Signal, Image and Video Processing (2024) 18:5647–5661
https://doi.org/10.1007/s11760-024-03261-7

ORIGINAL PAPER

Combining self-attention and depth-wise convolution for human pose estimation

Fan Zhang[1] · Qingxuan Shi[1,2] · Yanli Ma[1]

Received: 17 November 2023 / Revised: 26 March 2024 / Accepted: 2 May 2024 / Published online: 13 June 2024

Abstract

The implementation of convolutional neural networks (CNNs) and Transformers has significantly accelerated the booming advances of human pose estimation. However, challenges persist in accurately estimating target details. Pose estimation faces inherent difficulties when applied in complex environments marked by intricate conditions, including factors like motion blur and chaotic scenes. In this paper, we revisit the design of CNNs and Transformers, delving deeper into their internal structures. We sequentially utilize CNNs and Transformers, leveraging the proficiency of CNNs in extracting low-level features and the capability of Transformers in establishing long-range dependencies. Building upon this framework, we introduce modifications to both CNNs and Transformer-related structures, enhancing the overall expressive capacity of the model. The modification is made to the original CNNs section: we alter BasicBlock to AtBlock to maintain high-resolution information exchange to further excavate details of objects. The two following modifications are applied to the subsequent Transformer section: (1) we replace the self-attention layer in each encoder block with the local enhancement self-attention module to capture local information. (2) We propose a local perception feed-forward network to substitute for the feed-forward network layer in each encoder block, which employs the depth-wise convolution to enhance the correlation of neighbour information in the spatial dimension. Our modifications contribute to analyzing poses that are arduous to estimate due to occlusion. Our method combines self-attention and depth-wise convolution, named CSDNet. The experiments on both COCO2017 and MPII datasets show improved performance over the baseline. Compared to other models achieving the similar accuracy, our model has fewer parameters and requires less computation. Additionally, in complicated environments such as poor lighting conditions, our method can more accurately estimate fuzzy keypoints.

Keywords Computer vision · Human pose estimation · Transformer · Attention mechanism

1 Introduction

Human pose estimation (HPE) determines the spatial position of human keypoints from sensor input to obtain a representation of the human pose. HPE has wide-ranging practical applications, including but not limited to action recognition [1–3], action detection [4], human-object interaction [5, 6], and person re-identification [7, 8]. The synergistic cooperation of convolutional neural networks (CNNs) and Transformers leverages the strengths of both approaches, with CNNs excelling in extracting intricate low-level features, while Transformers excel in capturing long-range dependencies. This powerful combination drives its vigorous advancement in HPE.

The convolution operation considers the local spatial context at various levels, capturing information from simple low-level edges and textures to more complex higher-order semantic patterns. It can effectively establish connections between nearby keypoints, such as the ankle and knee. However, it is limited in modeling dependencies between keypoints that are further apart. This limitation arises due to the poor scalability of convolution. Repeatedly stacking convolutions to expand the receptive field does not adequately capture global dependencies. Moreover, that increases the complexity and computational cost of the model. The self-attention layers of Transformers can capture interactions between any pair of positions and excel at associating the

✉ Qingxuan Shi
qingxuanshi@hbu.edu.cn

[1] School of Cyber Security and Computer, Hebei University, Baoding 071002, China

[2] Hebei Machine Vision Engineering Research Center, Hebei University, Baoding 071002, China

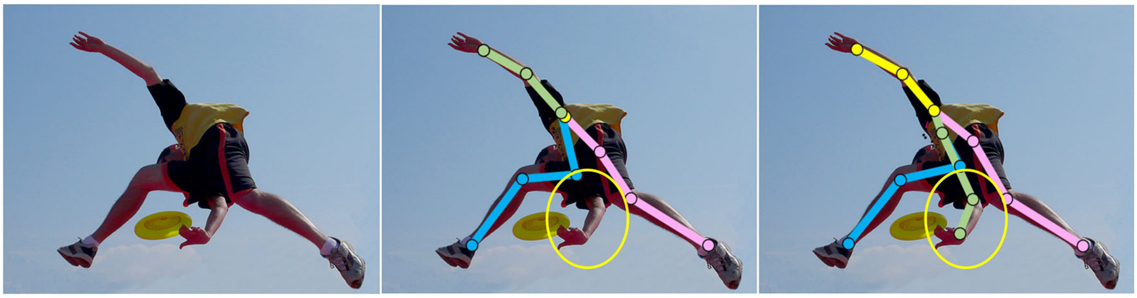

Fig. 1 An example of human pose estimation in complex poses. From left to right, the images represent the original image, prediction results of HRNet, and our prediction results. The area surrounded by yellow ellipses is a comparison of estimates between different methods. We observe that our method can better estimate complex poses compared to HRNet

