# Rethinking on Multi-Stage Networks for Human Pose Estimation

Wenbo Li[*1,2], Zhicheng Wang[*1], Binyi Yin[1], Qixiang Peng[1], Yuming Du[1,3], Tianzi Xiao[1,4], Gang Yu[1], Hongtao Lu[2], Yichen Wei[1], and Jian Sun[1]

[1]Megvii Inc. (Face++)
[2]Shanghai Jiao Tong University
[3]Beihang University
[4]Beijing University of Posts and Telecommunications
[1]{*liwenbo,wangzhicheng,yinbinyi,pengqixiang,duyuming,xiaotianzi,yugang,weiyichen,sunjian*}@megvii.com
[2]htlu@sjtu.edu.cn

## Abstract

*Existing pose estimation approaches can be categorized into* single-stage *and* multi-stage *methods. While a multi-stage architecture is seemingly more suitable for the task, the performance of current multi-stage methods is not as competitive as single-stage ones.*

*This work studies this issue. We argue that the current unsatisfactory performance comes from various insufficient design in current methods. We propose several improvements on the architecture design, feature flow, and loss function. The resulting multi-stage network outperforms all previous works and obtains the best performance on COCO keypoint challenge 2018. The source code will be released.*

## 1. Introduction

Human pose estimation problem has seen rapid progress in recent years using deep convolutional neural networks. Currently, the best performing methods [27, 14, 8, 39] are pretty simple, typically based on a *single-stage* backbone network, which is transferred from image classification task. For example, the COCO keypoint challenge 2017 winner [8] is based on Res-Inception [35]. The recent simple baseline approach [39] uses ResNet [15]. As pose estimation requires a high spatial resolution, up sampling [8] or deconvolution [39] is usually appended after the backbone networks to increase the spatial resolution of deep features.

Another category of pose estimation methods adopts a *multi-stage* architecture. Each stage is a simple light-weight
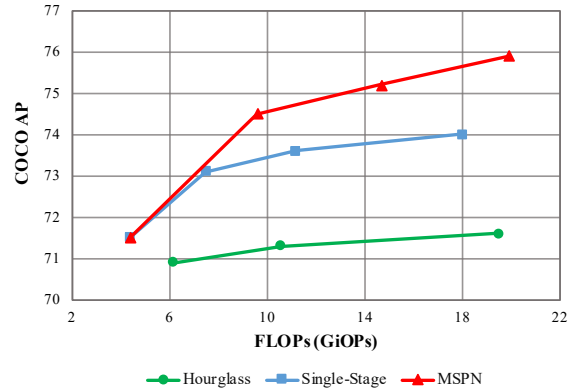
Figure 1. Pose estimation performance on COCO minival dataset of Hourglass [26], a single-stage model using ResNet [15], and our proposed MSPN under different model capacity (measured in FLOPs).

network and contains its own down sampling and up sampling paths. The feature (and heat) maps between the stages remain a high resolution. All the stages are usually supervised simultaneously to facilitate a coarse-to-fine, end-to-end training. Representative works include convolutional pose machine [38] and Hourglass network [26].

Pose estimation is more challenging than image classification. At an apparent look, the multi-stage architecture is more suited for the task because it naturally enables high spatial resolution, multi-level supervision and coarse-to-fine refinement. However, existing multi-stage methods do not perform as well as single-stage methods on COCO.

This work aims to study this issue. We point out that the current unsatisfactory performance of the multi-stage

methods is mostly due to the insufficiency in their design choices. With a number of improvements in the architecture, feature flow, and loss function, the potential advantage of a multi-stage architecture can be fully explored. New state-of-the-art performance is achieved, with a large margin compared to all previous methods.

Specifically, we propose a multi-stage pose estimation network (MSPN) with three novel observations and techniques. *First*, the single-stage module in the current multi-stage methods is far from optimal. For example, Hourglass uses equal width channels in all blocks for both down and up sampling. Such a design is inconsistent with the practice in current network architecture design such as ResNet [15]. We found that adopting the existing good network structure for down sampling path and a simple up sampling path is much better. *Second*, due to the repeated down and up sampling steps, information is more likely to lose and optimization becomes more difficult. We propose to aggregate features across different stages to strengthen the information flow and mitigate the difficulty in training. *Last*, observing that the pose localization accuracy is gradually refined during multi-stage, we adopt a coarse-to-fine supervision strategy in accordance. Supervision is also imposed over multi-scales to further improve training.

