Applied Intelligence
https://doi.org/10.1007/s10489-021-02383-6

CFENet: Content-aware feature enhancement network for multi-person pose estimation

Xixia Xu[1] · **Qi Zou**[1] · **Xue Lin**[1]

Accepted: 24 March 2021

Abstract
Multi-person pose estimation is a fundamental yet challenging task in computer vision. Although great success has been made in this field due to the rapid development of deep learning, complex situations (e.g., extreme poses, occlusions, overlapped persons, and crowded scenes) are still not well solved. To further mitigate these issues, we propose a novel Content-aware Feature Enhancement Network (CFENet), which consists of three effective modules: Feature Aggregation and Selection Module (FASM), Feature Fusion Module (FFM) and Dense Upsampling Convolution (DUC) module. The FASM includes Feature Aggregation Module (FAM) and Information Selection Module (ISM). The FAM constructs the hierarchical multi-scale feature aggregations in a granular level to capture more accurate fine-grained representations. The ISM makes the aggregated representations more distinguished, which adaptively highlights the discriminative human part representations both in the spatial location and channel context. Then, we perform FFM which effectively fuses high-resolution spatial features and low-resolution semantic features to obtain more reliable context information for well-estimated joints. Finally, we adopt DUC module to generate more precise prediction, which can recover missing joint details that are usually unavailable in common upsampling process. Comprehensive experiments demonstrate that the proposed approach outperforms most of the popular methods and achieves a competitive performance with the state-of-the-art methods over three benchmark datasets: the recent big dataset CrowdPose, the COCO keypoint detection dataset and the MPII Human Pose dataset. Our code will be released upon acceptance.

Keywords Multi-person pose estimation · Feature aggregation · Information selection · Feature fusion · Dense upsampling convolution

1 Introduction

Multi-person Pose Estimation devotes to locate body parts for multiple persons in an image, such as keypoints on the arms, torsos, and the face [10, 31, 56]. The related tasks contain pose estimation in videos [25] and 3D pose estimation [62].

✉ Qi Zou
qzou@bjtu.edu.cn

Xixia Xu
19112036@bjtu.edu.cn

Xue Lin
18112028@bjtu.edu.cn

[1] Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China

It's regarded as a fundamental task to deal with other related high-level tasks, such as human action recognition [45], human-computer interaction [30], motion capture [60] and emotion analysis [3].

Recently, due to the rapid development of convolutional neural networks (CNN) [14], most existing methods [7, 11, 18, 21, 46, 55, 57] have achieved remarkable advances in multi-person pose estimation. They can be roughly classified into two categories, i.e., the bottom-up [4, 17, 33] and top-down [7, 11, 21, 46]. The former first detects all body keypoints together and then assigns the detected joints to different human instances. In contrast, the latter one detects all human bounding boxes in one image at first and then estimates the pose based on the detected human proposals. Besides, other methods do exist, e.g., [35] can solve the pose estimation by turning the bottom-up method into an end-to-end single framework. With the prior information of human detection, top-down methods can get

Published online: 26 April 2021

more reliable and accurate estimations. Our model belongs to the top-down framework.

However, despite that current methods have achieved promising results on standard benchmarks such as MPII Human Pose [1] and COCO keypoint detection dataset [24], some challenges have not been well addressed. One of them is the missing detection of joints due to cluttered background like Fig. 1a, humans in multiple scales, or due to occlusions like Fig. 1e. Another is the false locating of joints due to overlapping in crowded scenes like Fig. 1c.

The main reasons lie in several points: **1)** multi-scale features are not fully exploited. Most existing methods only represent the multi-scale features in a coarse-grained layer-level, but neglect the fine-grained multi-scale information within each layer. Therefore, existing multi-scale techniques either are less robust to scale variations or have high computational costs. **2)** Most CNN-based methods treat the contributions of different joint locations and multiple channels equally, which results in lacking of discriminative spatial and semantic information across channels for pose estimation. **3)** Context information is not well represented in many existing methods, which results in failing to infer the invisible or indistinguishable joints, such as the left knee and ankle of the man on the motorbike in (Fig. 1e). **4)** Missing detailed informations in the downsampling process are not effectively recovered in most methods, which makes models tend to get the false or missing joints in the final output. Besides, the computational efficiency is also a problem. Some existing methods cost a large number of computing resources and design complex networks to address the above problems.

To handle the above issues effectively and efficiently, we propose a Content-aware Feature Enhancement Network (CFENet), which is designed on the ResNet [14] backbone, the whole framework consists of two parts: encoder and decoder network. In the encoder, a **Feature Aggregation**

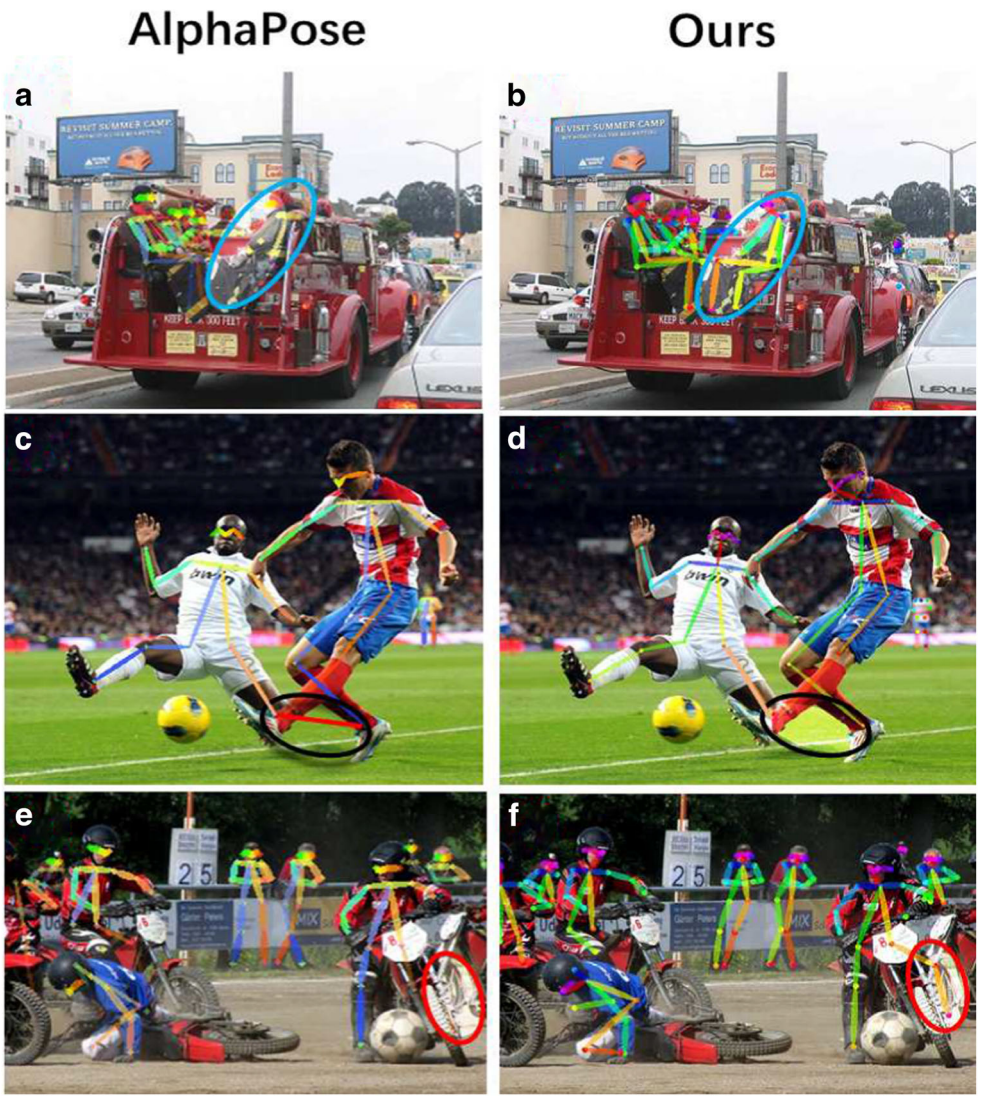

Fig. 1 Qualitative comparison of AlphaPose [11] and our method in complex scenes. The major difference between the two methods is marked with a circle

and Selection Module (FASM) is added in the bottleneck to enhance the multi-scale feature extractions and adaptively select the most discriminative human part regions. Before the decoding stage, we combine the low-level spatial information with the high-level semantic concepts in the **Feature Fusion Module (FFM)**. Based on the fused features, we attach a **Dense Upsampling Convolution (DUC)** [53] module in the decoder to explicitly recover the detailed keypoint information.

Our method has several benefits in comparison to existing widely-used methods for pose estimation. On one hand, **Feature Aggregation Module (FAM)** extracts features in a group-wise manner rather than in layer-wise manner. Therefore it can aggregate stronger features using a smaller number of parameters. On the other hand, **Information Selection Module (ISM)** learns the enhanced discriminative human part representations, and accordingly the predicted heatmaps are spatially more precise. Additionally, most existing fusion schemes aggregate low-level and high-level representations via the dense connections and cost much more computing resources. Instead, we combine the low-level structural features and high-level semantic information in an efficient way to capture rich context information adequately. Consequently, we can better infer the invisible or occluded joints with this context information and reduce the wrong joint regressions. Inspired by [53], we adopt DUC module on head of network rather than the common bilinear upsampling, which can decode more detailed information which are missing in the downsampling process. It can naturally fit the overall framework in an end-to-end manner, and increase the AP on the COCO dataset [24] significantly, especially on relatively small persons.

In summary, our main contributions are three-fold as follows.

- For better capturing the multi-scale fine-grained informations for body joints, Feature Aggregation and Selection Module (FASM) is proposed. Differing from most existing methods extracting aggregated coarse multi-level features, FASM can exploit more fine-grained information of body parts in a lightweight way, and adaptively select discriminative contents for joint locations.
- The Feature Fusion (FF) strategy in our method is leveraged to let spatial high-resolution information and semantic information interact and benefit each other, which has an outstanding advantage on inferring precise parts.
- Our method achieves a competitive performance and efficiency on the COCO keypoint dataset [24] and MPII human pose benchmark [1] and achieves comparable results on the complex CrowdPose dataset [21] without using any extra data.

2 Related work

2.1 Multi-person pose estimation

Recently, multi-person pose estimation attracts more and more attention due to the real-life demand. There are two kinds of methods for solving multi-person pose estimation. One is bottom-up approaches, which obtain all the keypoints of the humans in the input image and assemble the detected keypoints into different persons. The other is top-down approaches, which firstly adopt human detector to get bounding boxes and then feed the cropped bounding box to the pose estimation network.