long-range dependencies. It can progressively expand sizes of the receptive fields via aggregating information from other tokens. Based on these observations, we incorporate CNNs with Transformers to integrate the capability of CNNs in extracting local spatial context and the capability of Transformers in modeling long-range dependencies. Input images first pass through the convolution to extract low-level features, then resulting partitioned patches are fed into Transformer layers to acquire high-level global dependencies.

To enhance CNNs' performance on fine-grained tasks [9] and improve their ability to detect target details, we introduce attention mechanisms for fine-grained visual recognition tasks. This allows the model to adaptively focus on crucial areas in the context. Specifically for pose estimation, our aim is to better distinguish blurred human bodies and keypoints in complex environments, as well as distinguishing between human bodies and similar surrounding individuals or environments. Previous approaches, such as channel attention [10], explicitly model the dependencies between channels but may suffer from spatial information collapse due to pooling. While CBAM [11] compresses spatial and channel features separately, it fails to maintain high resolution internally. DANet [12], on the other hand, preserves high resolution in both attention branches but comes with a high computational cost. In contrary to CBAM and DANet, Polarized Self-Attention (PSA) [13] maintains high resolution in one direction of the branch while compressing features in its orthogonal branches. It leverages softmax and sigmoid functions to fit the output distribution. This preserves high resolution internally to mitigate potential resolution loss from downsampling while maintaining a reasonable computation overhead. In this paper, we design AtBlock which employs PSA to make the model more accurate in detecting object details.

Transformers excel at modeling the global dependencies among tokens using self-attention layers. However, the fixed token size limits their ability to account for the 2D structure and spatial local information within each block. To address this limitation, we introduce the local enhancement self attention (LESA) module. We leverage depth-wise convolutions to aggregate nearby information to enhance local information exchange. The original Multi-Head Attention (MHA) layer is replaced with the LESA module. In traditional feed-forward networks, fully connected layers are point-wise and may fail to learn cross-tagged information. Complementary to the MHA module, the feed-forward network module can supplement local information based on the long-range dependencies established by Transformers. Therefore, we introduce the local perception feed-forward network (LPFFN) layer, which enhances local information interaction using depth-wise convolutions. After the MHA module in each encoder, the original feed-forward network is replaced with the LPFFN.

The experiments are conducted on both COCO2017 and MPII datasets, yielding significant improvements over the baseline. In comparison to other models achieving similar accuracy, our model has fewer parameters and requires less computation. Ablation experiments are performed on each module to verify their effectiveness. HRNet [14] is a classic pose estimation network. It excels in estimating human body poses and is frequently used for benchmarking pose estimation results. A detailed introduction to HRNet is provided in Sect. 2.4. We visualize the experimental results and compare them with the visualization results of HRNet. As shown in Fig. 1, in complex poses, our model can more accurately estimate overlapping or occluded limbs to better estimate complete poses. Moreover, in challenging environments such as poor lighting conditions, our model performs better in estimating fuzzy keypoints.

2 Related work

2.1 Attention mechanisms

Attention mechanisms play a crucial role in human pose perception [15–17]. They assign substantial weights to the most informative features and suppresses less useful expressions. This adaptive focusing enables the model to emphasize the prominent parts effectively. SE-Net [10] first presents an effective channel attention learning mechanism that models dependencies between channels, resulting in an excellent performance. CBAM [11] combines both average and max pooling to aggregate features and incorporates spatial attention and channel attention. ACmix [18] reveals the powerful latent relationship between self-attention and convolution and elegantly integrates them. That combines their strengths and avoids computationally expensive operations. BiFormer [19] introduces a novel dynamic sparse attention via bi-level routing for more flexible computation allocation with content-awareness. However, the above methods either focus on designing more intricate attention modules that inevitably incur greater computational costs, or fail to establish a long-range channel dependency. In this paper, we employ PSA to preserve high-resolution semantics in attention calculation for information interaction while maintaining low computational cost. That helps address the fine-grained regression problem which were previously neglected.