Each technique above introduces performance improvement. When they play in synergy, the resulting multi-stage architecture significantly outperforms all previous works. This is exemplified in Figure 1. For the single-stage method, its performance becomes saturated while increasing the network capacity. For the representative multi-stage method Hourglass [26], only a small performance gain is obtained after using more than two stages. For our proposed network, performance continues to improve with more stages.

On COCO keypoint benchmark, the proposed method achieves 76.1 average precision (AP) on test-dev. It significantly outperforms state-of-the-art algorithms. We obtain 78.1 AP on test-dev and 76.4 on test-challenge dataset in MS COCO 2018 Skeleton Challenge, which is 4.3 AP improvement on test-challenge benchmark compared with MS COCO 2017 Challenge winner.

## 2. Related Work

Pose estimation has underwent a long way as a basic research topic of computer vision. In the early days, hand-crafted features are widely used in classical methods [2, 33, 12, 32, 11, 42, 19, 29]. Recently, many approaches [4, 13, 30, 18, 6, 3] take advantage of deep convolutional neural network (DCNN) [21] to enhance the performance of pose estimation by a large step. In terms of network architecture, current human pose estimation methods could be divided as single-stage [27, 14, 8, 39] and multistage [38, 5, 25, 26, 41, 20] two categories.

**Single-Stage Approach** Single-stage methods [27, 14, 8, 39] are based on backbone networks that are well tuned on image classification tasks, such as VGG [34] or ResNet [15]. Papandreou *et al.* [27] designs a network to generate heat maps as well as their relative offsets to get the final predictions of the keypoints. He *et al.* [14] proposes Mask R-CNN to first generate person box proposals and then apply single-person pose estimation. Chen *et al.* [8] which is the winner of COCO 2017 keypoint challenge leverages a Cascade Pyramid Network (CPN) to refine the process of pose estimation. The proposed online hard keypoints mining (OHKM) loss is used to deal with hard keypoints. Xiao *et al.* [39] provides a baseline method that is simple and effective in the pose estimation task. In spite of their good performance, these methods have encountered a common bottleneck. Simply increasing the model capacity does not give rise to much improvement in performance. This is illustrated in both Figure 1 and Table 2.

**Multi-Stage Approach** Multi-Stage methods[38, 5, 25, 26, 41, 20] aim to produce increasingly refined estimation. They can be bottom-up or top-down. In contrary, single-stage methods are all top-down.

Bottom-up methods firstly predict individual joints in the image and then associate these joints into human instances. Cao *et al.* [5] employs a VGG-19 [34] network as a feature encoder, then the output features go through a multi-stage network resulting in heat maps and associations of the keypoints. Newell *et al.* [25] proposes a network to simultaneously output keypoints and group assignments.