Bottom-up Bottom-up approaches [4, 17, 33, 41] firstly detect the human body keypoints and then group them with clustering algorithms to different human instances. Compared with the top-down approaches, they are faster in testing and lighter in building model. However, the bottom-up method loses the chance to amplify the details of each person, this results in its accuracy lower than the top-down methods. DeepCut [41] regards the keypoints localization as an Integer Linear Program (ILP) problem and assigns keypoints into different persons, and combines the person clusters with corresponding body parts to obtain the final results. DeeperCut [17] improves DeepCut [41] using ResNet [14] to extract stronger body parts representations and adopts image-conditioned pairwise terms to achieve better performance. Associative Embedding [33]produces confidence maps and groups the keypoint candidates into different instances to generate the final result. MultiPoseNet [20] jointly regresses the human detections and keypoint detections, which designs a pose residual network (PRN) to receive the detected results to obtain estimation results by assigning keypoints to different persons.

Top-Down Top-down approaches [7, 11, 38, 55] regard the process of predicting keypoints as a two-step operation. Actually, top-down methods firstly detect all persons in an image and crop the person regions, then feed the regions into the single person pose estimation model. G-RMI [38] predicts heatmaps and offsets of the points on the heatmaps to the labeled heatmaps, and then gets the final locations. Moreover, the CPN [7] divides keypoints into the hard and easy level. It uses the feature pyramid architecture [23] as backbone and divides the whole framework into the GlobalNet and RefineNet, to locate the easy and hard keypoints respectively. HRNet [48] proposes a state-of-the-art method and learns information from the same image with different scales to make the predictions more accurate. Our work follows a two-step framework [55] which based on a simple encoder-decoder architecture and adopts a few transposed convolution layers for generating high-resolution

representations. The top-down framework can achieve superior performance as development of object detection and single person pose estimation.

2.2 Multi-scale feature representations and fusions

From the traditional feature design [28] to deep learning [15, 50], the multi-scale features have been widely used in many research fields. A large receptive field is necessary for obtaining multi-scale feature representations to locate joints at different scales precisely. In the previous multi-person pose estimation methods, the large receptive field is achieved by the architecture in CPM [54] to implicitly capture the spatial relations among different parts, resulting in the progressively refined estimations. However, low-level information is not considered well. SHG [34] process the feature maps of all different scales to capture various spatial relationships of different resolutions, and adopt the skip connections to save information at each resolution. But this model works at a cost of much more computing resources and may not dig out the information exhaustively. By the above observations and inspired by the recent [12], we propose **FAM** to aggregate the features at a granular and lightweight level in the bottleneck and amplify the receptive fields for each layer.

The multi-level feature fusion is crucial for the keypoint localization as well. For example, SHG [34] fuse the different resolution feature maps in a dense way and achieve a good performance. The CPN [7] preserves the high-level and low-level information from the feature maps across different scales to obtain more information. The MSNP [22] improves the effect of the network via introducing feature fusion between adjacent stages. However, the above methods fuse the different level features with connections in complex forms and concepts. And they're still hard to make a better result. Thus, we propose a simple and effective **FFM** to combine the high-level semantic information with low-level features to enhance the global context information.

2.3 Attention mechanism

Visual attention has achieved great success in various tasks, such as image classification, object detection and semantic segmentation. SENet [16] proposes a "Squeeze-and-Excitation" block to leverage channel-wise feature by exploring channel attention. In [46], they design a Spatial and Channel-wise Attention Residual Bottleneck to strengthen the feature responses in the spatial and channel-wise context. The above methods only focus on the spatial and channel attention considering the single-scale feature representation extraction. However, few attention has been paid to adaptively distinguish the importance of the multi-scale aggregated keypoint representations. Our proposed **ISM** is inspired by the existing attention studies and we adaptively obtain the discriminative human parts representations based on the learned aggregated features.

2.4 Decoding of feature representation

The downsampling operations in CNN architectures lead to the essential fine details (low-level information) missing, which cannot be well recovered by the common up-sampling operations. There are many methods propose to decode detailed information which are generally lost in the downsampling process. Bilinear upsampling operation is commonly used [5, 59], due to its fast speed and efficient memory. Another common used operation is deconvolution, in which the unpooling operation adopts the stored pooling shifted from the pooling to recover the essential information for visualization [59]. As FCN [27] illustrates, in the decoding stage places the deconvolutional layer to generate results with the help of cascaded intermediate feature maps. DeconNet [37] adopts deconvolutional layers in unpooling operation by using the preserved pooled location. By the above observation, we find that the existing estimation methods pay little attention to the inevitable information missing during the encoding stage for the keypoint localization. Except that, simple baseline [55] method adopts a few transposed convolution layers for generating high-resolution representations, which are vital to recover more fine detailed information than the bilinear interpolation. However, we think it still can result in some detailed information missing, so we apply the **DUC** [53] module to replace the deconvolutional layers to make up for the missing details and obtain the final prediction. Different from the recent HRNet [48], it maintains the same high-resolution horizontally thus it doesn't need the upsampling operation let alone the information loss resulting from it. The DUC module is mainly designed for the most existing pose estimation methods that adopt the up-sampling operations.

3 Method

In this section, we creatively propose a CFENet to make the localization more precise. It can not only make the best use of the context information but extract the abundant discriminative multi-scale features. In addition, the model can recover more detailed information as much as possible that is mostly missing in the downsampling process. An overview of our proposed framework is illustrated in Fig. 2. Firstly, we briefly review the Simple Baseline Network (SBN) [55] framework. After that, we introduce the FASM, FFM and DUC module in detail.

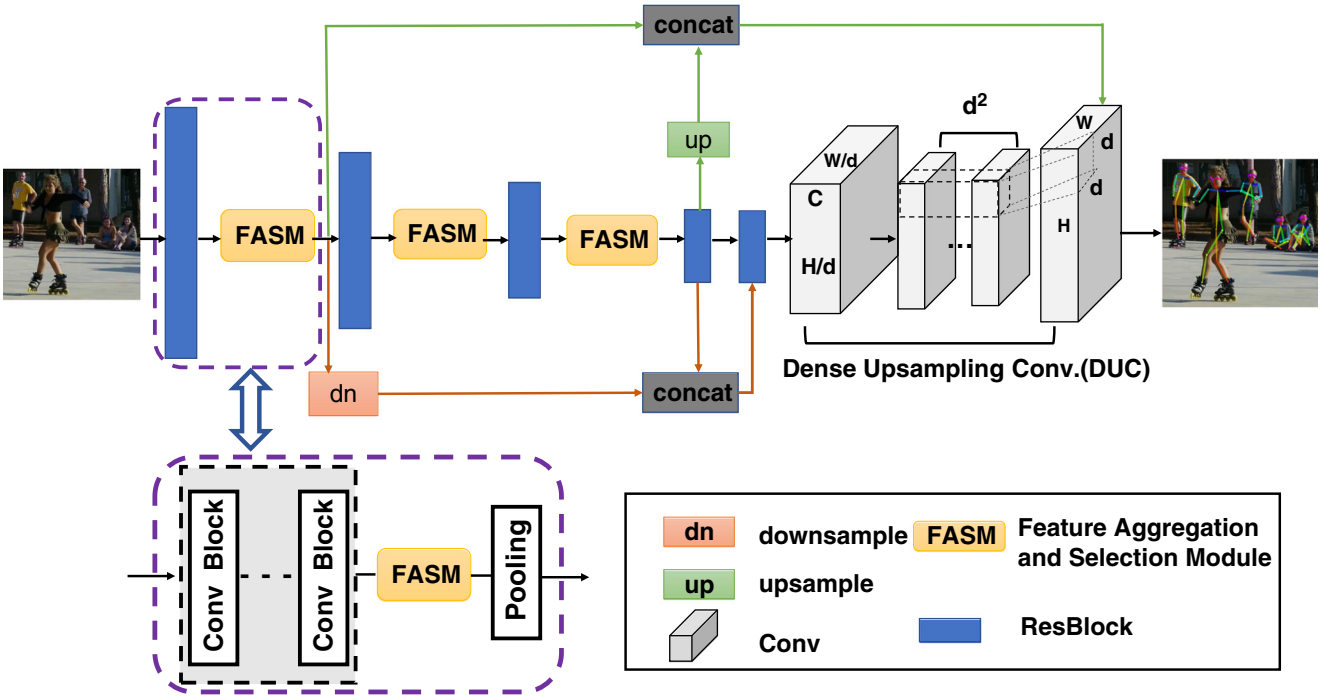

Fig. 2 Overview of the proposed CFENet for multi-person pose estimation. The proposed FASM is placed at every bottleneck of the encoder and regarded as input features extracted from the convBlock of ResBlock [14]. It is depicted in a purple dashed line box. The feature map from first resblock concatenates with feature map from the last resblock (the green line and the pink line represent the concatenated flow). The DUC [53] module is placed at the head of the network to decode the output feature maps to the stable resolutions which recovers rich and detailed contextual information

3.1 Review simple baseline network

Simple Baseline Network (SBN) [55] adopts three deconvolution layers after the last convolution stage of the ResNet [14], in which each deconvolution layer has 256 filters with 4×4 kernel size and stride is 2. Finally, a 1×1 convolution layer is added to the deconvolution module to predict the keypoint heatmaps. Build upon its design, we add our proposed components that refer to the FASM, FFM, DUC in the Fig. 2 to form our framework.

3.2 FASM: Feature Aggregation and Selection Module

In order to extract informational and discriminative feature of human parts, we design a Feature Aggregation and Selection Module in each bottleneck block as shown in Fig. 3. FASM consists of FAM and ISM. The FAM extracts more multi-scale features to enhance the feature representation ability of the human parts. ISM is followed by the FAM, which is mainly used for adaptively highlighting the discriminative joint features among different feature maps and locations. The detailed explanations are discussed as follows.

3.2.1 Feature aggregation module

Since human pose estimation is a typical structure prediction problem, it is essential to explore both local characteristics (such as some area around the torsos) and global context information. It is benefit to model the spatial relationship between joints. Accordingly, a larger receptive field of network is preferable. To realize this goal, FAM is proposed to extract robust features for multi-scale stimuli for the network. Superior to most existing methods that only consider the coarse multi-scale information across different layers of CNN framework, we also capture fine-grained multi-scale information within each layer in a lightweight way.

