ground truth JCMs $H_j^*(p)$ is generated as follows

$$H_j^*(p) \quad = \quad exp(-\frac{\|p - x_j)\|_2^2}{\sigma^2}) \tag{1}$$

where $p$ is a pixel location in image, and $x_j \in \mathbb{R}^2$ is the ground truth position of joint $j$. In prediction phase, JCM $H_j$ is generated by network. Then, joints coordinates are obtained by a $argmax$ function.

For $J$ joints, inspired by [1], we design $J$ unique JAMs as structure-aware map to help learning the association among joints. Strange pose, such as Fig. 1 (a), can be corrected by JAMs. Given a set of body parts, we try to build their relationship between two or the whole potential association among all parts. JAMs is a 2D vector fields for each limb which simultaneously considers location, size and orientation. Different from [1], our JAM is designed for joint. Thus, the number of JAMs is equal joints number while [1] is not. And the JAM for corresponding joint contains some limbs. That means one limb could appear in different JAMs, as long as the joint belong to the limb. In this way, some important joints (e.g., shoulder, hip, etc.) which connect more other joints could be strengthened specially by the corresponding JAMs.

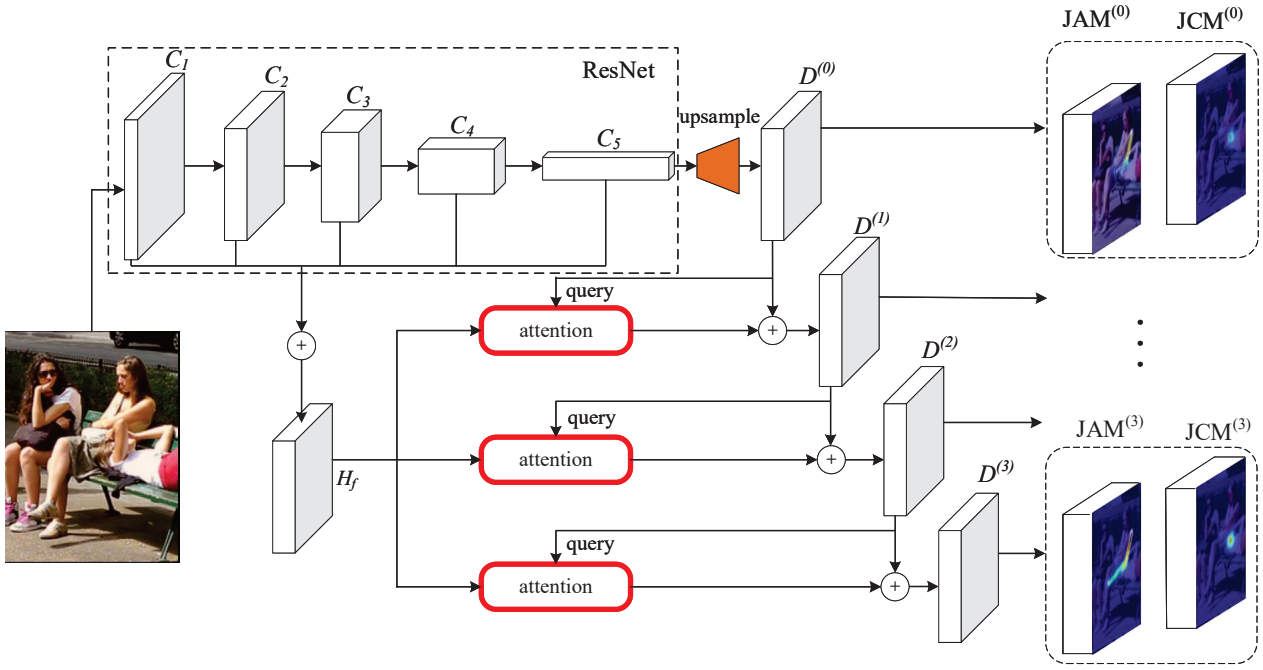


**Fig. 2**. The framework of the proposed recurrent SGAN for human pose estimation. Given a detected bounding box in the bottom-left, we extract deep features by ResNet (from $C_1$ to $C_5$). The encoded feature in $C_5$ is decoded into a structured feature map in $D^{(0)}$, which is further used to generate a joint confidence map (JCM) and a joint association map (JAM) by $1 \times 1$ convolution. Once $D^{(0)}$ has been obtained, we further strengthen this structured pose representation by calculating attention with the combination of all levels of features (i.e., $H_f$), and generate $D^{(1)}...D^{(3)}$ in a recurrent way. We can observe more significant joint relation prediction results in $JAM^{(3)}$ after refinement. "$\oplus$" denotes element-wise sum.

For one joint $j$, JAM $L_j$ defined as follows:

$$L_j = \sum_c^C Limb(c), \quad j \in Limb(c) \tag{2}$$

here, limb map $Limb$ is a human rigid limb connected by two joints, and is generated as:

$$Limb(p) = \begin{cases} v, & p \in Limb \\ 0, & p \notin Limb \end{cases} \tag{3}$$

The $Limb(c)$ fields is limited as follows:

$$0 \leq v \cdot (p - x_{j1}) \leq l_c \,\&\, |v_\perp \cdot (p - x_{j1})| \leq \sigma_l \tag{4}$$

where $\cdot$ denotes dot product, $\sigma_l$ is a distance of limb width in pixel, $v = (x_{j2} - x_{j1})/\|x_{j2} - x_{j1}\|_2$ is a unit vector in the direction of the limb and $v_\perp$ is a unit vector perpendicular to $v$ , $l_c = \|x_{j2} - x_{j1}\|_2$ is the length of limb in pixel.

### 3.2. Recurrent Structure-Guided Attention Network

Our JAMs are designed to model structure information of human pose. Spatial features are critical for JAMs. Therefore, we present recurrent structure-guided attention to extract specific spatial features. Different from CPN[6] that decoding multi-scale features to generate joint confidence maps, the recurrent SGAN automatically extract related spatial and structural features with an attention mechanism. Our model consists of a feed-forward module and 3 SGAN blocks.

**Feed-forward Module.** As shown in Fig. 2, feature maps($C_5$) are generated by ResNet, and was upsampled to the structural feature maps $D^{(0)}$ . In the last layer, JCMs $H_j$ and JAMs $L_j$ were outputted by $1 \times 1$ convolution, respectively. JCMs loss $E(H)$ is defined as follows:

$$E(H) = \sum_{j=1}^{J} \sum_p W_{H_j}(p) \cdot \|H_j(p) - H_j^*(p)\|_2^2 \tag{5}$$

JAMs loss $E(L)$ is defined as follows:

$$E(L) = \sum_{j=1}^{J} \sum_p W_{L_j}(p) \cdot \|L_j(P) - L_j^*(p)\|_2^2 \tag{6}$$

here, $H_j$ and $L_j$ are the outputs of feed-forward network. $H_j^*$ is the ground truth of JCMs. $L_j^*$ is the ground truth of JAMs. $W_H$ and $W_L$ are binary masks with $W = 1$ when joint $j$ is annotated at an image.

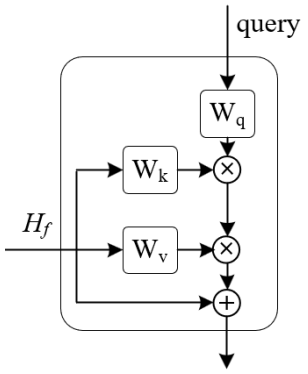


**Fig. 3**. A structure-guided attention block, which consists of a query (i.e., $D^{(0)}...D^{(2)}$ in Fig. 2), and a value (i.e., $H_f$). Different from [16], the goal of this design is to discover and complete related features to queries from all levels of features in multi-scale. "$\otimes$" denotes matrix multiplication. "$\oplus$" denotes element-wise sum. $W_q, W_k, W_v$ are $1 \times 1$ convolutions.