2.2 Transformers

Motivated by the remarkable success of Transformers in natural language processing (NLP) [20, 21], many attempts [22–24] have been made to introduce Transformer architectures to vision tasks. ViT [25] innovatively applies the Transformer architecture, inherited from NLP, to the field of computer vision. Specifically, ViT decomposes images into a series of fixed-length token sequences. Subsequently, DeiT [26] introduces token-based distillation to reduce the data required for training Transformers. LocalViT [27] enhances the locality of the Visual Transformer by incorporating depth-wise convolution into the feed-forward network. Swin Transformer [28] partitions inputs into non-overlapping windows, constraining self-attention within each local window. While it performs well, its computational complexity grows linearly with the number of input tokens. NomMer [29] can dynamically nominate the synergistic global–local context in vision Transformers. Global and local context can adaptively contribute based on different visual data and tasks. The deformable self-attention module in DAT [30] transfers candidate keys/values to crucial regions, allowing the self-attention module to focus on relevant areas. The work of CvT [31] aims to introduce convolution with image domain specific inductive bias into the Transformer architecture, achieving promising results by incorporating both. SMT [32] also integrates convolutional networks and vision Transformers. It introduces the Multi-Head Mixed Convolution (MHMC) module and the Scale-Aware Aggregation (SAA) module to further enhance convolutional modulation. In this paper, we carefully explore how to better utilize the advantages of different components of Transformers and compensate for their shortcomings in specific visual tasks.

2.3 Human pose estimation

Deep convolutional neural networks have demonstrated remarkable success in the field of human pose estimation [33–36]. This can be attributed to their inherent inductive bias, which includes translation invariance and locality, enabling them to efficiently extract local features from low-level images. Several sophisticated strategies have been incorporated into network designs, including multi-scale fusion [37–39], stacking [40, 41], and high-resolution representation [14]. These strategies lead to the development of several notable network architectures, such as CPM [42], Hourglass network [41], FPN [43], CPN [44], SimpleBaseline [40], HRNet [14] and many more. SimpleBaseline [40] designs a simple architecture by stacking transposed convolution and achieves promising results. The subsequent HRNet [14] maintains a high-resolution representation throughout the whole network to provide accurate heatmap estimation.

Capturing global dependencies is essential for accurate human pose estimation [45–47]. The introduction of Transformers based on self-attention enables the effective acquisition of remote dependency information required for visual tasks. This overcomes the limitations of CNNs, where stacking convolutions cannot effectively expand the receptive field and it is difficult to obtain global dependent information. Additionally, Transformers can better model the constraint relationships between keypoints. PRTR [48] gradually refines the location of the estimated critical points using a cascade approach. Similar to HRNet structure, HRFormer [49] proposes a parallel Transformer module to fuse multi-resolution features. ViTPose [50] utilizes plain and non-hierarchical vision Transformers as backbones for feature extraction. It shows the excellent capabilities of large plain vision Transformer models for pose estimation from various aspects. While these approaches improve model structures and achieve outstanding results, purely Transformer-based large models come with significant memory and computational costs. Other methods combine the advantages of convolution and Transformers. For example, TransPose [51] initially employs CNNs for extracting low-level features and then utilizes attention layers to capture global correlations. It reveals the long-term dependencies in predicting keypoints. Different from previous approaches that apply Transformers