In each resblock, bottleneck structure [14] is applied, which is a popular building block in many modern backbone CNNs architectures, as shown in the Fig. 3(a). Our proposed multi-scale FAM implements based on it. As shown in the purple rectangle enclosed by dashed line at the top of the Fig. 3(b). After the 1×1 convolution, we evenly split the feature maps into 4 feature map subsets, denoted by m_i, where $i \in 1, 2, 3, 4$. Compared to the input feature maps, each sub-map m_i has the same spatial size but the 1/4 number of channels. Besides m_1, each m_i is plugged into

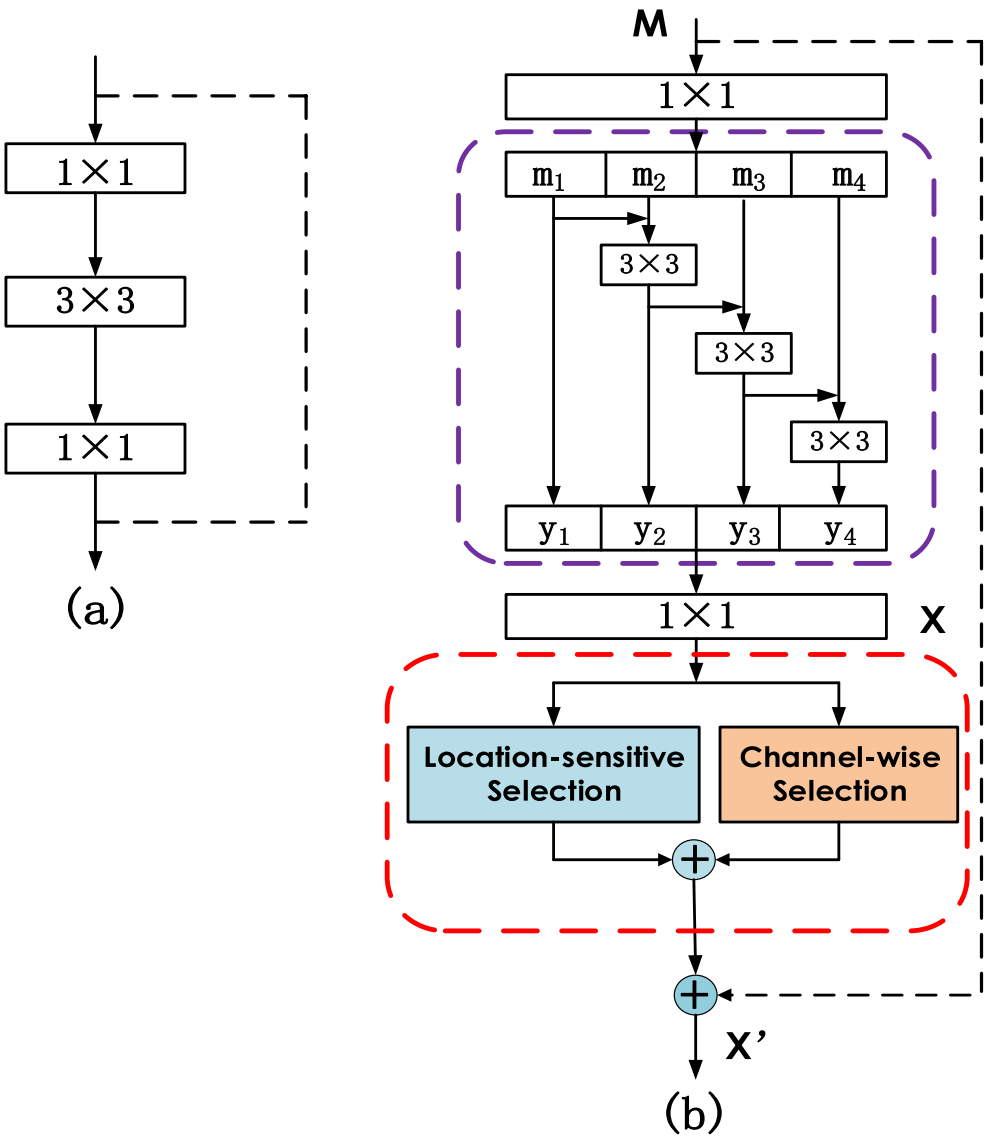

Fig. 3 The schema of the (**a**) Bottleneck and (**b**) FASM which is composed of the Feature Aggregation Module (the **purple** dashed rectangle) and Information Selection Module (the **red** dashed rectangle)

a corresponding 3×3 convolution $\mathbf{G}_i(\cdot)$, whose output is denoted as $y_i = \mathbf{G}_i(m_i)$. To characterize the fine-grained multi-scale information within each layer, we connect the output of each sub-map in a residual-like manner. The feature subset m_i is added with the output of $\mathbf{G}_{i-1}(\cdot)$, and then fed into $\mathbf{G}_i$. The form of the organization of features can improve the execution efficiency because we split features in a different way. If we want to reduce parameters, the subset numbers will increase. Thus, we don't apply the 3×3 convolution for m_1. Thus, y_i can be represented as:

$$y_i = \begin{cases} m_i, & i = 1, \\ G_i(m_i + y_{i-1}), & 1 < i \leq s. \end{cases} \tag{1}$$

Intuitively, benefit from the connections between sub-maps, each 3×3 convolutional operator $\mathbf{G}_i(\cdot)$ can potentially receive feature information from all feature splits m_j, $j \leq i$. Furthermore, the obtained feature y_i will correspond to a larger receptive field than the feature extracted by standard bottleneck structure without FAM. In summary, FAM is effective to obtain multi-scale features of detected persons, which takes full advantage of local information of joints and global relationship of different parts among human. In order to fully fuse information across different scales, we concatenate all branches and feed them into a 1×1 convolution, which can extract more robust features effectively. This module can reduce the model parameters and also improve the localization accuracy.

3.2.2 Information selection module

For the multi-person pose estimation, the extracted human features about the locations and semantics are possibly redundant. It's unreasonable to treat the contributions of different locations and channels equally. For the purpose of highlighting the discriminative information both in the spatial location and channel context, we propose ISM, which aims to learn location-sensitive weight β and channel-wise weight α for each feature map respectively. Inspired by the researches [16, 39], we design our ISM as shown in Fig. 3, the bottom half of the figure enclosed by the red rectangle dashed line.

location-sensitive selection Applying the whole feature maps may lead to sub-optimal results due to the irrelevant regions. Different from paying attention to the whole image region equally, **Location-Sensitive Selection (LSS)** mechanism attempts to adaptively highlight the part-related regions in the feature maps.

In specific, assuming the input of the LSS block is $\mathbf{X} \in R^{C\times H\times W}$ (the output of FAM), the output of the location-sensitive selection block is the location-sensitive weight $\beta \in R^{H\times W}$, the final discriminative enhanced feature maps map to the human part locations are $\mathbf{X}'_L \in R^{C\times H\times W}$. The location-sensitive weight β is generated by one convolution operation $\mathbf{W}_1 \in R^{1\times 1\times C}$ and another convolution operation $\mathbf{W}_2 \in R^{1\times 1\times 1}$, a Relu operation followed by a Sigmoid function in the input $\mathbf{X}$, i.e.,

$$\beta = Sigmoid(\sigma(W_2(\sigma(W_1(X))))) \tag{2}$$

where $\mathbf{W}$ denotes the convolution operation, the σ represents the Relu activation function and Sigmoid means the activation function.

Finally, the learned location-sensitive weight β is re-scaled on the input $\mathbf{X}$ and then adds back to the input to obtain the enhanced output $\mathbf{X}'_L$.

$$\mathbf{X}'_L = \mathbf{X} + \beta \otimes \mathbf{X}, \tag{3}$$

where $\otimes$ means the element-wise multiplication between $\mathbf{X}$ and β in the joint location context. In this way, we can capture the regional part-related representations in human context to reduce the perturbations resulting from the environment or background.

channel-wise selection Since convolutional filters operate as the pattern detectors. After the convolution, each channel of a feature map equals to the feature response of the corresponding filter. For the keypoint localization, different channels of features in CNNs generate different responses to different parts among the whole body. This selection mechanism produce a channel-wise vector by exploiting the inter-channel relationship of features. It can be viewed as a process of adaptively selecting the significant information for the human among different feature maps. After that, we can guide the network to eliminate the false alarms with the aid of the selected discriminative information.

The **Channel-wise Selection (CS)** module will assign larger weight to channels which show high responses to the human body parts.

As shown in Fig. 4, to aggregate the feature map among channels, we take $\mathbf{X}$ (the output of FAM) as the input of the channel-wise selection and the output of the channel-wise selection vector $\alpha \in R^C$, this vector softly encodes the global and local information for different feature maps. Thus we acquire the informative features $\mathbf{X}'_C \in R^{C\times H\times W}$ among the channel-wise context.

The CS weight matrix is computed in two steps analogous to the SE-Net [16] but differs from it. We utilize both average-pooled and max-pooled features simultaneously. We empirically confirm that exploiting both features greatly benefits eliminating the keypoint ambiguity instead of using them separately. Additionally, we choose element-wise add rather than other combination ways like concatenation by comparing the experimental results.

$$X = X_{avg} \oplus X_{max}, \tag{4}$$

where the $\oplus$ is the element-wise addition operation.

The channel-wise vector $z \in R^C$, where the c-th element of z is obtained by

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j), \tag{5}$$

where $x_c \in R^{H\times W}$ is the c-th element of the input $\mathbf{X}$.

In the next step, a selection mechanism with a sigmoid activation is acted on the channel-wise statistics z, i.e.,

$$\alpha = Sigmoid(W_2(\sigma(W_1(z)))), \tag{6}$$

where $\mathbf{W}_1 \in R^{C\times C}$ and $\mathbf{W}_2 \in R^{C\times C}$ represent two fully connected layers.

Finally, we get the final output $\mathbf{X}'_C$, similar to (3),

$$\mathbf{X}'_C = \mathbf{X} + \alpha \otimes \mathbf{X}, \tag{7}$$

where $\otimes$ means the element-wise multiplication in the channel-wise global context. After applying the above CS operation, we can better adaptively capture the fine-grained discriminative part-level representations for further boosting the localization performance.

Some examples can reflect the ISM refinement process are shown in Fig. 10. We can see false alarms for different body parts are reduced and heatmaps are corrected by selecting the significant human part regions. The ISM can correct the initial estimation of the occluded or similar joints relying on the discriminative location and channel-wise context information.

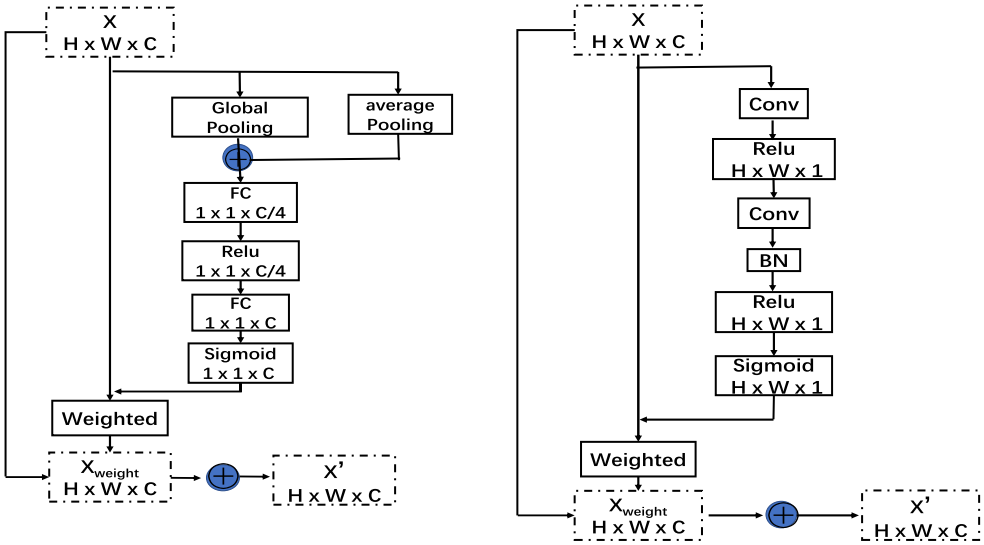

Fig. 4 Channel-wise Selection (left) and Location-Sensitive Selection (right), where X and X' mean the input feature and final enhanced feature respectively, X_{weight} means the weighted feature

3.2.3 Arrangement of the CS and LSS module

As shown in Fig. 3, assuming the input of residual bottleneck is $\mathbf{X} \in R^{C\times H\times W}$, the ISM is performed on the non-identity branch of the residual module after FAM. The ISM acts before adding up to the identity branch. The ISM is placed at the end of each ResNet [14] block. There are different combination strategies of the LSS and CS block in ISM. In particular, it can combine in serial or parallel. We empirically confirm that exploiting in parallel can greatly improve feature representation ability compared with in sequence. Thus we choose to combine them in parallel.

