A Cascaded Inception of Inception Network with Attention Modulated Feature Fusion for Human Pose Estimation

Submission ID: 2065

Abstract

Accurate keypoint localization of human pose needs diversified features: the high level for contextual dependencies and low level for detailed refinement of joints. However, the importance of the two factors varies from case to case, but how to efficiently use the features is still an open problem.

Existing methods have limitations in preserving low level features, adaptively adjusting the importances of different levels of features, and modeling the human perception process. This paper presents three novel techniques step by step to efficiently utilize different levels of features for human pose estimation. Firstly, an inception of inception (IOI) block is designed to emphasize the low level features. Secondly, an attention mechanism is proposed to adjust the importances of individual levels according to the context. Thirdly, a cascaded network is proposed to sequentially localize the joints to enforce message passing from joints of stand-alone parts like head and torso to remote joints like wrist or ankle. Experimental results demonstrate that the proposed method achieves state-of-the-art performance on both MPII and LSP benchmarks.

Human pose estimation aims to localize human parts, e.g., head, shoulders, wrists, and ankles, which plays a key role in analyzing human's behavior from images or videos. Accurate and efficient human pose estimation could facilitate varieties of applications such as video analysis (Zhang and Shah 2015), action recognition (Du, Wang, and Wang 2015), human computer interaction (Shotton et al. 2013) etc. But localizing human's joints is challenging due to deformation of human body, various appearances and occlusion.

To localize keypoints accurately, a model has to consider features from multiple scales. The high level features, which represent global position of human, indicate the contextual relationship between joints. They reveal a rough but robust pose especially in occluded cases, e.g. the left wrist of the man in Figure 1. The low level features provide detailed information that is useful for refining the positions of joints like ankles and wrists, such as the partially occluded left ankle of the man(Figure 1).

The trade-off between low- and high-level features is always depending on the situations. In Figure 1, to infer the position of the man's left wrist, which is completely occluded, high level features about the context of a larger region are required. Meanwhile, more details are required to be preserved in order to correctly locate the left ankle which is partially occluded.

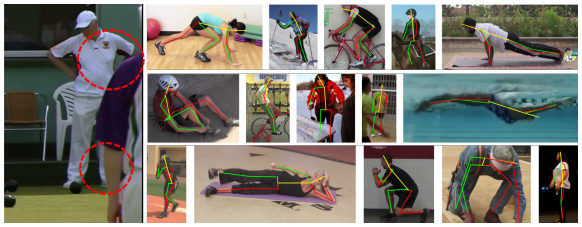

Figure 1: Results of our method on the MPII datasets. It is shown that the method is able to handle non-standard poses and resolve ambiguities when body parts are occluded.

Conventional human pose estimation approaches predict all joints at the same time from the same features (Wei et al. 2016; Newell, Yang, and Deng 2016). However, as we observed, some joints are distinct and less frequently occluded than others, which makes the estimation of them easier. So estimating each joint progressively and using features that rely on the predicted joints may simplify the learning task and provide more robust prediction of the harder cases. This cascaded framework has been adopted to deal with other visual tasks such as face alignment (Cao et al. 2014) and hand model fitting (Sun et al. 2015), but has not been used in human pose estimation yet.

According to the analysis above, in this work, we make three efforts to improve pose estimation by more efficiently capturing information from different scales. 1) As deeper network focuses on high-level features, some low-level features are lost. We design an inception of inception(IOI) block to preserve scale diversity. As the network goes deeper, the scale diversity could be preserved in each block. 2) The feature fusion is even dynamic in human pose estimation due to various appearance and occlusion. We upgrade IOI into an attention-modulated IOI(AIOI) block, which employs attention module to adapt the weights among features of different scales in each block. The adaptive feature scale makes it more robust to predict joints without structural context. 3) To enforce the message passing between joints, we

construct a cascaded joint network to predict joints progressively. Our framework sequentially treat the earlier estimated joints as an auxiliary feature, utilizing them to help estimating the latter joints. In our case, given the position of some joints can be estimated standalone, our network first estimates those joints such as head, body and neck. Then given rough positions of these joints, the remote joints with larger scale range are sequentially predicted.

Extensive experiments demonstrate the effectiveness of our model on MPII (Andriluka et al. 2014) and LSP (Johnson and Everingham 2010) datasets, and show that the proposed method consistently improves the performance.

Related Work

There is a large body of literature on pose estimation in the computer vision community. Models processing RGB images with CNN are most relevant to our work, in which we pay special attention to the issue of multi-scale preservation.