Top-down approaches first locate the persons using detectors [31, 23, 22]. And a single person pose estimator is then used to predict the keypoints locations. Wei *et al.* [38] employs deep convolutional neural networks as feature encoder to estimate human pose. This work designs a sequential architecture composed of convolutional networks to implicitly model long-range dependencies between joints. Hourglass [26] is proposed to apply intermediate supervision to repeated down-sampling, up-sampling processing for pose estimation task. [41] adopts Hourglass and further design a Pyramid Residual Module (PRMs) to enhance the invariance in different scales. Many recent works [20, 7, 10, 37] are based on Hourglass and propose various improvements. While these multi-stage methods work well on MPII [1], they are not competitive on the more challenging tasks on COCO [24]. For example, the winners of COCO keypoint challenge on 2016 [27], 2017 [8] are all single-stage based, as well as the recent simple baseline work [39]. In this work, we propose several modifications on existing multi-stage architecture and show that the multistage architecture is better.
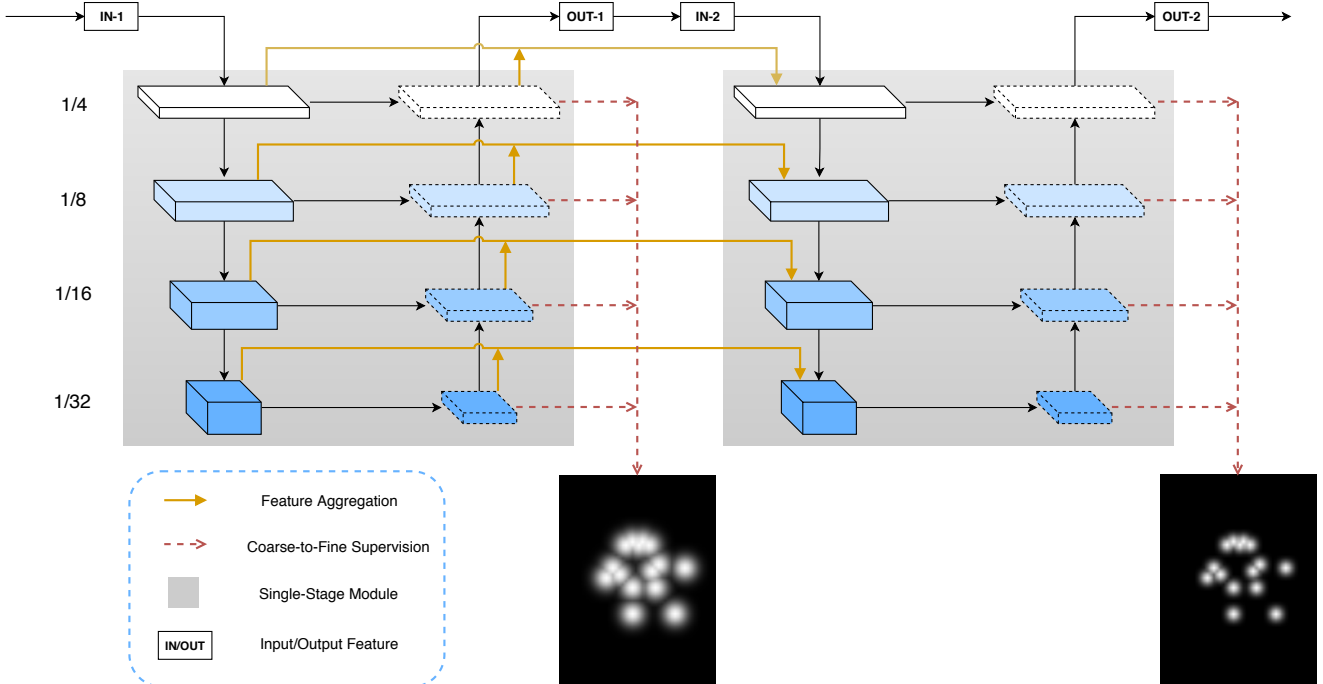
Figure 2. Overview of Multi-Stage Pose Network (MSPN). It is composed of two single-stage modules. A cross stage aggregation strategy (zoomed in Figure 3) is adopted between adjacent stages (Section 3.2). A coarse-to-fine supervision strategy further improves localization accuracy (Section 3.3).

## 3. Multi-Stage Pose Network

We adopt the top-down pipeline. It first performs human detection and then runs a single-person pose estimator on each detected human instance. A Multi-Stage Pose Network (MSPN) is proposed for pose estimation. A two-stage network is shown in Figure 2.

Our Multi-Stage Pose Network proposes three novel designs to boost the performance. First, we analyze the deficiency of the previous single-stage module and show that the state-of-the-art image classification network design can be exploited. Second, to reduce information loss, a feature aggregation strategy is proposed to propagate information from early stages to the later ones. At last, we introduce a coarse-to-fine supervision in our network. It gradually refines localization accuracy with more stages. Meanwhile, it makes full use of contextual information and enables more discriminative representations across scales. In the following sections, we will provide the details of each design respectively.