The FAM applies on the input of the residual bottleneck firstly and the feature selection is acted on the aggregated features. All processes are summarized as follows:

$$\begin{aligned} \mathbf{X} &= \mathbf{A}(\mathbf{M}), \\ \mathbf{X'} &= \mathbf{M} + (\mathbf{X} + \alpha \otimes \mathbf{X}) \oplus (\mathbf{X} + \beta \otimes \mathbf{X}), \\ \tilde{\mathbf{X}} &= \sigma(\mathbf{X'}), \end{aligned} \tag{8}$$

where the function $\mathbf{A}(\mathbf{M})$ represents the FAM module in the resnet bottleneck, and $\mathbf{X}$ represents the output of FAM from $\mathbf{A}(\mathbf{M})$. The $\tilde{\mathbf{X}}$ is the final output feature maps with the enhanced multi-scale discriminative feature information of the human.

3.3 Feature fusion module

Actually, the spatial-resolution information and semantic concepts can benefit each other. The deeper layers tend to encode a high-level semantic information, by contrast, the shallower layers are more likely to capture more spatial features and can better reconstruct spatial information just like the U-Net [43]. Feature fusion strategy is often employed in multi-person pose estimation in different forms. Many previous methods [7, 23, 34] combine the low-level and high-level features together. However, since the gap in semantic levels and spatial resolutions, making it's less effective to fuse different level features directly. Our approach is related to the ExFuse [61]. For semantic segmentation, it narrows the gap between spatial resolution and semantic-level features in a different way thereby progressively improves the performance. Motivated by this, we propose to take more spatial high-resolution structural feature into semantics and let more semantic information combine with high-resolution spatial features in semantic embedding way to enhance the global context information. In particular, for the spatial structural features, we assign semantic high-level information to combine with it at the head of the encoder (as depicted in green line) to make the spatial features encode more semantic concepts. For the semantic-level features, we involve more spatial information at the beginning of the encoder (as depicted in pink line) to guide the feature fusion in a semantic embedding way. When we obtain the fused global context information, we can better infer the invisible joints or disentangle the interacted joints.

Our FFM is shown in Fig. 2, the green line represents taking the semantics into low-level spatial features and the pink line represents taking the low-level spatial information into high-level features respectively.

3.4 Dense upsampling convolution

Multiple approaches can decode more accurate information which is generally lost in the downsampling process. For the pose estimation, different methods use the different upsampling forms to generate the final prediction heatmaps.

However, it's difficult for those common upsampling operations such as bilinear interpolation, deconvolution, transposed convolution, depth-to-space, bicubic upsampling etc to recover the detailed information better precisely due to their design isn't good at the fine-grained regression. Thus this will lead to the inaccurate keypoint localization inevitably.

Considering the flaws of the above methods, we find that the DUC [53] can better get over these obstacles and preserve the detailed information in the decoder. Concretely, given an input image $\mathbf{X} \in R^{C\times H\times W}$, the keypoint localization is to generate a heatmap with size $H \times W$ where each pixel is the coordinate of the location. We feed the image into ResNet [14] and output the feature maps $R^{h\times w\times c}$ before making predictions, where $h = \frac{H}{f}$, $w = \frac{W}{f}$, and f is the downsampling factor. The DUC directly adopts the convolution on the feature maps output by the encoder and then to generate the pixel-wise prediction map rather than pad zeros in the unpooling step just as the common upsampling methods. Figure 2 depicts our model architecture with a DUC module. The DUC is related to convolution operation, which is performed on the feature map output by ResNet [14] of $R^{h\times w\times c}$. Then we get further feature output of dimension $R^{h\times w\times f^2\times N}$, where N is the total number of joints. Each layer of the convolution operation learns the prediction for each joint. The output feature map is then reshaped to $R^{H\times W\times N}$ and apply a softmax function on it. In the end, we apply an element-wise argmax operation to get the final heatmaps. The core of DUC is to divide the whole heatmap into f^2 sub-parts which have the same height and width as the incoming feature map. That is, we transform the whole heatmap into a smaller map with multiple channels. This operation lets us apply the convolution operation directly between the input features and output heatmaps without inserting extra values in deconvolutional operation. Consequently, it avoids some detailed information missing in the process of the upsampling. We can improve the final prediction results by handling with the false alarms intrigued by inaccurate estimations to some samples with those information.

3.5 Objective function

The whole objective function for our overall framework in the training phrase is:

$$L_{pose} = \frac{1}{K}\sum_{K=1} \| h_k - h_{k'}\|_2^2, \tag{9}$$

where the h_k and $h_{k'}$ represent the confidence maps for the k-th joint predicted by our model and groundtruth heatmap, respectively. Here, the L_2 loss is chosen as the objective optimized function to train our model.

4 Experiments

Our multi-person pose estimation framework follows the top-down pipeline. First, we apply a human detector to produce all human bounding boxes in the image. Then for each bounding box, we apply our method to predict the corresponding human pose. The proposed method is evaluated on two recent standard multi-person datasets and a big new crowded dataset with many interactions and crowd cases: MPII Human Pose benchmark [1], COCO 2017 Keypoints Challenge dataset [24] and CrowdPose dataset [21].

From the perspective of the whole environment configurations, our model is implemented on the Pytorch [40]. For the training, 2 NVIDIA TITAN XP GPUs with the memory size of 12GB on a server are used and the BatchSize is set to 32. In our experiment, our ResNet backbones (ResNet-50, 101 and 152) are all initialized with the weights of the public-released Imagenet [44] pre-trained model. For the resolution of the input, we mainly adopt two different sizes (e.g., 256 × 192 and 388 × 284), the backbone ResNet-50 and the input size 256 × 192 is adopted in default. For the different datasets, we have few configuration differences in the training and the details are introduced in respective parts.

4.1 COCO keypoint detection

Datasets and Evaluation metric To evaluate the proposed method, we validate our model on the challenging COCO keypoint benchmark [24]. Our models are only trained on the COCO trainval dataset (includes $57K$ images and $150K$ person instances) without using any other extra data. Ablation studies are validated on the COCO minival dataset (includes $5K$ images). The final results are reported on the COCO test-dev dataset (includes $20K$ images) compared with the public state-of-the-art results. We use the standard evaluation metric [24] that reports the OKS-based AP (average precision) in the experiments, where the OKS (object keypoints similarity) defines the similarity between the predicted heatmap and the groundtruth heatmap.

Data Augmentation We apply random flip, rotation, and scale in our training stage. The flip value is 0.5. The scale range is ([0.7 ∼ 1.3]), and the random rotation range is ([-40° ∼ +40°]).

Training Adam [19] optimizer is adopted. The base learning rate is $5e-4$, and decreased by a factor of 0.1 at 90 and 120 epochs. Finally we train for 150 epochs on COCO dataset and we train ResNet-50 model which takes about 96 hours on MPII dataset. The input size of the image is resized to a fixed aspect ratio of height : width = 4 : 3 (e.g., 256 × 192 and 384 × 288), the same as the SBN [55].

Human Detector and Testing We adopt the same human detector as that in the SBN [55]. It is based on Faster-RCNN [42] with mAP of human category 56.4 AP. During testing, Soft-NMS [2] is used to suppress the duplicated bounding boxes. As a common practice like [7, 55], for the flipped image we average the predicted heatmap to get the keypoints location. Each keypoint location is predicted by adjusting the highest heatvalue location with offset from the maximum response to the second largest response.

Results on COCO val2017 We compare our model SHG [34], CPN [7], SBN [55] and HRNet [48] on the COCO minival dataset as shown in Table 1. It's noted that although the human detection result of the SHG and CPN (57.2%) is higher than ours (56.4%), our pose estimation accuracy still gains an obvious improvement than theirs. Based on the same input size of 256 × 192, compared with the 8-stage Hourglass, our model improves 5.9 AP and outperforms the CPN [7] and SBN [55] by 3.4 AP and 2.4 AP respectively. We also have an improvement of 2.2 AP and 1.6 AP for the input size of 384 × 288. Our method is only second to the HRNet [48] under the same condition. It designs a complex and robust deep high resolution network that can extract better multi-scale features of human across multi-stage fusions and thus achieves the best performance in the top-down methods. The qualitative visualization of the improvements between our model and the baseline model is shown in Fig. 5 on the COCO minival dataset.

As shown in Fig. 6, our approach can make better performance for pose estimation, although in challenging cases, (e.g., close-interactions in the first row, invisible joints generated by occlusions in the last row) where SBN [55] cannot well deal with.

Table 1 Comparison with the 8-stage Hourglass [34], CPN [7] SBN [55] and HRNet [48] on the COCO minival dataset

Models	Input Size	AP(OKS)
8-stage Hourglass [34]	256×256	67.1
8-stage Hourglass [34]	256×192	66.9
CPN [7]	256×192	68.6
CPN [7]	384×288	70.6
CPN* [7]	256×192	69.4
CPN* [7]	384×288	71.6
SBN [55]	256×192	70.4
SBN [55]	384×288	72.2
HRNet [48]	256×192	**73.4**
HRNet [48]	384×288	**75.8**
Ours*(ResNet-50)	256×192	*72.8*
Ours*(ResNet-50)	256×256	73.1
Ours*(ResNet-50)	384×288	*73.8*

'*' means the model training with the Online Hard Keypoints Mining. The best ones are bolded and the second best are underlined

Results on COCO test-dev 2017 Table 2 demonstrates the results of our method in the test-dev dataset of the COCO dataset. Our approach is obviously better than the existing promising approaches. Additionally, our CFENet (ResNet-50) achieves an AP of 72.8. It outperforms almost all the other top-down approaches under the same backbone network, provides a comparable result to the HRNet [48]. Compared to the SBN [55] with the same input size, our model still receives 1.3 improvements. Figure 7 illustrates some qualitative results generated using our method.