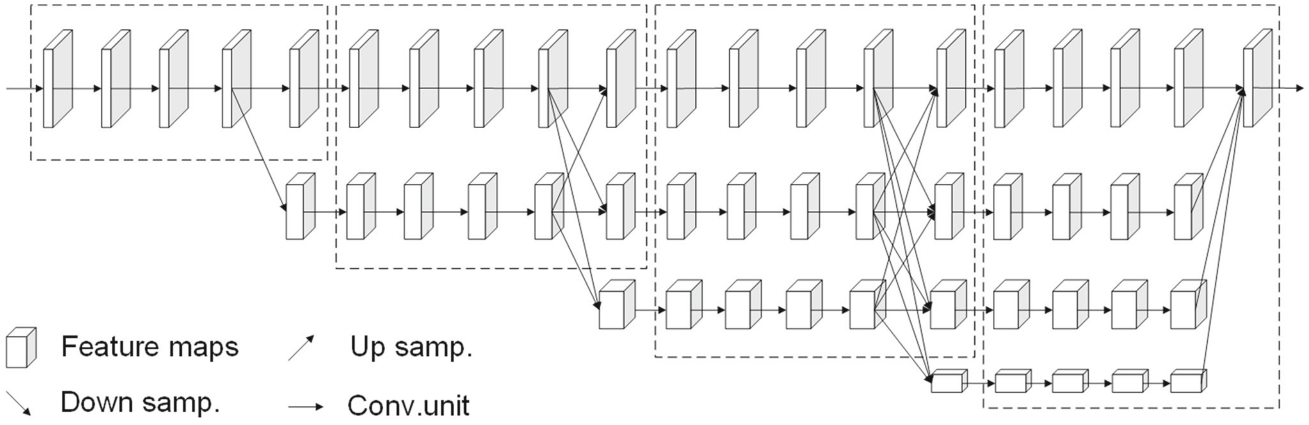

Fig. 2 The architecture of HRNet

[52, 53] to output a one-dimensional sequence of joint/vertex coordinates, TransPose utilizes Transformers to predict the 2D heatmaps represented with spatial distributions of keypoints. However, it lacks the ability to directly model the constraint relationships between keypoints. TokenPose [54] explicitly embeds keypoints as tokens, simultaneously learning visual cues and constraint relationships through self-attention interactions. UniFormer [55] employs different tokens in shallow and deep layers, seamlessly integrating the benefits of both convolutional and self-attention mechanisms via Transformers. In this paper, we re-examin the design of CNNs and Transformers, delving deeper into their internal structures. We first modify CNNs BasicBlocks to alleviate the resolution loss caused by downsampling. Then, we improve the Transformer layers to enhance its ability to extract local information while extracting global correlations. These explorations aim to enhance the overall expressive capability of the model and explore more possibilities in model architecture.

2.4 HRNet

As shown in Fig. 2, HRNet maintains high resolution throughout the process and employs parallel connections from high to low resolution. It starts with a high-resolution branch in the initial stage. It gradually adds branches from high to low resolution in each subsequent stage, with the resolution of the new branch being half of the lowest resolution in the previous stage. HRNet performs multiple multi-scale fusions, allowing each representation to repeatedly receive information from other parallel representations, thus obtaining rich high-resolution representations. This enhances the accuracy of predicted keypoint heatmaps. HRNet is frequently chosen as the backbone network in various pose estimation studies due to its outstanding performance. For meaningful comparisons and more intuitive representations, we also select HRNet as the backbone network for our research.

3 Method

We present a hybrid network architecture as shown in Fig. 3 that effectively combines the strengths of CNNs and Transformers. CNNs is utilized to extract low-level image features efficiently. Meanwhile, Transformers facilitate higher-order information exchange, allowing for the effective capture of global dependencies. That allows Transformers to operate directly on feature maps processed by CNNs, rather than requiring access to the original image. To further improve model performance, we introduce several modifications to both the corresponding CNN and Transformer sections. Specifically, we replace the BasicBlock in HRNet with our AtBlock in the CNN section. In the Transformer section, we propose local enhancement self-attention (LESA) and local perception feed-forward network (LPFFN) to replace traditional Transformer layers.

3.1 Image to tokens

We aim to model long-range dependencies at high resolution and minimize information loss caused by resolution reduction. Attention mechanisms have shown promising results in fine-grained tasks, allowing models to adaptively focus on important areas in the context. Therefore, we incorporated an attention mechanism into our approach. As shown in Fig. 4, we compare several commonly used channel attention and spatial attention mechanisms, and ultimately adopt Polarized Self Attention (PSA).

As shown in Fig. 4b, e, the first channel attention mechanism and spatial attention mechanism compress the internal resolution and lead to a significant loss of image information, thereby hindering subsequent image feature extraction. On the other hand, as shown in Fig. 4c, f, the second channel attention mechanism and spatial attention mechanism preserve high resolution in the spatial dimension but involves redundant computations when generating a large attention