### 3.1. Effective Design of Single-Stage Module

Good design of single-stage module is crucial for the multi-stage network. Most recent multi-stage methods [41, 20, 7, 10, 37] are based on the Hourglass architecture [26]. However, as shown in Table 1, Hourglass simply stacks convolutional layers and the number of features remains

| Scale | Hourglass | | MSPN | |
|---|---|---|---|---|
| | top-down | bottom-up | top-down | bottom-up |
| 1/4 | 256 | 256 | 256 | 256 |
| 1/8 | 256 | 256 | 512 | 256 |
| 1/16 | 256 | 256 | 1024 | 256 |
| 1/32 | 256 | 256 | 2048 | 256 |

Table 1. Number of feature channels on each scale for single-stage modules of Hourglass and MSPN. 'Top-down' and 'bottom-up' represent down-sampling and up-sampling procedures respectively.

consistent during repeated down and up sampling procedures in a single stage. It results in a relatively poor performance seen from Figure 1. An effective single-stage module as shown in Figure 2 is a U-shape architecture in which features extracted from multiple scales are utilized for predictions. Drawing experience from popular backbones [34, 15, 40, 17, 9, 16] such as VGG [34], we double the number of features every time there is a down-sampling operation, which could effectively reduce the information loss. Besides, computing capacity is mainly allocated to the down-sampling unit rather than up-sampling unit. It is reasonable since we aim to extract more representative features in the down-sampling process and the lost information can hardly be recovered in the up-sampling procedure. Therefore, increasing the capacity of down-sampling unit is usu-
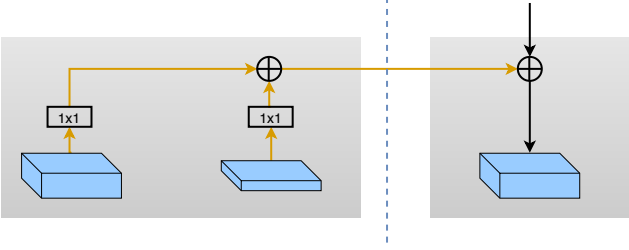
Figure 3. Cross Stage Feature Aggregation on a specific scale. Two $1 \times 1$ convolutional operations are applied to the features of previous stage before aggregation. See Figure 2 for the overall network structure.

ally more effective.

### 3.2. Cross Stage Feature Aggregation

A multi-stage network is flavored in that information lost during repeated up and down sampling. To mitigate this issue, a cross stage feature aggregation strategy is used to propagate multi-scale features from early stages to the current stage in an efficient way.

As is shown in Figure 2, for each scale, two separate information flows are introduced from down-sampling and up-sampling units in the previous stage to the down-sampling procedure of the current stage. It is noted that a $1 \times 1$ convolution is added on each flow as shown in Figure 3. Together with down-sampled features of current stage, three components are added to produce fused results. With this design, the current stage can take full advantage of prior information to extract more discriminative representations. In addition, the feature aggregation could be regarded as an extended residual design, which are helpful dealing with the gradient vanishing problem.

### 3.3. Coarse-to-fine Supervision

In the pose estimation task, context is crucial for locating the challenging poses since it provides information for invisible joints. Besides, we notice that small localization errors would seriously affect the performance of pose estimation. Accordingly, we design a coarse-to-fine supervision, as illustrated in Figure 2. Specifically, the ground truth heat map for each joint is realized as a Gaussian in most previous works. In this work, we further propose to use different kernel sizes of the Gaussian in different stages. That is, an early stage uses a large kernel and a late stage uses a small kernel. This strategy is based on the observation that the estimated heat maps from multi-stages are also in a similar coarse-to-fine manner. Figure 4 shows a illustrative example. It demonstrates that the proposed supervision is able to refine localization accuracy gradually.

Besides, we are inspired that intermediate supervision could play an essential role in improving the performance of deep neural network from [36]. Therefore, we introduce
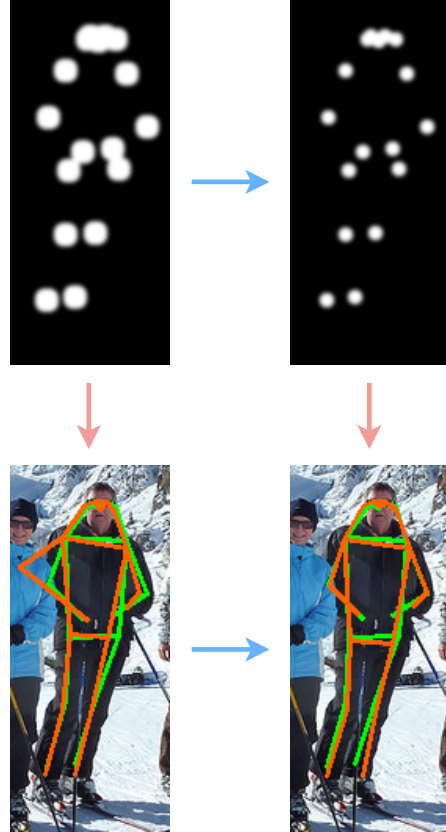