4.2 MPII human pose estimation

Datasets and Evaluation metric The MPII Human Pose dataset [1] consists of images taken from real-world activities with full-body pose annotations. There are about $25K$ images with $40K$ objects, where there are $12K$ objects for testing and the remaining objects for the training set. The data augmentation and training strategy keep same with COCO, except that the input size is cropped to 256 × 256 for fair comparison with other methods. For evaluation metric, the standard metric PCKh [1] (head-normalized probability of correct keypoint) score is used. The PCKh@0.5 score is reported by ours, 50% of the head size for normalization.

Training and Testing Adam [19] optimizer is adopted. The base learning rate is set to 0.001, and is decreased by a factor of 0.1 at 90 and 120 epochs, and finally we train for 120 epoch. Actually, we train ResNet-50 which takes about 12 hours. The input size of the image is resized to a fixed size (e.g., 256 × 256 and 384 × 384), the same as the SBN [55]. The testing procedure is almost the same as that in COCO except that we adopt the standard testing strategy using the provided person boxes instead of detected person boxes.

Results on MPII Human Pose Table 3 shows the PCKh@0.5 results, we reimplement the SBN [55] by using ResNet-152 as the backbone with the input size 256 × 256. Our CFENet achieves 92.0 PKCh@0.5, and outperforms the baseline and makes a comparable results with other methods [9, 18, 47, 51, 58] in the table. It proves that out model can obtain a relatively good performance on this dataset.

4.3 CrowdPose

Datasets and Evaluation metric CrowdPose dataset [21] contains $20K$ images in total and $80K$ human instances. The CrowdPose dataset divides into three crowding levels by **Crowd Index**: easy ($0 \sim 0.1$), medium ($0.1 \sim 0.8$) and hard

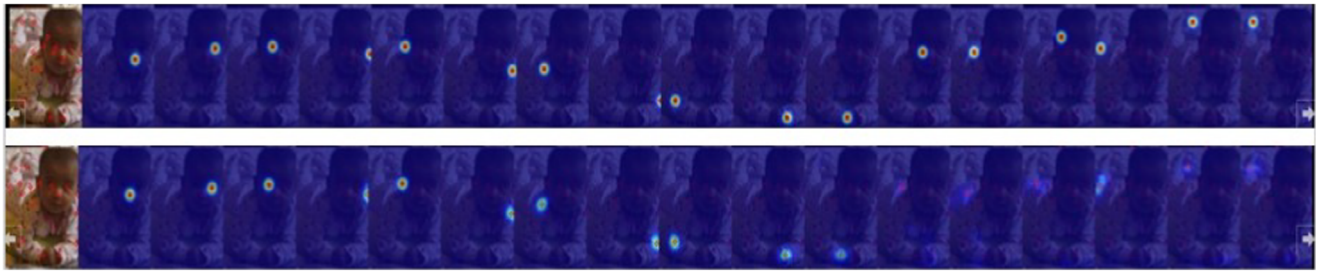

Fig. 5 Visual heatmaps of SBN [55] and our model on the COCO minival dataset. From left to right are input images, predicted heatmaps. The first row is generated by our method, the second row is generated by SBN [55]. Best viewed in color

(0.8 $\sim$ 1). Its Crowd Index satisfies uniform distribution in [0, 1]. CrowdPose dataset aims to promote performance in crowded cases and make models generalize to different scenarios. It uses the same evaluation metric as COCO.

Training and Testing The data augmentation and the training strategy are almost the same as COCO, except that the input size is cropped to 256 $\times$ 192 for fair comparison with other methods. The model is only trained on the CrowdPose training set for about 200 epochs and taking about 100 hours without using any extra data. The testing procedure is almost the same as that in COCO except that the provided human detector of SBN [55] is used for all methods to ensure a fair comparison. We extend 30% along the height and width directions of the detected human proposals for ensuring that the human parts can be extracted completely.

Results on CrowdPose The Table 4 show the quantitative results on CrowdPose test set. We can see that our method improves 4 $\sim$ 5 mAP compared with most of the advanced methods. It demonstrates the effectiveness of our proposed method in crowded scenes.

We also report the results on three crowding levels as stated before in Table 5, i.e., **uncrowded**, **medium crowded** and **extremely crowded**. It demonstrates that our method

Fig. 6 Comparison of SBN [55] and our model on the COCO minival dataset. From left to right are input images, SBN results, ours results. Best viewed in color

Table 2 Comparison of final results on the COCO test-dev dataset

Method	Backbone	Input Size	AP	AP .5	AP .75	AP (M)	AP (L)	AR
CMU-Pose [4]	-	-	61.8	84.9	67.5	57.1	68.2	66.5
Mask-RCNN [13]	ResNet-50-FPN	-	63.1	87.3	68.7	57.8	71.4	-
AE [33]	-	512×512	65.5	86.8	72.3	60.6	72.6	70.2
G-RMI [38]	ResNet-101	353×257	64.9	85.5	71.3	62.3	70.0	69.7
CPN [7]	ResNet-Inception	384×288	72.1	91.4	80.0	68.7	77.2	78.5
RMPE [11]	PyraNet	320×256	72.3	89.2	79.1	68.0	78.6	-
Integral Pose [49]	ResNet-101	256×256	67.8	88.2	74.8	63.9	74.0	-
MultiPoseNet [20]	-	-	69.6	86.3	76.6	65.0	76.3	73.5
CSANet [46]	ResNet-101	384×288	73.8	91.7	81.4	70.4	79.6	80.3
SBN [55]	ResNet-152	384×288	73.7	91.9	81.1	70.3	80.0	79.1
HRNet [55]	HRNet-W32	384×288	74.9	92.5	82.8	71.3	80.9	80.1
HRNet [48]	HRNet-W48	384×288	**75.5**	**92.5**	**83.3**	**71.9**	**81.5**	**80.5**
Mask R-CNN* [32]	ResNet-101-FPN	-	90.4	77.0	64.9	76.3	75.2	
G-RMI* [32]	ResNet-152	353×257	71.0	87.9	77.7	69.0	75.2	75.8
oks* [32]	-	-	72.0	90.3	79.7	67.6	78.4	77.1
bangbangren*$^{+}$ [32]	ResNet-101	-	72.8	89.4	79.6	68.6	80.0	78.7
CPN^{+} [32]	ResNet-Inception	384×288	73.0	91.7	80.9	69.5	78.1	79.0
HRNet* [48]	HRNet-W48	384×288	**77.0**	**92.7**	**84.5**	**73.4**	**83.1**	**82.0**
Ours	ResNet-50	256×192	72.8	91.5	80.5	69.0	77.1	76.4
Ours	ResNet-50	384×288	73.8	91.6	81.5	70.7	79.6	77.4
Ours	ResNet-101	256×192	73.5	91.6	81.6	70.6	78.2	77.2
Ours	ResNet-101	384×288	74.5	91.7	81.9	71.1	78.7	78.6
Ours	ResNet-152	256×192	74.2	91.8	82.1	71.3	79.1	77.9
Ours	ResNet-152	384×288	*75.0*	*92.0*	*82.2*	*71.6*	*80.2*	*79.4*

Top: methods in the literature, trained only with the COCO trainval dataset. **Middle**: results submitted to the COCO test-dev leaderboard [32]. "*" means that the method involves extra data for the training. "+" indicates the results using the ensembled models. **Bottom**: the results of our single model, trained only with the COCO trainval dataset

get more accurate results across all crowding levels, meanwhile as the scenes are more crowded the relative improvement increases.

Compared with the baseline SBN [55], our proposed approach has a significant improvement and we also provide comparable results on CrowdPose [21] across three crowding

Fig. 7 Qualitative results of our model on the COCO test-dev dataset. Our model can achieve more accurate result in pose estimation although there exist many problems such as extreme cases (the bottom-left result), complex and crowded poses (the top-right result)

Table 3 Results on MPII test set

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Insafutdinov et al. [17]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei et al. [54]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Newell et al. [34]	98.2	96.3	91.2	87.2	89.8	87.4	83.6	90.9
Sun et al. [47]	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
Tang et al. [52]	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
Ning et al. [36]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Luvizon et al. [29]	98.1	96.6	92.0	87.5	90.6	88.0	82.7	91.2
Chu et al. [9]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou et al. [8]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Chen et al. [6]	98.1	96.5	92.5	88.5	90.2	**89.6**	86.0	91.9
Yang et al. [58]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke et al. [18]	98.5	96.8	92.7	88.4	90.6	89.3	**86.3**	92.1
Tang et al. [51]	98.4	96.9	92.6	88.7	**91.8**	89.4	86.2	92.3
SBN et al. [55]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
HRNet et al. [48]	**98.6**	**96.9**	**92.8**	**89.0**	91.5	89.0	85.7	**92.3**
Ours	**98.6**	96.8	92.3	88.2	91.3	88.6	85.0	*92.0*

levels. CrowdPose [21] surpasses our final model by 1 $\sim$ 2 AP. This can be attributed to their special design for handling crowded scenarios. They add an additional global association algorithm after performing the modified single person pose estimation. Compared with them, our model can be trained in an end-to-end manner without using any extra post-processing method. We achieve comparable results with them, which can be attributed to our elaborate feature representations and effective decoding. Additionally, there are no other methods reporting experiment results on CrowdPose to our knowledge. The qualitative visualization on CrowdPose dataset is depicted in Fig. 8.

4.4 Ablation study

In this section, we will decompose our proposed CFENet to analyse the effect of each sub module. Moreover, we make all comparisons by experiment on COCO val2017 dataset, MPII valid set and CrowdPose valid set. The ablated results of the components are showed in Tables 6, 7 and 8 respectively. We adopt the input size 256 $\times$ 192 as the default setting, and the ResNet-50 is adopted for our default backbone.

4.4.1 Analysis of FASM&DUC modules

In this subsection, we will discuss the effect of the proposed FASM&DUC components respectively.

Feature Aggragation Module Pose estimation is a kind of sparse object detection. The multi-scale representation ability is vital for object detection. Therefore, it's important for the regression to have sufficient multi-scale representations. We use SBN [55] as the baseline, and replace the bottleneck of the ResNet [14] with our proposed FAM, while keeping other configurations unchanged. We can see that the model with the FAM achieves the AP of 70.9 AP in Table 6. It indicates that when only using the FAM, our model outperforms the baseline by 0.5 AP. The performance of the PCKh@0.5 achieves 0.4 higher than the baseline model on the MPII valid set in Table 7. Even for the crowded dataset, our FAM still achieves a large improvement of 3.0 mAP than the baseline in Table 8. From the results on the three datasets, it can be proved that our FAM plays an important role in improving the localization accuracy indeed.