**Recurrent Structure-Guided Attention Mechanism.** In order to extract more useful spatial information and structural information for $D$ and refine outputs, we design a recurrent structure-guided attention network. The inputs of single $SGAN_k$ are multi-scale features $H_f$ and the structural features $D^{(k-1)}$. $D^{(k-1)}$ is used as a query to find specific features in multi-scale features($H_f = \sum_{i=1}^{5} C_i$) by attention module. Then, use the specific features to refine $D^{(k-1)}$, hence, JCMs and JAMs are refined. SGAN also works on the challenging human instances by recurrently stacking multiple SGAN blocks. This recurrent attention mechanism can significantly improve the accuracy of prediction due to more accurate spatial information and the strengthened structural features. The process of a single SGAN loop working is formulated as follows:

$$D^{(k)} = F(D^{(k-1)} + \Psi(D^{(k-1)}, H_f)) \quad (7)$$

here, $C_i$ represents the output of the $i$th layer in ResNet. $F(\cdot)$ represents $conv$ operation. $\Psi(\cdot)$ represents an structure-guided attention operation.

JCMs loss and JAMs loss are used in each SGAN block. For the entire network, total loss is

$$Loss = \sum_{k=0}^{K}(E^{(k)}(H) + E^{(k)}(L)) \quad (8)$$

**Attention Module.** As human joints can be mixed from $C_5$ due to its large receptive fields. We further integrate the features of $C_1 - C_5$ by structure-guided attention module. The architecture of structure-guided attention block is inspired by [16] and is shown in Fig.3. The structural features $D$ are inputted as a query, then computed a dot-product similarity with $H_f$. Then, transform the similarity matrix into feature maps. Finally, output the sum of residual $H_f$ and the transformed feature maps.

**Table 1**. Ablation study of the network we proposed on COCO $val2017$ dataset. SGAN means recurrent structure-guided attention blocks. JAM means JAMs loss. Backbone is ResNet-101. Input size is $256 \times 192$. No flip test.

| Method | SGAN | JAM | AP |
|---|---|---|---|
| SGAN-A | × | × | 70.4 |
| SGAN-B | × | ✓ | 70.5 |
| SGAN-C | 3 | × | 71.2 |
| SGAN-D | 1 | ✓ | 70.7 |
| SGAN-E | 2 | ✓ | 71.5 |
| SGAN-F | 3 | ✓ | **71.9** |

## 4. EXPERIMENTS

### 4.1. Datasets and Baseline

The COCO Keypoint Detection Task [17] requires localization of person joints in challenging uncontrolled conditions. There are more than 200k images and 250k person instances labeled with joints in the COCO train, validation and test dataset. The train and validation sets with 150k instances are publicly available. In experiments, only COCO $train2017$ dataset is used to train our model, which includes 57k images and 150k person instances, and COCO $val2017$ dataset is used to test and study ablation. MPII dataset includes around 25k images which contains over 40k people. We achieve the public state-of-the-art results on both COCO $val2017$ and MPII datasets with a fair comparison.

Object keypoint similarity (OKS) is defined on COCO evaluation. And the mean average precision (mAP) over ten OKS thresholds is employed as main competition metric, which is calculated from the distance between predicted joints and ground truth kepoints normalized by scale of the person. For MPII dataset, the mAP of joints based on PCKh is also used as comparison metrics.

Our baseline is [5], which gets a results of 70.4 AP(with flip test) based on ResNet-50 on COCO $val2017$ dataset.The architecture of [5] is a single feed-forward network. ResNet is the backbone of this network, and three deconvolution layers are connected behind $C_5$ to generate joint confidence maps. Then, joints coordinates are obtained by $argmax$ operation on JCMs. Although a great performance has been achieved by [5], there still is a lack of structural constraint for human pose. Thus, we propose the recurrent structure-guided attention network based on [5].