Human pose estimation The use of CNN has pushed the boundary of accuracy in pose estimation (Tompson et al. 2014; Pishchulin et al. 2016; Wei et al. 2016; Fan et al. 2015; Newell, Yang, and Deng 2016). DeepPose was the first practical attempt by (Toshev and Szegedy 2014) that directly regressed the coordinates via CNN. However, the structure of human body was supposed to be learned implicitly. (Yang et al. 2016) further regularized the learning process by incorporating geometric relationships among body parts into the CNN model. (Bulat and Tzimiropoulos 2016) proposed a detection-followed-by-regression CNN cascaded to learn relationships between parts and spatial context, in which heatmaps produced by the detection network could guide the regression network where to focus. To obtain long range dependencies from the output of previous stages, a multiple iterative stage is proposed, known as the Convolutional Pose Machine (CPM) (Wei et al. 2016). Furthermore, (Chu et al. 2016b; 2016a)creatively pointed out that exploring the relationships between feature maps of joints was more beneficial than modeling structures on score maps, because the former preserved substantially richer descriptions of body joints. However, most of these works focused on modeling the structural information of human body and paid little attention to the details.

Multi-scale fusion Multi-scale is a built-in advantage of CNN. In theory, as the network goes deeper, the neurons automatically obtain larger receptive field, which allows the stacked stages in CPM(Wei et al. 2016) to gradually capture long-range spatial dependencies among body parts. However, the local information is inevitably lost along the way. The stacked hourglass network (Newell, Yang, and Deng 2016) avoided this problem by adding skip layers to preserve multiple spatial information at each resolution. They also showed how repeated bottom-up, top-down processing was critical to improving the performance. Even though they started to preserve details, more attention was paid to long distance dependencies as they constructed an eight-stage network.

Besides pose estimation, merging of information across multiple scales is known to improve the performance on many other tasks that produce pixel-wise outputs, such as object detection (Gidaris and Komodakis 2015), image segmentation (Chen et al. 2016), scene parsing (Lin et al. 2016), and depth estimation (Eigen and Fergus 2015). An intuitive idea is to consolidate scales from different layers (Newell, Yang, and Deng 2016; Zagoruyko et al. 2016; Long, Shelhamer, and Darrell 2015; Xie and Tu 2015). It is often referred to as skip layer topology, where feature responses from different layers are combined in a shared output layer. Another typical way of multi-scale consolidation is having multiple scaled inputs, and then averaging all the outputs (Chen et al. 2015; Farabet et al. 2013). Obviously, it would suffer from the increasing running time and memory usage caused by the extra computation at both the training and testing stage. Another limitation it shares with skip layer is that only a fixed number of scales are selected in advance, depending on the number of skip layers or input scales. Multi-scale can also happen in the elaborate-designed computational modules, such as residual module (He et al. 2015), inception module (Szegedy, Ioffe, and Vanhoucke 2016), and Atrous Spatial Pyramid Pooling (Chen et al. 2016). Multi-scale fusion could also be performed by extracting more global features. Parsenet (Liu, Rabinovich, and Berg 2015) utilized pooling to aggregate features over the whole image. In (Eigen and Fergus 2015), fully-connected layer was considered to have contribution in providing full-image field of view, because each spatial location in the output is connected to all the image features. However, their methods still focused on more global features.

Sequential Prediction Apart from predicting body joints independently, the design of information flow between body joints is an effective way to enhance the overall performance, since it optimizes the estimation of a single joint by referring to the information from related joints. In (Chu et al. 2016b), learning structural relationship among parts at the feature level was studied. They proposed a bi-directional tree structured model to connect correlated joints, which allows every joint to receive information from all the correlated joints and update its features. Similarly, for message transmission among body joints, (Chu et al. 2016a) provided a CRF-CNN framework which can simultaneously model structural information in both output and hidden feature layers in a probabilistic way. Such a design is novel in pose estimation, because it takes feature-output, output-output, and feature-feature relationships into consideration. Besides, (Chen and Yuille 2014) explored pairwise spatial relationships between joints represented by a mixture model over types of spatial relationships. Those models are different from ours in three ways. Firstly, the relationship is built upon all joints, not just pairwise or neighbor joints. Secondly, the information flows unidirectionally, i.e., from easy-level joints (head and neck) to medium-level joints (shoulder and hip), and finally to difficult-level joints (elbow, wrist, ankle and knee). Thirdly, simply by concatenating predic-

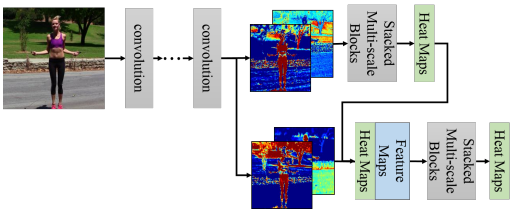

Figure 2: A two-stage pose estimation architecture constructed by the multiple iterative stage strategy. The proposed multi-scale blocks are stacked to build the pose estimation model, *i.e.* IOI, AIOI and CJN.