To further interpret how the multi-scale features are aggregated and its impact on the keypoint prediction, we visualize the specific examples in Fig. 9. We can find that

Table 4 Results on CrowdPose test set

Models	mAP	mAP @0.5	mAP @0.75	mAR	mAR @0.5	mAR @0.75
Mask R-CNN [13]	57.2	83.5	60.3	65.9	89.5	69.4
AlphaPose [11]	61.0	81.3	66.0	67.6	86.7	71.8
SBN [55]	60.8	81.4	65.7	67.3	86.3	71.8
CrowdPose [21]	**66.0**	**84.2**	**71.5**	**72.7**	**89.5**	**77.5**
Ours	*64.2*	*82.2*	*69.5*	*70.7*	*87.7*	*75.5*

Table 5 Results on CrowdPose test set

Method	AP_{easy}	AP_{medium}	AP_{hard}
OpenPose [4]	62.7	48.7	32.3
Mask R-CNN [13]	69.4	57.9	45.8
AlphaPose [11]	71.2	61.4	51.1
SBN [55]	71.4	61.2	51.2
CrowdPose [21]	**75.5**	**66.3**	**57.4**
Ours	*73.6*	*63.7*	*55.6*

Test set is divided into three parts. Since the dataset is quite young, only a few works report results on it

Table 6 Ablation study of each component on the COCO minival dataset

Model	Baseline	FAM	ISM	FFM	DUC	AP(OKS)
1	√					70.4
2	√	√				70.9
3	√		√			71.5
4	√			√		70.9
5	√				√	70.6
6	√	√	√	√	√	**72.8**

different receptive fields in our FAM help learn the multi-granular human features (i.e., the human parts with large scale or smaller scale). At the same time, the aggregated results show that it can better aggregate the learned multi-granular feature representations between different keypoint positions.

Information Selection Module As shown in Table 6, the ISM achieves the AP of 71.5. It indicates that when only using ISM, our model outperforms the baseline SBN [55] by 1.1 AP. Similarly, it achieves an improvement of 0.3, 2.4 on the MPII and CrowdPose respectively. This demonstrates that it makes a big difference for achieving more precise pose estimation.

Moreover, we also explore the effects of different implementation orders of the LSS and CS block in the residual bottleneck. And we arrive at a conclusion that the best-combining strategy is in parallel, which can largely improve the accuracy in Table 9.

To interpret our ISM clearly, we visualize the estimated heatmaps with the ISM as shown in Fig. 10. Besides, we observe that ISM can solve many false alarms that happen in previous heatmaps caused by intertwined people, occlusion, and inaccuracy to achieve more precise location.

Dense Upsampling Convolution Here, we analyze the effect of DUC based on baseline framework. In this module, we can change the dimension of the top output feature maps of ResNet. We assume that the dimension of the output layer is $64 \times 64 \times 17$ in the baseline model (17 is the joint numbers), and for the same layer with DUC dimension is $64 \times 64 \times (r^2 \times 17)$. The r is the downsampling rate ($r = 8$ in our paper). We reshape the predicted heatmap to $512 \times 512 \times 17$. This module increases few parameters compared to the previous model. When we train this model of DUC, we adopt the similar set with the baseline and train for 50 epochs, and finally achieve a AP of 70.6 on the COCO minival dataset as shown in Table 6, a 0.2 AP increase compared to the baseline. As for the MPII and CrowdPose, it achieves 0.1 and 0.4 improvements respectively compared with the baseline in Tables 7 and 8. Although the improvement is slight, it still behaves better than the normal upsampling methods to help in pose estimation.

Fig. 8 Qualitative results of our model on the CrowdPose dataset. Our method can deal with challenges include appearance variation (the top-row), occlusion, clustered (the top-left results), and non-standard poses (the top-right results)

Table 7 Ablation experiments about each proposed module on MPII validation dataset (PCKh@0.5)

Model	Baseline	FAM	ISM	FFM	DUC	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
1	√					96.7	95.6	90.8	86.2	89.7	86.8	82.8	90.2
2	√	√				97.5	96.0	91.2	87.0	90.0	87.4	83.2	90.6
3	√		√			97.3	95.9	91.1	86.6	90.2	87.3	83.3	90.5
4	√			√		97.1	95.8	91.0	86.4	90.0	87.2	83.2	90.4
5	√				√	97.0	95.7	90.9	86.3	89.9	87.0	83.0	90.3
6	√	√	√	√	√	**97.9**	**96.2**	**91.2**	**87.1**	**90.0**	**87.4**	**83.5**	**90.8**

Table 8 Ablation experiments about each proposed module on CrowdPose validation dataset

Model	Baseline	FAM	ISM	FFM	DUC	mAP	mAR	AP_{easy}	AP_{medium}	AP_{hard}
1	√					61.1	67.5	71.5	61.4	51.4
2	√	√				64.1	70.4	73.2	63.4	55.6
3	√		√			63.5	69.6	72.8	62.8	54.8
4	√			√		62.0	68.4	72.3	62.4	52.8
5	√				√	61.5	67.9	71.9	61.8	51.9
6	√	√	√	√	√	**64.6**	**71.1**	**73.9**	**63.9**	**56.0**

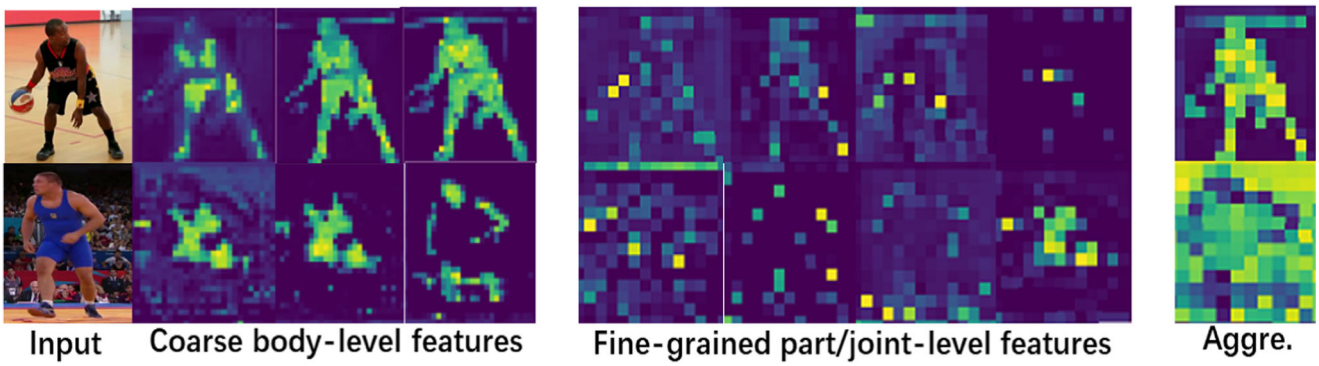

Fig. 9 The specific examples to demonstrate the intuitive interpretation of the multi-scale feature aggregation on the pose estimation. The deeper the color, the stronger the response of the learned human features. (Best viewed in color)

Table 9 Ablation study of the different implementations of CS and LSS on the COCO minival dataset, MPII valid set and CrowdPose valid set

Method	mAP(COCO)	mAP(CrowdPose)	PCKh@0.5(MPII)
SBN [55](baseline)	70.4	61.1	90.2
SBN [55]+CS+LSS	70.8	61.8	90.2
SBN [55]+LSS+CS	71.1	62.4	90.3
SBN [55]+CS&LSS in parallel	**71.5**	**63.5**	**90.5**

CS denotes the Channel-wise Selection, LSS denotes the Location-sensitive Selection

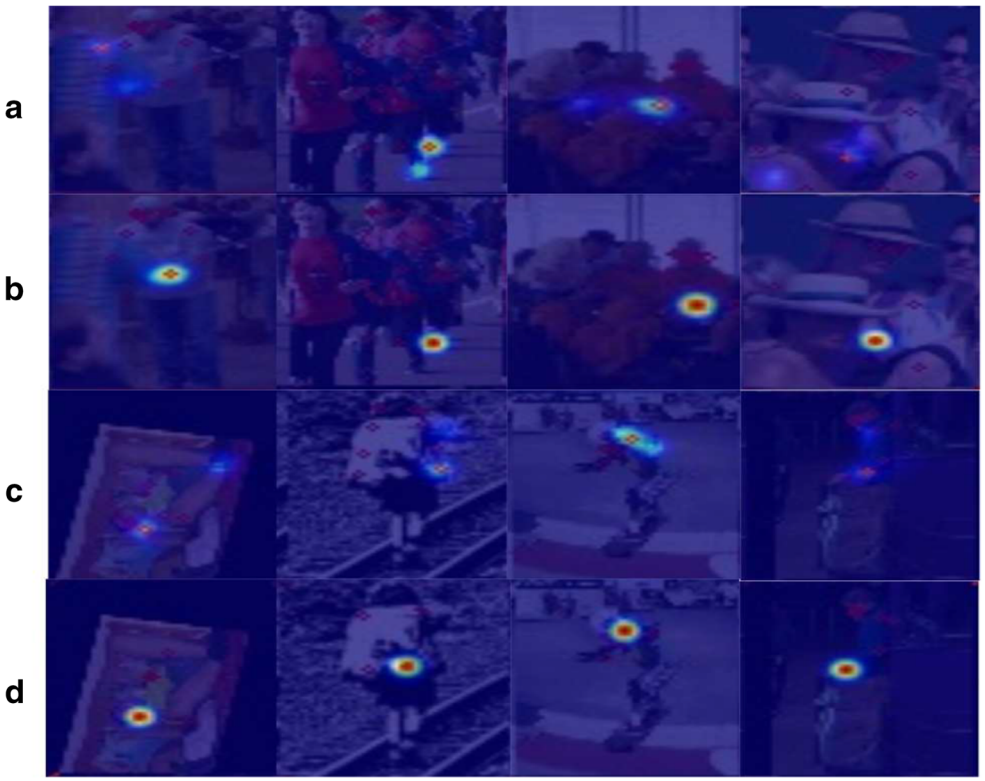

Fig. 10 Instances of the ISM solve various false alarms occurred in previous heatmaps. **A** and **C** show the previous heatmaps of baseline w/o ISM. (B) and (D) show the refined heatmaps of different human body parts through adaptively selecting the discriminative locations and channel-wise context information, which can reduce false alarms intrigued by incorrect estimations. The deeper the color, the larger value of the weight is. (Best viewed in color)

4.4.2 Is feature fusion enhanced ?

Here we discuss the influence of **FFM** used in our network. From the above table, the fusion strategy cannot always achieve the good performance only if the proper low-level encoding layer concatenates with the high-level decoder layer. The feature maps with different scales extracted from the ResNet [14] backbone denote as Conv-2 $\sim$ 5. In the decoder stage, we combine the low-level (Conv2) feature maps with the high-level (Conv5) feature maps together. The proposed CFENet makes a considerable improvement of 0.5 mAP, 0.2 score and 0.9 mAP on the three datasets respectively comparing with the baseline model. The comparison results also validate the effectiveness of the feature fusion strategy to some extent.