Figure 4. Illustration of coarse-to-fine supervision. The first row shows ground-truth heat maps in different stages and the second row represents corresponding predictions and ground truth annotations. The orange line is the prediction result and the green line indicates ground truth.

a multi-scale supervision to perform intermediate supervisions with four different scales in each stage, which could obtain substantial contextual information in various levels to help locate challenging poses. As shown in Figure 2, an online hard keypoints mining (OHKM) [8] is applied to the largest scale supervision in each stage. L2 loss is used for heat maps on all the scales.

## 4. Experiments

### 4.1. Dataset and Evaluation Protocol

MS COCO [24] is adopted to evaluate the performance of our framework. It consists of three splits: train, validation and test. Similar in [8], we aggregate the data of train and validation parts together, and further divide it into trainval dataset (nearly 57K images and 150K person instances) and minival dataset (5k images). They are separately utilized for training and evaluating. OKS-based mmAP (AP for short) is used as our evaluation metric [24].

## 4.2. Implementation Details

**Human Detector.** We adopted a state-of-the-art object detector MegDet [28] to generate human proposals. The MegDet is trained with full categories of MS COCO dataset[1]. Only human boxes out of the best 100 ones of all categories are selected as the input of single-person pose estimator. All the boxes are expanded to have a fixed aspect ratio of $4 : 3$.

**Training.** The network is trained on 8 Nvidia GTX 1080Ti GPUs with mini-batch size 32 per GPU. There are 90k iterations. Adam optimizer is adopted and the linear learning rate gradually decreases from 5e-4 to 0. The weight decay is 1e-5.

Each image will randomly go through a series of data augmentation operations including cropping, flipping, rotation and scaling. As for cropping, instances with more than eight joints will be cropped to upper or lower bodies with equal possibility. The rotation range is $-45° \sim 45°$, and scaling range is $0.7 \sim 1.35$. The image size is set $256 \times 192$ in Section 4.3, and $384 \times 288$ in Section 4.4.

**Testing.** A post-Gaussian filter is applied to the estimated heat maps. Following the same strategy as [26], we average the predicted heat maps of original image with results of corresponding flipped image. Then, a quarter offset in the direction from the highest response to the second highest response is implemented to obtain the final locations of keypoints. The pose score is the multiplication of box score and average score of keypoints, same as in [8].

## 4.3. Ablation Study

In this section, we provide an in-depth analysis of each individual design in our framework.

In order to show the effectiveness of our method in a clear way, we also perform corresponding experiments on Hourglass [26]. All results are reported on minival dataset. The input image size is $256 \times 192$.

### 4.3.1 Multi-Stage Architecture

First, we evaluate how the capacity of backbone affects the performance of pose estimation. In terms of the single-stage network in Table 2, we observe that its performance gets quickly saturated with the growth of backbone capacity. It is obvious that Res-101 outperforms Res-50 by 1.6 AP and costs a more 3.1G FLOPs, but there is only 0.5 gain from Res-101 to Res-152 at the cost of additional 3.7G FLOPs. For further exploration, we train a Res-254 network by adding more residual blocks on Res-152. Although the FLOPs of the network increases from 11.2G to 18.0G, there is an only 0.4 AP improvement. Therefore, it is not effective to adopt Res-152 or larger backbones for a single-stage network.

---

[1]We will release the detection results in the future.

| Method | Res-50 | Res-101 | Res-152 | Res-254 |
|---|---|---|---|---|
| AP | 71.5 | 73.1 | 73.6 | 74.0 |
| Flops(G) | 4.4 | 7.5 | 11.2 | 18.0 |

Table 2. Results of single-stage networks with different backbones on COCO minival dataset.