**Table 2**. Compared with Simple Baseline [5] on MPII dataset. 15 joints are annotated on MPII. The AP of symmetrical joints is the mean of left joint and right joint. Mean is mean AP based on PCKh threshold of 0.5.

| Method | Backbone | Head | Shoulder | Elbow | Wrist | Hip | Knee | Ankle | Mean |
|---|---|---|---|---|---|---|---|---|---|
| Simple-Baseline | ResNet-50 | 96.42 | 94.85 | 88.15 | 82.15 | 87.28 | 83.34 | 79.07 | 87.87 |
| Ours | | 96.39 | 95.41 | 88.77 | 83.25 | 88.02 | 84.51 | 80.56 | 88.67 |
| Simple-Baseline | ResNet-101 | 96.56 | 95.04 | 88.44 | 83.01 | 87.73 | 83.34 | 79.88 | 88.25 |
| Ours | | 96.18 | 95.11 | 88.68 | 83.07 | 87.83 | 84.28 | 80.49 | 88.46 |
| Simple-Baseline | ResNet-152 | 96.90 | 95.65 | 89.25 | 83.64 | 88.07 | 84.28 | 80.66 | 88.87 |
| Ours | | 96.76 | 95.81 | 89.84 | 84.34 | 88.9 | 85.19 | 81.48 | **89.41** |

### 4.2. Experiments Details

The person bounding boxes used in training are ground truth. The area overlapped with the fixed bounding box in the image is cropped and re-sized to a fixed resolution. ResNet [18] network is initialized by pre-trained model on ImageNet [19]. Adam optimizer is used for all models and the initially learning rate is 1e-3. We trained model for 140 epochs and learning rate was decreased to 1e-4 after 90 epochs, 1e-5 after 120 epochs. In test phase, a faster-RCNN human detector with the mAP 56.4 is used to get human bounding boxes. As introduced in section 3.1 , the coordinates of human joints are predicted by the $argmax$ operation on the JCMs of $SGAN_3$ . We finished our ablation experiments on COCO dataset. Test results on COCO and MPII dataset are provided by the model of just training on single dataset. Special explanation, no flip test is used in our all experiments for fair comparison.

**Table 3**. Comparison with CMU-Pose[1], Hourglass[4] , CPN[6] and Simple-Baseline[5] on COCO $val2017$ dataset. OHKM means Online Hard Keypoints Mining used in CPN.The results of [5] is obtained by the code released by the authors and no flip test is used for a fair comparison. Thus, it's lower than [5] reported in paper with flip test.

| Method | Backbone | Input Size | AP |
|---|---|---|---|
| CMU-Pose | - | $256 \times 192$ | 61.0 |
| 8-Hourglass | - | $256 \times 192$ | 66.9 |
| CPN | ResNet-50 | $256 \times 192$ | 68.6 |
| CPN(OHKM) | | | 69.4 |
| Simple-Baseline | | | 69.2 |
| Ours | | | **71.4** |
| Simple-Baseline | ResNet-101 | $256 \times 192$ | 70.4 |
| Ours | | | **71.9** |

### 4.3. Ablation Experiment

JAMs loss is designed to constrain the association between human joint and joint. And we hope JAMs loss could solve the problem that predicting some another person's joints to compose a strange pose because the model could learn more information about pose structure. As Fig. 1 (a) shown, baseline model SGAN-A falsely predicted another person's upper limb joints as right joints cause a very strange pose. However, our methods with adding JAMs structure-aware loss successfully force the predicted joints transfer to right person from near wrong person. And as their JAMs are shown in Fig. 1 (c) and (d), we can see that the JAMs has a right human pose that all JAMs are focus on the middle person, and all joints are on the middle person due to JAMs loss constraining. As a result, our SGAN have the ability to learn right human pose because it could learn a great deep structured feature representation for human pose with the supervision of JAMs.