For demonstrating the effectiveness of FFM intuitively, we visualize the specific examples in Fig. 11 to help readers understand the effect of our feature fusion strategy on the feature learning of the human parts. It shows that the high resolution low-level feature map often represents more detailed spatial information such as edge, texture and silhouette, but lacks high-level semantic knowledge. High-level feature maps always include more semantic knowledge such as a discriminative part-related response and the human body context information that can help infer the joint locations. By fusing the above two kinds of complementary information via the FFM and making them benefit each other to provide more comprehensive information for the final keypoint predictions.

4.4.3 Bottleneck: The effective location to apply FASM.

We empirically exploit the most effective location to apply our FASM is the bottleneck of the model. The previous work SE-Net [16] mostly concentrate on making their contribution in the **convolution blocks** instead of the **bottlenecks**. By comparing different model performance in different locations on COCO minival dataset. We can conclude that applying the FASM at the bottleneck can achieve best result considering the speed and accuracy trade-offs just as shown in Table 10. It demonstrates that the right location to place the FASM is also important for the localization performance.

4.4.4 The discussion about the efficiency.

To further analyze the efficiency of our method, we conduct experiments to compare the parameters and GFLOPs of our model with others. The result is gathered in the Table 11. We can find that the parameters of our network are lower and less than the other counterparts except the SBN [55], while the number of GFLOPs are also lower with others and ours is slightly larger than the SBN [55] and CPN [7]. Especially, our model only with the FASM compare with the other

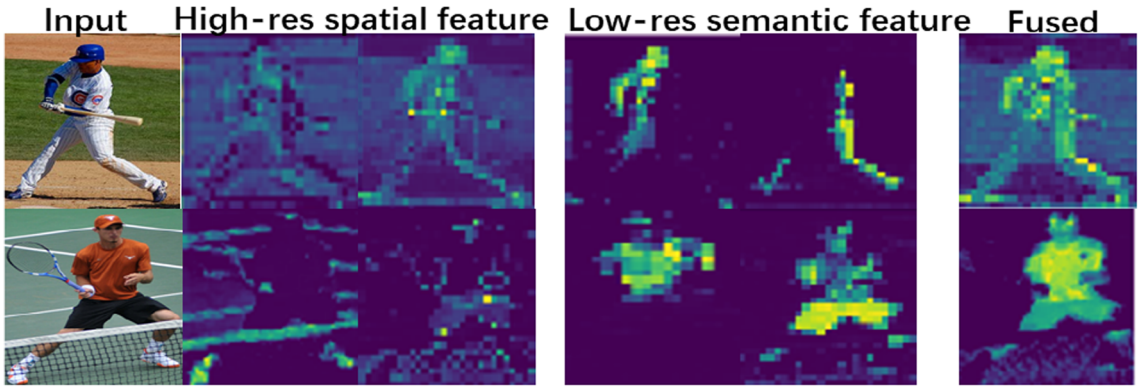

Fig. 11 Examples to illustrate the feature fusion strategy

multi-scale feature learning methods like CPN [7], HRNet [48] and SHG [34] behaves more efficient and lightweight.

By comparison, we can observe that our model can achieve better accuracy and speed trade-off at a little overhead cost on the whole.

4.4.5 Online hard keypoints mining

We also examine the impact of the OHKM strategy in our model. In particular, the mean squared error (MSE) is adopted as out loss function of the model to compare the predicted heatmaps and groundtruth heatmaps difference. We adopt 2D Gaussian with standard deviation of 1 pixel centered on the groundtruth location of each keypoint to produce the groundtruth heatmaps. The OHKM strategy in our method is adopted to calculate the top R ($R < K$) keypoint loss of K (K is the total number of labeled keypoints, say 17 in this case). The effect of value R is shown in Table 12. It shows that we achieve the best performance when $R = 8$. It also lets our model behave more competitive by considering the hard samples in experiment. Additionally, we set the value of R is 8 in this case.

4.4.6 Auxiliary strategy analysis

For further making the experiment details clear and proving the effectiveness of our proposed method, we investigate the relevant experimental strategies in detail. It mainly refers to the choice of human detector, the data preprocessing and the impact of the different backbones.

Table 10 Comparison of FASM in different points

Model	Params	GFLOPs	AP
SBN [55]	34.0M	8.90	70.4
SBN [55]+FASM-C	39.28M	10.52	70.8
SBN [55]+FASM(Our)	37.0M	9.23	**71.5**

FASM-C denotes where the module is inserted to convolution block

Human Detection Performance Table 13 shows the relationship between the human detection result and the corresponding pose estimation performance on the COCO minival dataset. Our model and SBN [55] are compared in this experiment. Both models are trained with the ResNet-50 backbone and the 256 × 192 input size. The SBN [55] adopts the Faster-RCNN [42] as the human detector, which reports the human detection AP 56.4 in their paper. We adopt the same human detector with them for fair comparison and the result illustrates the robustness of our method. Additionally, we also use the human detection results provided by the CPN [7] (reports the human detection AP 57.2) for comparison.

Effect of Data Pre-processing Here, we discuss the effect of the input sizes of our CFENet. All methods use the same backbone of the ResNet-50. As Table 14 illustrates, we find that our model outperforms the SBN [55] by 2.4 AP when the input size of 256 × 192. The AP increases 1.6 when the input size is 384 × 288.

Effect of Backbone Network To our acknowledgment, the backbone model becomes more complex and the performance

Table 11 Comparison of the model parameters and GFLOPs with the other methods

Model	Params	GFLOPs	AP
G-RMI [38]	42.6M	57.0	64.9
8-Stacked Hourglass [34]	89.0M	19.4	66.9
CPN [7]	102.0M	**6.20**	69.4
HRNet-w48 [48]	63.6M	14.60	75.1
SBN [55]	**34.0M**	8.90	70.4
SBN [55]+FASM	37.0M	9.23	71.5
Ours	44.0M	10.30	72.8

Table 12 Comparison of models of hard keypoints numbers in online hard keypoints mining strategy

R	7	8	11	13	15	16
AP(OKS)	71.3	**72.8**	71.2	71.6	71.2	72.3

will be better. We adopt the ResNet-50, ResNet-101, and ResNet-152 backbones respectively to conduct experiments with the input image size of 384 × 288. Table 15 shows that the backbone changes from ResNet-50 to ResNet-101 and ResNet-152, the corresponding AP increase is 0.7 and 1.2 respectively. It also further proves our proposed strategies are effective in achieving more precise keypoint localizations.

4.4.7 Qualitative failure cases analysis

Failure Cases of Others From the experimental results, it can be seen that the proposed method cannot achieve the best performance. We make a more comprehensive comparison further analyze that whether the failure tests of all the methods are same and then show what kind of scenarios the proposed method is unsuitable. The qualitative comparison results of our proposed method with the other existing methods are illustrated in Fig. 12, we can find that most of the existing methods can deal with the common situations that do not exist the challenging samples. However, the comparison results show some differences in the face of some difficult samples (i.e., overlapped parts or occluded/crowded humans), the failure tests of all the methods aren't the same due to the different models are designed for dealing with the different challenges.

Failure Cases of Ours Additionally, we also visualize some failure cases by our model in Fig. 13, it can be seen that our model cannot behave well when the scenarios that the humans are extreme crowded or severe occlusions exist between different humans. Although our method achieves an accurate keypoint localization on the above datasets, it still exists some limitations for dealing with much more challenging samples. The reason for failures mainly come from the imperfect design of our module and we will make further improvements to mend this.

Table 13 Comparison of pose estimation methods with different human detectors on the COCO minival dataset

Pose Method	Det Method	Human AP	Pose AP
SBN [55]	Faster-Rcnn [42]	56.4	70.4
CPN [7]	– [7]	57.2	69.4
RMPE [11]	SSD-512 [26]	55.5	70.3
Ours	Faster-Rcnn [42]	56.4	**72.8**
Ours	– [7]	57.2	**73.5**

All pose estimation methods are trained with the ResNet-50 backbone and the 256×192 input size

Table 14 Effect of input size of the image

Models	Input Size	AP(OKS)
SBN [55]	256×192	70.4
SBN [55]	384×288	72.2
Ours*(ResNet-50)	256×192	**72.8**
Ours*(ResNet-50)	256×256	**73.1**
Ours*(ResNet-50)	384×288	**73.8**

"*" means the model training with the Online Hard Keypoints Mining

4.5 Discussion and future work

Can this algorithm be generalized to other tasks? Our algorithm relies on the design of the proposed feature enhancement module. In general, from the perspective of the architecture, our proposed FASM and FFM can largely enhance the feature representations, which benefits most of the tasks that need the stronger feature representations. (i) From the perspective of the task, our algorithm may bring improvements on the tasks similar to ours (i.e., the structured prediction problems), such as segmentation and facial landmark detection. (ii) From the perspective of the feature learning, our proposed feature enhancement module such as FAM and ISM has good generalization ability for the other pixel-level computer vision tasks (i.e., semantic segmentation, landmark detection, video prediction and so on). F or example, the recent work [64] adopts the information selection mechanism (ISM) to boost the feature representations for video prediction and leading to a considerable improvement. Further, the proposed feature aggregation strategy can well represent the multi-scale features at a granular level and increases the range of receptive fields for each network layer. It constructs a hierarchical residual-like connections within one single residual block and can be plugged into the state-of-the-art backbone CNN models, e.g. , ResNet, ResNeXt, and DLA. Actually, we also evaluate the FAM block on

Table 15 Comparison of models with different backbone

Models	Input Size	AP(OKS)
8-stage Hourglass [34]	256×192	66.9
Ours(ResNet-50)	384×288	**73.8**
Ours(ResNet-101)	384×288	**74.5**
Ours(ResNet-152)	384×288	**75.0**

Fig. 12 Qualitative failure tests of the recent other methods. The samples are randomly sampled from the our adopted datasets

all these models and demonstrate consistent performance gains over baseline models on widely-used datasets, e.g. , CIFAR-100 and ImageNet. Similarly, our proposed ISM can selectively learn more useful information from the abundant features and filter out some distractive information, further enhancing the feature discrimination. It can largely enhance the feature representation ability for many tasks.

Limitations As we noted, our proposed method cannot achieve the best performance. The main limitations of our method include (i) The practical deployment of the model may be limited due to the complexity of the model. (ii) More attempts of the feature learning strategy can be adopted to obtain more abundant and general knowledge in practical and theoretical. For example, the proposed FAM mainly splits the features in parallel on average and extracts the multi-scale features in a group-wise manner and fuses them directly, while more different variants are underexplored actually. The FFM strategy may be too simple to fuse the multi-level information. The adopted DUC cannot be compared to the best method from our experimental results, although it alleviates the information loss to some extent.