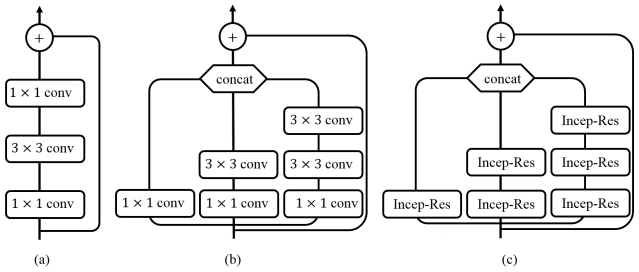

Figure 3: Blocks performed to the pose estimation model. (a) Residual block with a skip connection in each block, (b) Inception-Residual block which contains multi-scale branches, (c) the proposed Inception of Inception block that contains multiple branches of Inception block.

tion layers from the previous-level branch, we realize efficient information sharing. Whereas other models adopt relatively complex procedures, like convolution with transform kernels (Chu et al. 2016b), as well as probability computations (Chu et al. 2016a; Chen and Yuille 2014).

Method

In this paper, we construct a two-stage network following the multiple iterative stage strategy(Wei et al. 2016), as shown in Figure 2. All the output heat maps of the previous stage are forwarded into the latter stage to capture long term relationship efficiently. Moreover, on the top of each stage, we stacked several multi-scale blocks to predict heat map for each joint.

In this section, we first present the design of IOI block and then perform an attention model to each block to fuse multi-scale features adaptively. At last, the cascaded joint network structure is presented to predict the joints progressively.

Inception of Inception

For a standard chain-like neural network, the deeper network is, the higher level information it summarizes. By adding skip layer, one may build many shortcuts between feature maps (residual block). Multi-Pass strategy has been applied to various tasks to fuse multi-scale features. In this kind of strategy, different branches are constructed with different theoretical receptive fields (Szegedy et al. 2015). These strategies have been adopted in many networks (Chen et al. 2015; Bulat and Tzimiropoulos 2016), and two well-known models are residual block (He et al. 2015) and inception-residual block (Szegedy, Ioffe, and Vanhoucke 2016). In what follows, the two well-known models are briefly introduced as they will be our baseline models for comparison.

Residual block The residual block is proposed by (He et al. 2015) and has been devoted to lots of computer vision tasks. As shown in Figure 3 (a), a skip connection combines features from different scales. However, in this block, features from only two layers were mixed together. The low-level features are decreased when more residual blocks are stacked.

Inception-Residual block (Incep-Res) The inception-v1 block, proposed in (Szegedy et al. 2015), preserves more scales information by employing a multi-scale structure. (Szegedy, Ioffe, and Vanhoucke 2016) further introduced hybrid inception-residual block, which combines the inception architecture with residual connections, brings more benefits. Specifically, Figure 3 (b) shows the inception-residual block, which contains four branches, fused at the end of the block. By employing their design of inception-resNet-A module, we build our basic inception-residual block for 32×32 grid feature map, which achieved a trade-off between localization accuracy and processing efficiency. Differently, to extract features from different scales equally, we construct an inception-residual block with the same channels for each scale. The inception-residual block increases the scale diversity, but the feature variation is still limited by the difference of the possible number of convolutions in each branch.

In a multi-pathway neural network, the output contains a mixture of different scales of information extracted from the input. To calculate the proportions of different scales of information in the output, we can understand a model as an ensemble of many one-pathway networks(Veit, Wilber, and Belongie 2016). E.g. we can expand a K stacked residual block as an ensemble of 2^K networks and each network has different scale of feature(Figure 4 (a)). For each one-pathway network p, we use n_p to represent the number of convolution layers for it and k_i indicates the kernel size of the ith convolution layer. Its theoretical receptive field $R(p)$ can be simply computed by $R(p) = \sum_{i<n_p}(k_i - 1) + 1$. The larger $R(p)$, the higher feature scale that this pathway has. Summarizing all paths, we can estimate the distribution of receptive field. Formally, let P be the set of all possible paths, the proportion $P(r)$ of pathways that have a theoretical receptive filed of r could be computed by $P(r) = |\{p|R(p) = r\}| / |P|$. $|.|$ indicates the cardinality of each set. We can also compute the distribution of receptive field for stacked inception-residual and the proposed IOI blocks in the same way.

In Figure 4, we find that: (1) inception-residual block has larger scale variance than the residual block; (2) as we stack more inception blocks, the low level features are reduced(Figure 4 (c)).