Then, we demonstrate the effectiveness of multi-stage architecture based on the proposed single-stage module. From Table 3, we can see that the performance of single-stage Hourglass [26] is poor. Adding one more stage introduces a large AP margin. It shows that a multi-stage network is potential. However, the improvement becomes small when four or eight stages are employed. This indicates the necessity of a more effective single-stage module. Our single-stage model is discussed in Section 3.1 and the performance with 71.5 AP on minival dataset demonstrates the superiority of our single-stage module. And our two-stage network further leads to a 3.0 improvement and obtains 74.5 AP. Introducing the third and fourth stage maintains a tremendous upward trend and eventually brings an impressive performance boost. These experiments indicate that MSPN successfully pushes the upper bound of existing single-stage and multi-stage networks. It obtains noticeable performance gain with more network capacity.

| Stages | Hourglass | | Stages | MSPN | |
|---|---|---|---|---|---|
| | FLOPs(G) | AP | | FLOPs(G) | AP |
| 1 | 3.9 | 65.4 | 1 | 4.4 | 71.5 |
| 2 | 6.2 | 70.9 | 2 | 9.6 | 74.5 |
| 4 | 10.6 | 71.3 | 3 | 14.7 | 75.2 |
| 8 | 19.5 | 71.6 | 4 | 19.9 | 75.9 |

Table 3. Results of Hourglass and MSPN with different number of stages on COCO minival dataset. MSPN adopts Res-50 in each single-stage module.

| Method | Res-50 | 2×Res-18 | L-XCP | 4× S-XCP |
|---|---|---|---|---|
| AP | 71.5 | 71.6 | 73.7 | 74.7 |
| FLOPs | 4.4G | 4.0G | 6.1G | 5.7G |

Table 4. Results of MSPN with smaller single-stage modules on COCO minival dataset. "L-XCP" and "S-XCP" respectively represent a small and a large Xception backbone.

Finally, we testify that our single-stage module can effectively adopt other backbones. We conduct more experiments on ResNet-18 and Xception [9] architectures. Results are illustrated in Table 4. It is clear that the two-stage network based on Res-18 obtains a comparable result with Res-50 with smaller FLOPs. Moreover, we design two Xception [9] backbones with different capacity, a large one (L-

| Method | Backbone | Input Size | AP | $AP^{50}$ | $AP^{75}$ | $AP^{M}$ | $AP^{L}$ | AR | $AR^{50}$ | $AR^{75}$ | $AR^{M}$ | $AR^{L}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CMU Pose [5] | - | - | 61.8 | 84.9 | 67.5 | 57.1 | 68.2 | 66.5 | 87.2 | 71.8 | 60.6 | 74.6 |
| Mask R-CNN [14] | Res-50-FPN | - | 63.1 | 87.3 | 68.7 | 57.8 | 71.4 | - | - | - | - | - |
| G-RMI [27] | Res-152 | 353×257 | 64.9 | 85.5 | 71.3 | 62.3 | 70.0 | 69.7 | 88.7 | 75.5 | 64.4 | 77.1 |
| AE [25] | - | 512×512 | 65.5 | 86.8 | 72.3 | 60.6 | 72.6 | 70.2 | 89.5 | 76.0 | 64.6 | 78.1 |
| CPN [8] | Res-Inception | 384×288 | 72.1 | 91.4 | 80.0 | 68.7 | 77.2 | 78.5 | 95.1 | 85.3 | 74.2 | 84.3 |
| Simple Baseline [39] | Res-152 | 384×288 | 73.8 | 91.7 | 81.2 | 70.3 | 80.0 | 79.1 | - | - | - | - |
| **Ours (MSPN)** | 4×Res-50 | 384×288 | **76.1** | **93.4** | **83.8** | **72.3** | **81.5** | **81.6** | **96.3** | **88.1** | **77.5** | **87.1** |
| CPN+ [8] | Res-Inception | 384×288 | 73.0 | 91.7 | 80.9 | 69.5 | 78.1 | 79.0 | 95.1 | 85.9 | 74.8 | 84.6 |
| MSRA+* | Res-152 | 384×288 | 76.5 | 92.4 | 84.0 | 73.0 | 82.7 | 81.5 | 95.8 | 88.2 | 77.4 | 87.2 |
| Ours (MSPN*) | 4×Res-50 | 384×288 | 77.1 | 93.8 | 84.6 | 73.4 | 82.3 | 82.3 | 96.5 | 88.9 | 78.4 | 87.7 |
| **Ours (MSPN+*)** | 4×Res-50 | 384×288 | **78.1** | **94.1** | **85.9** | **74.5** | **83.3** | **83.1** | **96.7** | **89.8** | **79.3** | **88.2** |