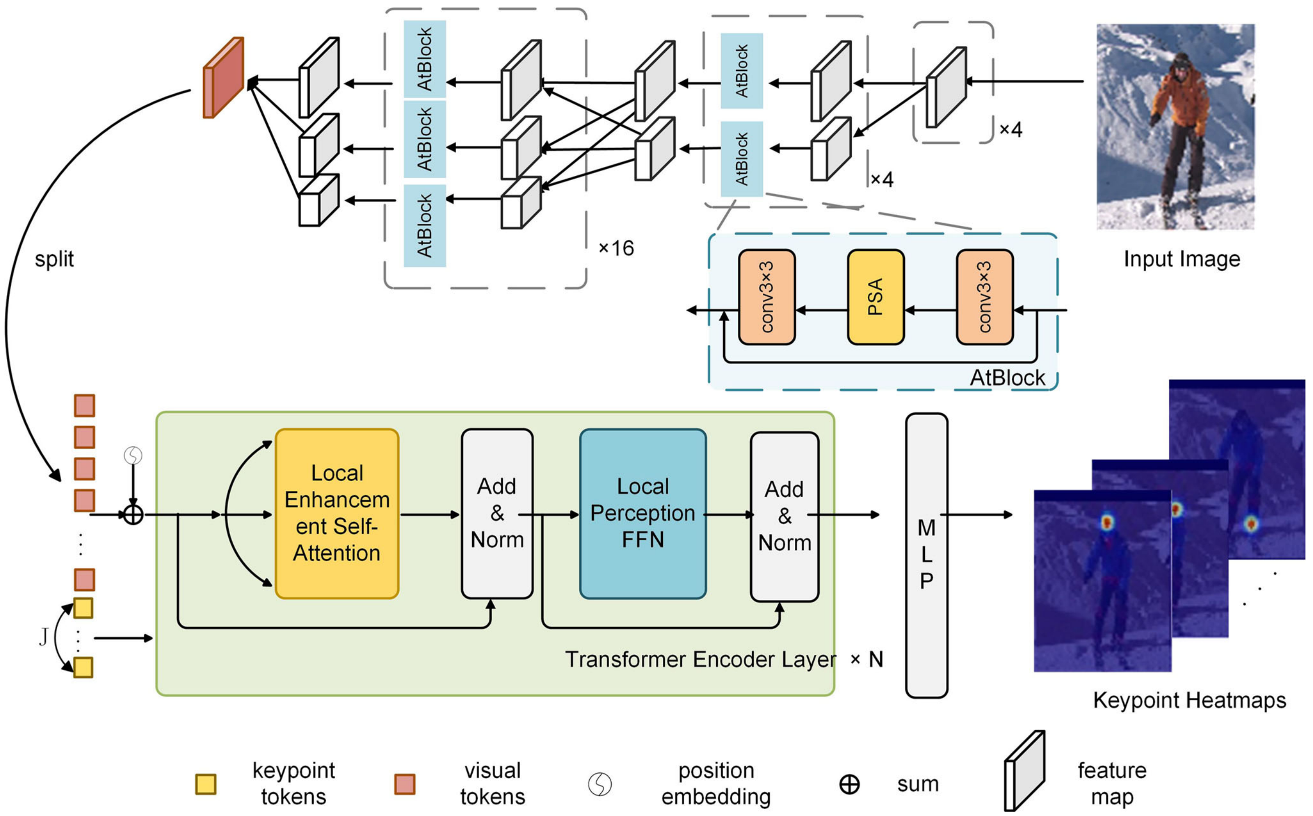

Fig. 3 The pipeline of our proposed CSDNet. We replace BasicBlock with AtBlock in the initial three stages of HRNet for feature extraction. The extracted feature map is then partitioned into equally-sized patches and transformed into a sequence of one-dimensional vectors, referred to as visual tokens. These tokens are fed into Transformer encoding layers, comprising LESA and LPFFN, which effectively capture the global dependencies of tokens while enhancing the interaction with local information. Finally, the output keypoint tokens pass through the MLP head to generate the predicted heatmap of keypoints

matrix of size $H \times W \times C$. In contrast, PSA strikes a balance between memory consumption and high-resolution feature extraction. Specifically, as shown in Fig. 4a, in the channel dimension, it maintains a C/2 channel resolution on one branch while compressing channels on the other branch to generate an attention matrix of size $1 \times 1 \times C$. It maintains a high spatial resolution of $H \times W$ internally, avoiding the complete collapse of spatial information. Similarly, as shown in Fig. 4d, in the spatial dimension, PSA preserves high spatial resolution in one branch while suppressing the spatial dimensions via global pooling operations, and ultimately obtain an attention matrix of size $H \times W \times 1$. That also uses C/2 channel resolution to minimize memory consumption.