We propose our IOI block that generalizes the inception structure design. Instead of mixing ones, it preserves features from different scales in different branches. As shown in Figure 3 (c), the proposed block is a nested structure. By stacking the inception-residual block following the incep-

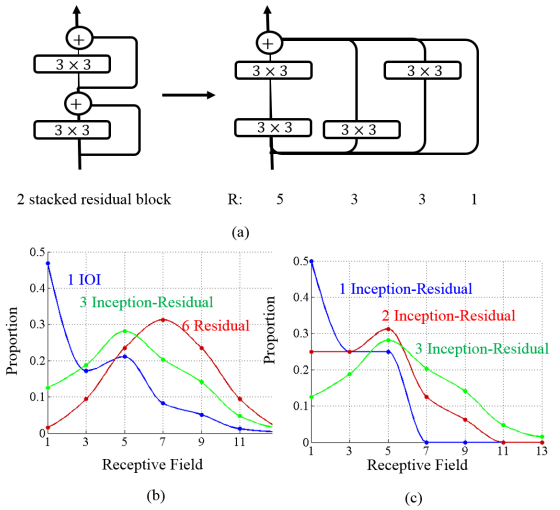

Figure 4: Scale distribution of different blocks. (a) when computing the scale distribution of 2 stacked residual blocks, we can understand it as a 4 pathway block. The 4 pathway block has the same number of possible branches with the 2 stacked residual blocks and shares same convolution layers for each possible pathway. (b) Scale proportion of three blocks, 6 stacked Residual blocks, 3 stacked Inception-Residual blocks and 1 IOI block. (c) Scale variation after Inception-Residual blocks are stacked.

tion manner, the proposed IOI block contains different number of basic units for each branch to maintain multi-scale features. By using strong inception-residual(Szegedy, Ioffe, and Vanhoucke 2016) structure as building block, the proposed module enables the IOI block achieving geometric growth of the scale distribution with multiplied branches. It also has an elongated skip connection across as far as nine convolution layers. This design also helps the proposed block to go deeper without losing detailed feature. The feature fusion is performed at the end of each block across different branches to preserve intermediate multi-scale information. Figure 4 (b) shows that the proposed IOI block presents to have more detailed features among all the blocks.

To improve the training of the networks, batch normalization (Ioffe and Szegedy 2015) is adopted right after each convolution and followed by the nonlinearity activation. But for simplicity, we just omit them in Figure 3.

Attention to Scale in Each Block

It is obvious that for predicting the location of a joint, not every level of features are equally important; rather it depends on the context, which varies in different scenarios. We employ an attention module in each IOI block to select from different scale features adaptively. The resulting block is called attention-modulated IOI (AIOI) (Figure 5).

The attention model of the AIOI block takes the multi-scale features as input and outputs the weight maps for each scale. Each weight is rescaled to the range between zero and one using sigmoid. Element-wise product is performed between feature maps and weight maps. The weighted feature

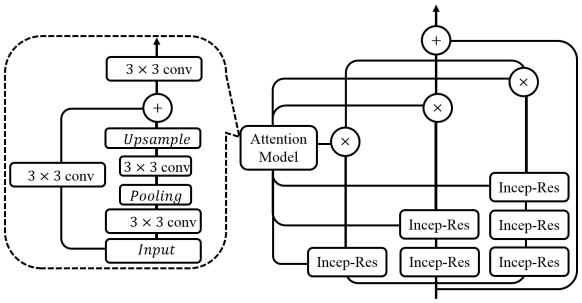

Figure 5: The proposed attention model is constructed following the Hourglass manner to generate weights with respect to high level knowledge.

maps are then summed across scales together to form the fused feature maps. Given multi-scale features $f^s_{i,c}$ where i, c, s are the index of positions, channel and scale respectively. The attention model F is denoted as $h^s_i = F(f^1_i, ..., f^S_i)$. The final fused feature map can be written as

$$\hat{f_{i,c}} = \sum_{s=1}^{S} \hat{w}^s_{i,c} \cdot f^s_{i,c} \quad (1)$$

Unlike previous work (Chen et al. 2015), which uses softmax to select features from different scales $w^s_i = \frac{exp(h^s_i)}{\sum_{t=1}^{S} exp(h^t_i)}$, we use sigmoid weight $\hat{w}^s_{i,c} = \frac{1}{1+e^{-h^s_{i,c}}}$. Since our feature fusion is performed across middle layers rather than the final layer in (Chen et al. 2015), sigmoid is a better choice to fuse multi-scale features for further processing.

As scale selection needs higher level knowledge, we design the attention model following the Hourglass manner , as shown in Figure 5. First, a 3×3 convolution is performed across the input feature maps. Then it is downsampled once with a pooling layer. After another 3×3 convolution on the smaller feature maps, its output is upsampled to the input feature map size. Similar to Hourglass, a skip branch is added to keep detail information before weighting the input feature maps with element-wise product.

Cascaded Joint Network

Multiple iterative stage strategy could capture long term relationship efficiently, but all joints are processed equally in each stage. The dependence between different joints in the same stage, which is also important is ignored. Intuitively, some joints are more easily to be predicted than others and it is reasonable to assume that inferring the easily identified joints first than difficultly identified joints would be a better strategy than inferring all joints simultaneously.