Table 5. Comparisons of results on COCO test-dev dataset. "+" indicates using an ensemble model and "*" means using external data.

| Components | | | **Hourglass** | **MSPN** |
|---|---|---|---|---|
| BaseNet | CTF | CSFA | | |
| √ | | | 71.3 | 73.3 |
| √ | √ | | 72.5 | 74.2 |
| √ | √ | √ | 73.0 | 74.5 |

Table 6. Ablation Study of MSPN on COCO minival dataset. 'BaseNet' represents a four-stage Hourglass or two-stage MSPN based on Res-50 with similar complexity, see Table 3. 'CTF' indicates the coarse-to-fine supervision. 'CSFA' means the cross stage feature aggregation.

| Setting | 1 | 2 | 3 |
|---|---|---|---|
| Kernel Size 1 | 7 | 5 | 7 |
| Kernel Size 2 | 7 | 5 | 5 |
| AP | 74.2 | 74.0 | 74.5 |

Table 7. Results of a two-stage MSPN with different supervision strategies on COCO minival dataset. The kernel size controls the fineness of supervision and a smaller value indicates a finer setting.

XCP) and a small one (S-XCP). The four-stage S-XCP outperforms the single large model with 1.0 in AP with similar complexity. These results demonstrate the generality of our single-module backbone.

#### 4.3.2 Cross Stage Feature Aggregation

To address the issue that a deep multi-stage architecture is vulnerable by information losing during repeating up-down sampling procedures, we propose a cross stage feature aggregation strategy. It is adopted to fuse different level features in adjacent stages and ensure more discriminative representations for current stage. Table 6 shows that the proposed feature aggregation strategy brings about a 0.3 gain from 74.2 to 74.5 for MSPN and a 0.5 improvement in terms of Hourglass, which demonstrates its effectiveness on dealing with aforementioned problems. At the same time, we can draw a conclusion that Hourglass tends to lose more information during forward propagation and our feature aggregation strategy can effectively mitigate this issue.

#### 4.3.3 Coarse-to-fine Supervision

In this part, we evaluate our coarse-to-fine supervision for both MSPN and Hourglass. The results are shown in Ta-

ble 6. It is clear that this strategy improves the performance of our network by a large margin from 73.3 to 74.2. First of all, this design aims to realize a coarse-to-fine detection procedure and the result demonstrates its effectiveness on further improving the accuracy of keypoint localization. In addition, it is reasonable that intermediate supervisions can take full advantage of contextual information across different scales. To demonstrate the applicability of this supervision in other multi-stage networks, we further apply this strategy to a four-stage Hourglass that is comparable with our two-stage MSPN in complexity, and finally obtains a 1.2 improvement in AP. In a word, the proposed coarse-to-fine supervision could largely boost the performance of pose estimation and be well adapted to other multi-stage networks.

Furthermore, we conduct several experiments to verify which level of supervision will have a higher efficiency in our network. As described in Section 3.2, we apply a Gaussian blur operation to each point on a heat map and a smaller kernel corresponds to a finer supervision. As shown in Table 7, we could see that either setting-1 or setting-2 will degrade the performance compared with the proposed coarse-to-fine supervision (setting-3). Especially, setting-2 even leads to a worse performance than the setting-1, which indicates that an appropriate supervision could make a difference to the final result.