PSA preserves high-resolution information in both channel and spatial domains, which facilitates accurate estimation of highly nonlinear pixel-level semantics. Moreover, the computational complexity of PSA is CHW, which is less than or equal to that of other methods. Therefore, we employ PSA and transform the BasicBlock into AtBlock in our work.

We feed the image into the modified HRNet to obtain the feature map. Next, we partition the feature map into a range of flattened 2D patches with a size of 4 × 4 and apply a trainable linear projection to obtain patch embeddings. To preserve positional information, we subsequently add position embeddings to the patch embeddings. Afterwards, feed the obtained embedding vectors into the Transformer layers.

3.2 Local enhancement self-attention

The self-attention layer of Transformers can capture interactions between any pair of positions, providing a significant advantage in representing global interactions. However, due to the fixed size of tokens, capturing local information can be challenging. To address this issue, we leverage depth-wise convolution to aggregate information from nearby positions and enhance local information exchange. Specifically, we utilize a 3 × 3 depth-wise convolution that can aggregate information from 8 adjacent patches in this paper.

The Transformer layer takes a one-dimensional embedding vector of tokens as input. Therefore, it is necessary to process the 2D feature maps extracted by the CNN backbone. Firstly, the features $X \in R^{H \times W \times C}$ are divided into N

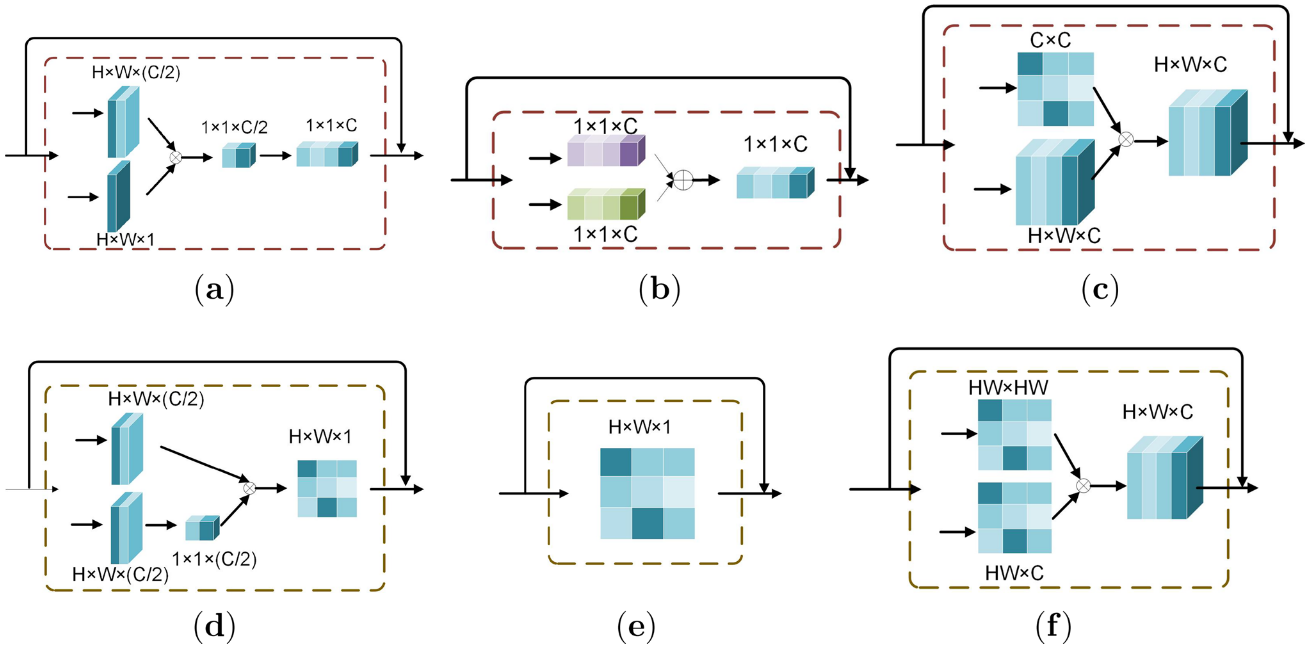

Fig. 4 Polarized Self-Attention (PSA) and two classic mechanisms of the channel attention and spatial attention. **a** The channel attention part of PSA. **b** The first type of channel attention mechanism. **c** The second type of channel attention mechanism. **d** The spatial attention part of PSA. **e** The first type of spatial attention mechanism. **f** The second type of spatial attention mechanism. The figures illustrate how attention maps vary during different attention operations. The final attention maps are all added to the input feature map to obtain an output of size $H \times W \times C$