We construct a two-stage network following the multiple iterative stage strategy, as shown in Figure 2. Moreover, on the top of AIOI block, we construct a cascaded joint network to realize the aforementioned strategy. We divide joints into three groups in order: (1) Head and neck are more likely to be visible, contrast to the other joints which are always partially occluded. So in our network, we first localize head

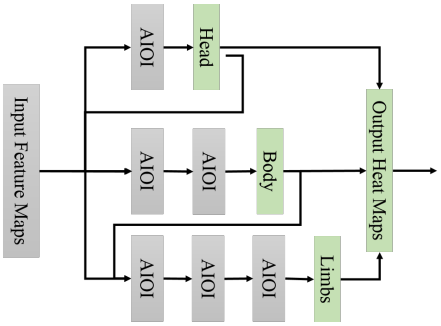

Figure 6: A two-stage pose estimation architecture constructed by the multiple iterative stage strategy. The proposed blocks are stacked to build the pose estimation model.

and neck in the first branch. (2) As the shoulders and hips are relatively rigid than the joints on the limbs, we predict joints near body in the second branch (3) The joints on the limbs are more difficult to localize than the others due to occlusion and nonrigid deformation. The third branch finally inferences the ankles and wrists. When we compute the joints' heatmaps for one group, we concatenate it with the feature map, and go through several AIOI blocks, to predict joints' heatmaps of the next group. The number of the stacked AIOI blocks is different for each branch to construct a memory efficient cascaded joint network. As head and neck are easy to localize, we perform only one AIOI block in the first branch. While predicting shoulders and hips, the second branch needs a lager receptive field to capture the results of head and neck. The input of the cascaded joint network is 32×32, so two AIOI blocks (with a theoretical receptive field of 23×23) are stacked in the second branch. To capture all the context information of the former branches, the third branch has three AIOI blocks with a theoretical receptive field 35×35 which is larger than the input size.

The joint estimation benefits from the proposed sequential prediction in two aspects: (1) Heapmaps of joints can be treated as a rich semantic feature. (2) Given the fact that the former groups are more easily to be detected than the latter ones. Their heatmap are more robust and constrain the output of latter groups. By reducing supervision information in intermediate layers, each branch of our model concentrates on different joints. This can be justified by an ablation study (Table 5).

Experiment

Datasets and Evaluation Criterion

We analyze the proposed models on the MPII validation set and present the comparison of our algorithm with the state-of-the-art methods on both the standard MPII (Andriluka et al. 2014) and 'Leeds Sports Poses' (LSP) (Johnson and Everingham 2010) benchmarks.

LSP dataset. The Extended Leeds Sports Dataset (Johnson and Everingham 2011) extends the original LSP dataset and ends up with 11000 images for training and the same 1000 images for testing. The Percentage Correct Keypoints (PCK) metric (Yang and Ramanan 2013) is used on LSP where the models are evaluated according to the torso size.

Table 1: PCKh@0.5-based comparison of multi-scale blocks on MPII validation set.

Models	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Res	97.3	94.6	87.3	81.9	87.0	81.9	77.4	87.3
Incep-Res	97.5	95.2	88.5	83.2	88.2	83.4	79.3	88.4
IOI	97.5	95.4	89.1	83.9	88.1	84.4	80.2	88.8

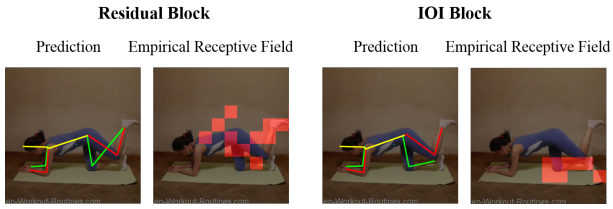

Figure 7: Visualization examples of receptive field perceived by the neurons that have high responses to the right ankle on the MPII validation set. The results show that the IOI block is able to capture details around the related joints.

MPII dateset. More than 28000 training images of the MPII dataset are used to train the proposed model. The dataset comprises of 10000 annotated persons and 14 keypoints for each person are labeled for evaluation. The results are measured by PCKh score (Andriluka et al. 2014), where the error tolerance is normalized with respect to half of the head size for each person. The proposed models are analyzed on a subset of MPII training set containing 25925 images, and evaluated on the validation set of 2958 images.

Comparison of Multiscale Blocks

We first compare the results of the IOI block with those of the residual block and inception-residual block. To obtain a fair comparison, we only replace 'multi-scale blocks module' with stacked baselines or proposed blocks in Figure 2, and the depth of blocks stacked in each architecture is designed to keep the same theoretical receptive field. In the experiment, 18 residual blocks, 9 inception-residual blocks and 3 IOI blocks are carried out.

Table 1 reports results for each architecture. We observed: (1) IOI outperforms all other architectures and achieves 1.5% overall accuracy higher than the baseline. (2) For certain joints such as ankle and wrist, IOI block exceeds the residual block by 2.8% and 2.0%. The difficulty of predicting them comes from the large degree of freedom and the occlusion of limbs. And our method shows good capability of handling this.