| Method | Backbone | Input Size | AP | $AP^{50}$ | $AP^{75}$ | $AP^M$ | $AP^L$ | AR | $AR^{50}$ | $AR^{75}$ | $AR^M$ | $AR^L$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mask R-CNN* [14] | Res-50-FPN | - | 68.9 | 89.2 | 75.2 | 63.7 | 76.8 | 75.4 | 93.2 | 81.2 | 70.2 | 82.6 |
| G-RMI* [27] | Res-152 | 353×257 | 69.1 | 85.9 | 75.2 | 66.0 | 74.5 | 75.1 | 90.7 | 80.7 | 69.7 | 82.4 |
| CPN+ [8] | Res-Inception | 384×288 | 72.1 | 90.5 | 78.9 | 67.9 | 78.1 | 78.7 | 94.7 | 84.8 | 74.3 | 84.7 |
| Sea Monsters+* | - | - | 74.1 | 90.6 | 80.4 | 68.5 | 82.1 | 79.5 | 94.4 | 85.1 | 74.1 | 86.8 |
| MSRA SB+* [39] | Res-152 | 384×288 | 74.5 | 90.9 | 80.8 | 69.5 | 82.9 | 80.5 | 95.1 | 86.3 | 75.3 | 87.5 |
| **Ours (MSPN+*)** | 4×Res-50 | 384×288 | **76.4** | **92.9** | **82.6** | **71.4** | **83.2** | **82.2** | **96.0** | **87.7** | **77.5** | **88.6** |

Table 8. Comparisons of results on COCO test-challenge dataset. "+" indicates using an ensemble model and "*" means using external data.

### 4.4. Comparison with State-of-the-art Methods

As shown in Table 5, our single model trained by only COCO data achieves 76.1 AP on test-dev and outperforms other methods by a large margin in all metrics. Advocated by external data, MSPN leads to a 1.0 improvement resulting in 77.1 AP. And the ensemble model finally obtains 78.1. From Table 8, it is clear that our framework obtains 76.4 AP on the test-challenge dataset and shows its significant superiority over other state-of-the-art methods. Eventually, our method surpasses COCO 2017 Challenge winner CPN [8] and MSRA Sample Baseline [39] by 4.3 and 1.9 AP in test-challenge dataset respectively.

Finally, some pose estimation results generated by our method are shown in Figure 5. We can see that our MSPN handles crowd and occlusion situations as well as challenging poses effectively.

## 5. Conclusion

In this work, we propose a Multi-Stage Pose Network (MSPN) to perform multi-person pose estimation. It breaks the performance ceiling of the current methods and achieves state-of-the-art results on MS COCO datasets. We first verify the effectiveness of the multi-stage pipeline with well-designed single-stage modules in MSPN. Additionally, a coarse-to-fine supervision and a cross stage feature aggregation strategy are proposed to further boost the performance of our framework. Extensive experiments have been conducted to demonstrate its superiority over other current methods as well as its generalizability.

## References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. 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] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1014–1021. IEEE, 2009.

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

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

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016.

[6] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016.

[7] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[8] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. *arXiv preprint*, 2018.

[9] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint*, pages 1610–02357, 2017.

[10] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. *CoRR*, abs/1702.07432, 2017.

[11] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3041–3048, 2013.

[12] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik. Articulated pose estimation using discriminative armlet classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3342–3349, 2013.

[13] G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. In *European Conference on Computer Vision*, pages 728–743. Springer, 2016.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Figure 5. Visualization of MSPN results on COCO minival dataset.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[17] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 7, 2017.

[18] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[19] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, pages 1465–1472. IEEE, 2011.

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

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[22] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun. Detnet: A backbone network for object detection. *arXiv preprint arXiv:1804.06215*, 2018.

[23] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[25] A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, pages 2277–2287, 2017.

[26] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[27] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, volume 3, page 6, 2017.

[28] C. Peng, T. Xiao, Z. Li, Y. Jiang, X. Zhang, K. Jia, G. Yu, and J. Sun. Megdet: A large mini-batch object detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6181–6189, 2018.

[29] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013.

[30] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016.

[31] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[32] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 422–429. IEEE, 2010.

[33] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3681, 2013.

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[37] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas. Quantized densely connected u-nets for efficient landmark localization. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[38] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[39] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. *arXiv preprint arXiv:1804.06208*, 2018.

[40] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5987–5995. IEEE, 2017.

[41] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, 2017.

[42] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011.