Further improvements According to the above analysis, we have further reconsidered our module to think what can be improved in the future. It can be mainly described in the following two aspects: (i) from the need of the practical application, the computational calculations and FLOPs can be further reduced to relieve the burden of the platform. This needs more lightweight feature extraction modules,

Fig. 13 Qualitative failure results of our model. Our method cannot better deal with the extreme crowded scenarios (the top-row), severe human interactions/occlusions (the bottom-left/right results), and extreme poses (the bottom-middle results)

accelerating the speed of our model; (ii) from the view of solving the challenging issues of the pose estimation, some specific improvements based on our proposed method can be explored in the future work. Correspondingly, we will try more variants for the design of FAM, such as to split the features randomly rather than on average and aggregate the multi-scale features not only in parallel but also along the vertical directions. In this way, we can learn more comprehensive multi-scale features in a global and local view. Before the ISM module, we can try the feature shuffle operations to further increase the diversity of features for the better discriminative feature learning. Moreover, we will explore much better strategy (i.e., like the HRNet or recent CARAFE [63]) to replace the DUC for eliminating or making up for the lost information in decoder. At the same time, more elaborate and effective feature fusion strategies can also be explored in the future research.

In addition to improving the effectiveness of the proposed module, we can also try to design more effective feature learning mechanisms that can be applied in many other tasks. For example, recently we consider the feature learning from a new perspective. Concretely, we explore the feature relations between neighboring feature maps with a dynamic graph to learn the latent relationships between keypoint locations. It can effectively improve the localization accuracy. In summary, we will continuously explore more effective and efficient ways to accelerate the feature learning. They can solve the specific issues of the pose estimation, or provide a general feature enhancement strategy.

5 Conclusion

In this paper, we handle the multi-person pose estimation with the top-down pipeline. The Feature Aggregation and Selection Module (FASM) embedded in residual bottleneck is designed to adaptively highlight the discrimination of the multi-scale feature representations, for precise joint location and rich spatial context. The Feature Fusion (FF) strategy bridges the gap between semantic information and spatial structural features in semantic embedding to make them reinforce each other. The Dense Upsampling Convolution (DUC) module is adopted to produce elaborate predictions on feature maps to help recover detailed keypoint information. Overall, our model achieves competitive performance consistently on popular multi-person pose estimation datasets, especially a dataset in crowded scenes. Although our method cannot achieve the best performance, we hope that it makes progress in inspiring more effective and reliable feature learning strategies for the multi-person pose estimation.

Acknowledgements This research is supported by National Nature Science Foundation of China under Grants (No.61473031), the Fundamental Research Funds for the Central Universities (2019JBM019).

References

1. Andriluka M, Pishchulin L, Gehler PV, Schiele B (2014) 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3686–3693
2. Bodla N, Singh B, Chellappa R, Davis LS (2017) Soft-nms improving object detection with one line of code. In: Proceedings of the IEEE International Conference on Computer Vision, pp 5562–5570
3. Buitelaar P, Wood I, Negi S, Arcan M, Mccrae JP, Abele A, Robin C, Andryushechkin V, Sagha H, Schmitt M (2018) Mixedemotions: an open-source toolbox for multi-modal emotion analysis. IEEE Trans Multimed 20(9):2454–2465
4. Cao Z, Simon T, Wei S, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1302–1310
5. Chen L, Papandreou G, Kokkinos I, Murphy KP, Yuille AL (2018a) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
6. Chen Y, Shen C, Wei X, Liu L, Yang J (2017) Adversarial posenet: A structure-aware convolutional network for human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1221–1230
7. Chen Y, Wang Z, Peng Y, Zhang Z, Yu G, Sun J (2018b) Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7103–7112
8. Chou C, Chien J, Chen H (2018) Self adversarial training for human pose estimation. In: Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pp 17–30
9. Chu X, Yang W, Ouyang W, Ma C, Yuille AL, Wang X (2017) Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5669–5678
10. Dong L, Chen X, Wang R, Zhang Q, Izquierdo E (2018) Adore: an adaptive holons representation framework for human pose estimation. IEEE Trans Circ Syst Video Technol 28(10):2803–2813
11. Fang H, Xie S, Tai Y, Lu C (2017) Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2353–2362
12. Gao S, Cheng M, Zhao K, Zhang X, Yang M, Torr PHS (2019) Res2net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell:1–1
13. He K, Gkioxari G, Dollar P, Girshick RB (2017) Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2980–2988
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
15. Hou Q, Cheng M, Hu X, Borji A, Tu Z, Torr PHS (2017) Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 815–828

16. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7132–7141
17. Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B (2016) Deepercut: a deeper, stronger, and faster multi-person pose estimation model. Eur Conf Comput Vis:34–50
18. Ke L, Chang M, Qi H, Lyu S (2018) Multi-scale structure-aware network for human pose estimation. Eur Conf Comput Vis:731–746
19. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. Computer Science
20. Kocabas M, Karagoz S, Akbas E (2018) Multiposenet: Fast multi-person pose estimation using pose residual network. Eur Conf Comput Vis:437–453
21. Li J, Wang C, Zhu H, Mao Y, Fang H, Lu C (2019a) Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
22. Li W, Wang Z, Yin B, Peng Q, Su J (2019b) Rethinking on multi-stage networks for human pose estimation. arXiv:190100148
23. Lin T, Dollar P, Girshick RB, He K, Hariharan B, Belongie SJ (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 936–944
24. Lin T, Maire M, Belongie SJ, Hays J, Perona P, Ramanan D, Dollar P, Zitnick CL (2014) Microsoft coco: Common objects in context. Eur Conf Comput Vis:740–755
25. Liu S, Li Y, Hua G (2019) Human pose estimation in video via structured space learning and halfway temporal evaluation. IEEE Trans Circ Syst Video Technol:2029–2038
26. Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu C, Berg AC (2016) Ssd: Single shot multibox detector. Eur Conf Comput Vis:21–37
27. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3431–3440
28. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
29. Luvizon DC, Tabia H, Picard D (2017) Human pose regression by combining indirect part detection and contextual information. Computers & Graphics
30. Marcosramiro A, Pizarro D, Marronromera M, Gaticaperez D (2015) Let your body speak: Communicative cue extraction on natural interaction using rgbd data. IEEE Trans Multimed 17(10):1721–1732
31. Martinezgonzalez A, Villamizar M, Canevet O, Odobez J (2019) Efficient convolutional neural networks for depth-based multi-person pose estimation. IEEE Trans Circ Syst Video Technol:1–1
32. MS-COCO, 2017 Coco keypoint leaderboard. http://cocodataset.org/
33. Newell A, Huang Z, Deng J (2017) Associative embedding: End-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems, pp 2277–2287
34. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. Eur Conf Comput Vis:483–499
35. Nie X, Feng J, Zhang J, Yan S (2019) Single-stage multi-person pose machines. In: Proceedings of the IEEE International Conference on Computer Vision, pp 6951–6960
36. Ning G, Zhang Z, He Z (2018) Knowledge-guided deep fractal neural networks for human pose estimation. IEEE Trans Multimed:1246–1259
37. Noh H, Hong S, Han B (2015) Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1520–1528
38. Papandreou G, Zhu T, Kanazawa N, Toshev A, Tompson J, Bregler C, Murphy KP (2017) Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3711–3719
39. Park J, Woo S, Lee J, Kweon IS (2018) Bam: Bottleneck attention module. arXiv:180706514
40. Facebook (2017) Pytorch. https://pytorch.org/
41. Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler PV, Schiele B (2016) Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4929–4937
42. Ren S, He K, Girshick RB, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp 91–99
43. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. Med Image Comput Comput Assist Intervent:234–241
44. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein MS et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
45. Mahshid M, Reza S (2019) A motion-aware convLSTM network for action recognition. Appl Intell:1–7
46. Su K, Yu D, Xu Z, Geng X, Wang C (2019) Multi-person pose estimation with enhanced channel-wise and spatial information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5674–5682
47. Sun K, Lan C, Xing J, Zeng W, Liu D, Wang J (2017) Human pose estimation using global and local normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp 5600–5608
48. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5693–5703
49. Sun X, Xiao B, Wei F, Liang S, Wei Y (2018) Integral human pose regression. Eur Conf Comput Vis:536–553
50. Szegedy C, Liu W, Jia Y, Sermanet P, Reed SE, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1–9
51. Tang W, Yu P, Wu Y (2018a) Deeply learned compositional models for human pose estimation. Eur Conf Comput Vis:197–214
52. Tang Z, Peng X, Geng S, Wu L, Zhang S, Metaxas DN (2018b) Quantized densely connected u-nets for efficient landmark localization. Eur Conf Comput Vis:348–364
53. Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, Cottrell GW (2018) Understanding convolution for semantic segmentation. Workshop Appl Comput Vis:1451–1460
54. Wei S, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4724–4732
55. Xiao B, Wu H, Wei Y (2018) Simple baselines for human pose estimation and tracking. Eur Conf Comput Vis:472–487
56. Xiao Y, Lu H, Sun C (2015) Pose estimation based on pose cluster and candidates recombination. IEEE Trans Cir Syst Video Technol 25(6):935–943
57. Yang B, Ma AJ, Yuen PC (2019) Body parts synthesis for cross-quality pose estimation. IEEE Trans Circ Syst Video Technol 29(2):461–474
58. Yang W, Li S, Ouyang W, Li H, Wang X (2017) Learning feature pyramids for human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1290–1299

59. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. Eur Conf Comput Vis:818–833
60. Zhang X, Ding M, Fan G (2017) Video-based human walking estimation using joint gait and pose manifolds. IEEE Trans Circ Syst Video Technol 27(7):1540–1554
61. Zhang Z, Zhang X, Peng C, Xue X, Sun J (2018) Exfuse: Enhancing feature fusion for semantic segmentation. Eur Conf Comput Vis:273–288
62. Dinh D, Lim M, Thang N, Lee S, Kim T (2014) Real-time 3D human pose recovery from a single depth image using principal direction analysis. Appl Intell 41(2):473–486
63. Wang J, Chen K, Xu R, Liu Z, Joy C, Lin D (2019) CARAFE: Content-Aware ReAssembly of FEatures. Int Conf Comput Vis:3007–2016
64. Lin X, Zou Q, Xu X, Huang Y, Tian Y (2020) Motion-Aware Feature enhancement network for video prediction. IEEE Trans Circ Syst Video Technol PP(99):1–1

Xixia Xu received the B.S. degree in software engineering from the Lanzhou Jiao tong University, Lan Zhou, China, in 2018. She is currently pursuing the Ph.D. degree with the Department of Computer and Information Technology, Beijing Jiao tong University, Beijing, China, in 2019. Her current research interests include computer vision, image processing, and machine learning with the applications on 2d multi-person pose estimation analysis and human.

Qi Zou received the Ph.D. degree in computer science from Beijing Jiao Tong University, Beijing, China, in 2006. She is currently a Professor and a Doctoral Supervisor with the School of Computer and Information Technology, Beijing Jiao Tong University. She has authored or co-authored over 30 papers in peer-reviewed journals and conferences. Her research interests include computer vision, pattern recognition, and intelligent.

Xue Lin received the B.S. and M.E. degree in School of Information Science and Engineering from University of Jinan, Jinan, China, in 2015 and 2018, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer and Information Technology, Beijing Jiao tong University, Beijing, China, in 2018. Her research interests are computer vision, image processing, and machine learning, as well as their applications on human-centric activity.