To better understand what details the IOI blocks capture, we visualize the empirical receptive field (Zhou et al. 2014; Liu, Rabinovich, and Berg 2015) for the output neurons that have high responses to ankle. Figure 7 shows the comparison of the empirical receptive fields between IOI block and baseline residual block: (1) When predicting the right ankle, the residual block has a larger receptive field (the red region of the image) that covers almost half of the entire body. In this case, though large receptive field has access to global

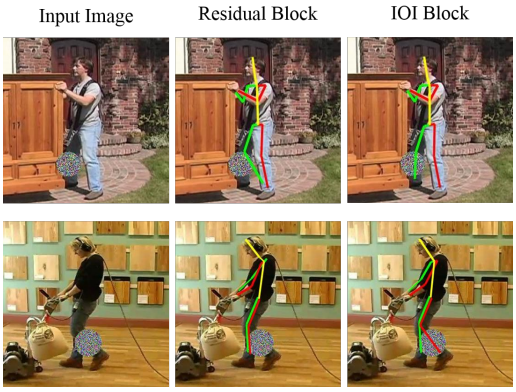

Figure 8: Comparison on the artificial limb-occlusion dataset. When one side of the limb is occluded by the circular noise, residual block tends to confuse the occluded part with the exposed counterpart since the structure information they refer to is lost. Our IOI block still recognizes the right position by focusing on the critical details, such as the exposed foot connected with the occluded calf.

Table 2: Comparison of different nonlinear functions for the attention model.

Models	Wrist	Ankle	UBody	Mean
AIOI-softmax	82.7	79.0	88.5	87.9
AIOI-sigmoid	83.2	80.0	88.7	88.4

context, some incomplete but critical clues could have been overwhelmed by a massive dump of structural information, which would have a negative influence on the outputs. (2) We also observe that the receptive field of the IOI block is relatively small, but keeps pinpointing to the joint. Although the right ankle is partially occluded, our model still precisely locates it with the help from the exposed right leg as a more effective receptive field.

Based on the above evaluation, the IOI block turns out to be the optimal among all the possible combinations. We hereafter restrict our attention to the comparison of only the IOI with other state-of-the-art models.

Effect of Attention on IOI

We evaluate the attention model on the MPII validation set and first compare two kinds of activations: (1) A softmax function applied in (Chen et al. 2015) to generate weights. (2) A model follows the same structure but replaces the softmax function with the sigmoid function. In Table 2, we observed that the attention model with the sigmoid activation performed better than softmax as it preserves more features in the feature fusion for each block.

In Table 3, the model with an attention structure introduced in Figure 5 achieves better performance and improves

Table 3: PCKh@0.5-based comparison of the proposed attention model on MPII validation set.

Models	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
IOI	97.5	95.4	89.1	83.9	88.1	84.4	80.2	88.8
AIOI	97.6	95.5	89.2	84.3	88.6	84.4	80.9	89.1

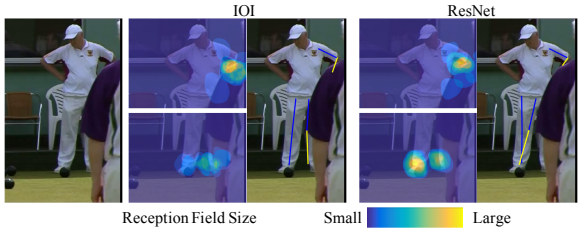

Figure 9: Empirical receptive fields for two joints: wrist and ankle. The color indicates the size of the receptive field (RF) for each pixel, the lighter the larger. Different joints need various features. As for the case of complete occlusion like left wrist, both the baseline and IOI acquires larger RF to judge the position using global information. However, for the partially occluded left ankle, IOI dynamically captures the unobstructed details and has a smaller RF around the ankle. While the baseline method insists larger RF and leads to erroneous prediction using global information.

Table 4: Comparisons of PCKh@0.5 scores on the augmented MPII validation dataset. The limbs in the augmented dataset are randomly occluded.

Models	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Res	97.2	94.7	87.0	81.6	86.3	76.1	64.7	84.9
Incep-Res	97.4	95.1	88.3	83.2	87.3	78.8	67.8	86.3
IOI	97.6	95.2	88.6	83.2	87.2	79.0	68.7	86.5
AIOI	97.6	95.4	88.9	84.1	87.7	79.9	71.3	87.2

the performance of the IOI block by 0.3%.

Robustness of the AIOI Block

We are also interested in the robustness to bring by the attention module. In Figure 9, we visualize the empirical receptive fields of both residual block and AIOI based models. As the previous network focuses on high-level features, the partially visible left leg is ignored. On the other hand, the AIOI block has an attention model to modulate features adaptively, so it predicts the left wrist with larger receptive field and detects the left ankle correctly with smaller receptive field.