However, as table 1 shown, model SGAN-B with adding JAMs loss has a little improvement compared with SGAN-A. The reason is that human joints could be mixed in deep features($C_5$) due to the large receptive fields. In other words, the human structural representation is not strong enough to include all structural features for the target person. Therefore, we propose a structured-guided attention network to discover the missed structural features. Then, integrate them with the deep structured representation. Compared with SGAN-A and SGAN-C, the results denoted that 1.2 AP is improved by our SGAN architecture. Compared with SGAN-C and SGAN-F, there is an improvement of 0.7 AP with adding JAMs loss based on the SGAN architecture. It verifies that JAMs help network learn a deep structured representation, but it needs to be strengthened by SGAN.

From the results of SGAN-D, SGAN-E and SGAN-F, we can find that there is a little improvement from stacking two SGAN blocks to three SGAN blocks. Thus, we choose SGAN-F in our other experiments.

### 4.4. Results on the MPII Dataset

We evaluate our SGAN-F on MPII dataset based on ResNet. The results are summarized on table 2. Compared with [5], our model achieve an improvement of 0.7 AP based on ResNet-50, an improvement of 0.2 AP based on ResNet-101, an improvement of 0.6 AP based on ResNet-152. These results verify the effectiveness of SGAN in MPII dataset.

### 4.5. Results on MS COCO Dataset

We evaluate our SGAN-F based on ResNet and compare with other methods on COCO $val2017$. Table 3 summarizes the results. We get an improvement of 2.2 AP based on ResNet-50 and an improvement of 1.4 AP based on ResNet-101 under a fair comparison with the sate-of-the-art method [5]. These results verify the effectiveness of SGAN in COCO dataset. And the case in Fig. 1 demonstrates the effectiveness of SGAN in complex situation(e.g., occlusions in crowed scenarios).

## 5. CONCLUSION

In this paper, we propose a recurrent structure-guided attention network for human pose estimation. To solve the challenges of large pose variations and occlusions, we propose 1) a structure loss (optimized by both JCM and JAM) to learn structured representation for human poses, 2) a structure-guided attention network to strengthen this structured representation from multi-scale features, 3) a recurrent fashion to stack SGAN into a holistic framework. Extensive experiments have shown superior performance of the proposed model (i.e., $3.5\%$ relative mAP gains by a single model on the COCO Keypoint dataset). In future, we will make deep exploration on the learning of human joint relations in an automatic way, rather than relying on human design.

## 6. REFERENCES


[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017, pp. 7291–7299.

[2] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *CVPR*, 2016, pp. 4929–4937.

[3] Alejandro Newell, Zhiao Huang, and Jia Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *NIPS*, 2017, pp. 2277–2287.

[4] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016, pp. 483–499.

[5] Bin Xiao, Haiping Wu, and Yichen Wei, "Simple baselines for human pose estimation and tracking," in *ECCV*, 2018.

[6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun, "Cascaded pyramid network for multi-person pose estimation," in *CVPR*, 2018, pp. 7103–7112.

[7] Xianjie Chen and Alan L Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *NIPS*, 2014, pp. 1736–1744.

[8] Ben Sapp and Ben Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *CVPR*, 2013, pp. 3674–3681.

[9] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool, "Human pose estimation using body parts dependent joint regressors," in *CVPR*, 2013, pp. 3041–3048.

[10] Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," in *CVPR*, 2016, pp. 845–853.

[11] Ita Lifshitz, Ethan Fetaya, and Shimon Ullman, "Human pose estimation using deep consensus voting," in *ECCV*. Springer, 2016, pp. 246–260.

[12] Tomas Pfister, James Charles, and Andrew Zisserman, "Flowing convnets for human pose estimation in videos," in *ICCV*, 2015, pp. 1913–1921.

[13] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *NIPS*, 2014, pp. 1799–1807.

[14] 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 *CVPR*, 2017, pp. 4903–4911.

[15] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu, "Rmpe: Regional multi-person pose estimation," in *ICCV*, 2017.

[16] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.