To quantitatively gauge the model's robustness to partial occlusion, we augment MPII's validation set by adding random occlusions on the limbs as test data. Under this setting, calves are randomly covered by a circular patch of random noise. The diameter of each circle is set to 80% of the calf length in our experiment, so as to make the circular patch large enough to extensively occlude the structure without hiding the ankle joint. We show the occluded samples in Figure 8. Notice that the augmented data are not used for training.

Table 4 shows that: (1) Compared with Table 1, performances for all algorithms at ankle point dropped around 10%, which demonstrates the challenge of unseen occlusion. (2) The AIOI blocks dropped 9.6%, which is significantly smaller than the residual block 12.7%, inception-residual block 11.5% and IOI block 11.5%. This gap is even larger than that on the standard MPII validation set. The result shows that the AIOI block is more robust to the loss of

Table 5: The proposed cascaded joint network is evaluated on the MPII validation set and compared with three models: (1) MS-AIOI model, (2)CJN-Random: cascaded joint network with random selected joint for each branch, (3)CJN-all:cascaded joint network that predicts all joints in each branch.

Models	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
MS-AIOI	97.6	95.0	89.5	84.6	88.3	84.2	81.7	89.1
CJN-Random	97.7	95.0	88.7	83.6	88.2	84.4	81.1	88.8
CJN-all	97.7	95.6	89.4	84.4	89.0	85.2	82.0	89.4
CJN	97.6	95.3	89.5	85.1	88.7	85.4	83.1	89.6

structure.

As for the visual performances exhibited in Figure 8, the proposed AIOI-based model localizes keypoints accurately even when the structure information is missing, which suggests the block's capability in detail preservation through multiplied branches and longer skip connections.

Effect of the Cascaded Prediction

The cascaded strategy and deeply supervised method were widely employed by previous works. Here we compare with several baselines to evaluate the advantage of the proposed order of each group. (1) MS-AIOI, in which cascade we feed forward the former heatmaps as an additional feature. Here we remove this feed path and each branch directly outputs one group of joints, as a baseline of multiple scale AIOI. (2) We also randomly divide joints into three ranked groups. The model is indicated by CJN-Random. (3)CJN-all: In this baseline we implement a structure similar to Convolutional Pose Machine (Wei et al. 2016). In each step CJN-all predicts all joints simultaneously instead of sequentially.

In Table 5, we observed that CJN-all contains completely information and gets better result than MS-AIOI does, since the MS-AIOI has no message passing between each branch and the harder joints could not benefit from the distinct ones. In addition, the CJN outperformed all the other models. The carefully designed structure allows it to further improve the performance of AIOI(compared with Table 3) by 0.5 %.

Comparison with the State-of-the-Art

We compare our algorithm with the state-of-the-art methods on the MPII and LSP benchmarks. Seeing the benefits of adding MPII data to boost the model performance in the previous practice (Wei et al. 2016), we train the model on the united training sets of the extended LSP and MPII. The result in Table 6 indicates that the proposed model surpasses the state of the art on this well studied dataset with a PCK as high as 93.6% and outperforms all the previous methods for all joints, especially for wrists and ankles.

The results on MPII dataset are summarized in Table 7. The proposed method achieves the state-of-the-art performance. The method with the closest performance is proposed by (Chu et al. 2017), which uses a different strategy to fuse multi-scale features. It does not perform as well as our method, and one potential reason is that it does not preserve more detailed features. The AUC of our method exceeds by 0.8%, which indicates more accurate keypoints localization.

Table 6: PCK@0.2-based person-centric comparison with the state-of-the-art methods on LSP test set

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Rafi *et al.*	95.8	86.2	79.3	75.0	86.6	83.8	79.8	83.8
Yu *et al.*	87.2	88.2	82.4	76.3	91.4	85.8	78.7	84.3
Belagiannis *et al.*	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2
Lifshitz *et al.*	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Pishchulin *et al.*	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Insafutdinov *et al.*	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Wei *et al.*	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bulat *et al.*	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Chu *et al.*	**98.1**	93.7	89.3	86.9	**93.4**	94.0	92.5	92.6
AIOI+CJN	**98.1**	**94.0**	**91.0**	**89.0**	**93.4**	**95.2**	**94.4**	**93.6**

Table 7: PCKh@0.5-based comparison with the state-of-the-art methods on MPII test set

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean	AUC
Tompson	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0	54.9
Hu	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4	51.1
Pishchulin	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4	56.5
Lifshitz	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0	56.8
Gkioxary	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1	57.3
Rafi	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3	57.3
Insafutdinov	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5	60.8
Wei	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5	61.4
Bulat	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7	59.6
Newell	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9	62.9
Chu	**98.5**	96.3	91.9	**88.1**	90.6	88.0	85.0	91.5	63.8
AIOI+CJN	98.4	**96.4**	**92.0**	87.8	**90.7**	**88.3**	**85.3**	**91.6**	**64.6**

Conclusion

In this paper, we tackle human pose estimation with a two-stage architecture built on the novel multi-scale Inception of Inception blocks. In order to preserve more features from different scales in the block, the IOI block is developed in a nested inception-residual structure, which makes IOI wider and deeper compared with other multi-scale blocks. Owing to that, architecture constructed by the IOI block turns out to be the most effective in localizing keypoints. In order to adjust the scale distribution for each sample, an attention model is designed for each block to fuse multiscale feature adaptively. As the proposed block preserves diverse features, it is more robust to structural information loss caused by occlusion. Finally, to enforce dependency between different joints, we proposed a multiscale cascaded joint network to process different joints in different scales and different order and further improves the result. Overall, our model achieves the state-of-the-art performances on both LSP and MPII benchmarks.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 3686–3693. IEEE.

Bulat, A., and Tzimiropoulos, G. 2016. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 717–732. Springer.

Cao, X.; Wei, Y.; Wen, F.; and Sun, J. 2014. Face alignment by explicit shape regression. *IJCV* 107(2):177–190.

Chen, X., and Yuille, A. L. 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 1736–1744.

Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; and Yuille, A. L. 2015. Attention to scale: Scale-aware semantic image segmentation. *arXiv preprint arXiv:1511.03339*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*.

Chu, X.; Ouyang, W.; Li, H.; and Wang, X. 2016a. CRF-CNN: Modelling structured information in human pose estimations. In *NIPS*.

Chu, X.; Ouyang, W.; Li, H.; and Wang, X. 2016b. Structured feature learning for pose estimation. *arXiv preprint arXiv:1603.09065*.

Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A. L.; and Wang, X. 2017. Multi-context attention for human pose estimation. In *CVPR*.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 1110–1118.

Eigen, D., and Fergus, R. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2650–2658.

Fan, X.; Zheng, K.; Lin, Y.; and Wang, S. 2015. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *CVPR*, 1347–1355. IEEE.

Farabet, C.; Couprie, C.; Najman, L.; and LeCun, Y. 2013. Learning hierarchical features for scene labeling. *PAMI* 35(8):1915–1929.

Gidaris, S., and Komodakis, N. 2015. Object detection via a multi-region and semantic segmentation-aware cnn model. In *ICCV*, 1134–1142.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR* abs/1502.03167.

Johnson, S., and Everingham, M. 2010. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*.

Johnson, S., and Everingham, M. 2011. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 1465–1472. IEEE.

Lin, L.; Wang, G.; Zhang, R.; Zhang, R.; Liang, X.; and Zuo, W. 2016. Deep structured scene parsing by learning with image descriptions. *arXiv preprint arXiv:1604.02271*.

Liu, W.; Rabinovich, A.; and Berg, A. C. 2015. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. *arXiv preprint arXiv:1603.06937*.

Pishchulin, L.; Insafutdinov, E.; Tang, S.; Andres, B.; Andriluka, M.; Gehler, P.; and Schiele, B. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*.

Shotton, J.; Sharp, T.; Kipman, A.; Fitzgibbon, A.; Finocchio, M.; Blake, A.; Cook, M.; and Moore, R. 2013. Real-time human pose recognition in parts from single depth images. *Communications of the ACM* 56(1):116–124.

Sun, X.; Wei, Y.; Liang, S.; Tang, X.; and Sun, J. 2015. Cascaded hand pose regression. In *CVPR*, 824–832.

Szegedy, C.; Liu, W.; Jia, Y.; and Sermanet, P. 2015. Going deeper with convolutions. In *CVPR*, 1–9.

Szegedy, C.; Ioffe, S.; and Vanhoucke, V. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*.

Tompson, J. J.; Jain, A.; LeCun, Y.; and Bregler, C. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 1799–1807.

Toshev, A., and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 1653–1660.

Veit, A.; Wilber, M. J.; and Belongie, S. 2016. Residual networks behave like ensembles of relatively shallow networks. In *NIPS*, 550–558.

Wang, C.; Wang, Y.; and Yuille, A. L. 2013. An approach to pose-based action recognition. In *CVPR*, 915–922.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *CVPR*.

Xie, S., and Tu, Z. 2015. Holistically-nested edge detection. In *ICCV*, 1395–1403.

Yang, Y., and Ramanan, D. 2013. Articulated human detection with flexible mixtures of parts. *PAMI* 35(12):2878–2890.

Yang, W.; Ouyang, W.; Li, H.; and Wang, X. 2016. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*.

Zagoruyko, S.; Lerer, A.; Lin, T.-Y.; Pinheiro, P. O.; Gross, S.; Chintala, S.; and Dollár, P. 2016. A multipath network for object detection. *arXiv preprint arXiv:1604.02135*.

Zhang, D., and Shah, M. 2015. Human pose estimation in videos. In *ICCV*, 2012–2020.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2014. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*.