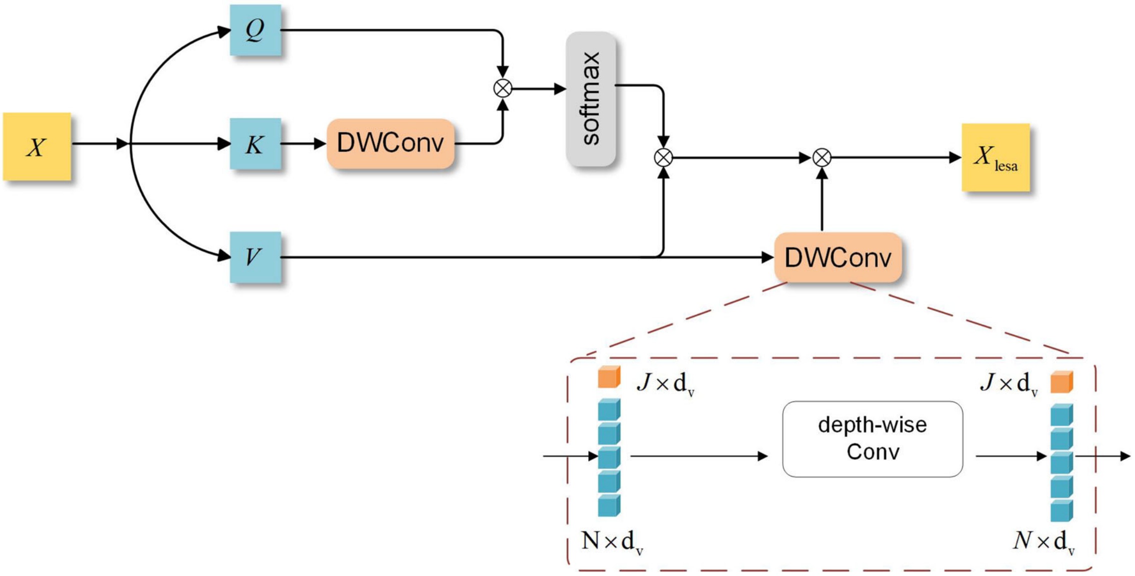

Fig. 5 The structure of the Local Enhancement Self-Attention. To obtain the output, we apply a depth-wise convolution operation on K and multiply it with Q. The result then passes through softmax and is multiplied with V that also undergoes depth-wise convolution

patches of size $P_h \times P_w$, where N is $\frac{H}{P_h} \times \frac{W}{P_w}$. Subsequently, each patch is flattened into a 1D vector of size $P_h \times P_w \times C$ and then mapped to a d-dimensional embedding. Then we obtain visual tokens $X_v \in R^{N \times d}$. We randomly initialize J learnable d-dimensional embedding vectors to represent keypoints, called keypoint tokens $X_j \in R^{J \times d}$. Finally, we add the two to obtain input $X \in R^{(N+J)\times d}$.

Figure 5 depicts the details of the Local Enhancement Self-Attention operation. A linear transformation is first performed on the input $X \in R^{(N+J)\times d}$ to obtain the key $K \in R^{(N+J)\times d_k}$, query $Q \in R^{(N+J)\times d_k}$, and value $V \in R^{(N+J)\times d_v}$, where d, d_k and d_v respectively represent the dimensions of input, key (query) and value. In DWConv, $V^{'} \in R^{N\times d_v}$ passes through a depth-wise convolution, while $V^{''} \in R^{J\times d_v}$ remains unaffected. In the end, $V^{'}$ and $V^{''}$ are concatenated to form V. The same applies to K as well. Afterwards, K passes through the DWConv and then matrix multiplied with Q. Then, a $softmax$ operation is applied, and the resultant output is subjected to matrix multiplication with V to obtain $X^{'}$. Subsequently, V passes through the DWConv and multiplied by $X^{'}$ to obtain $X_{lesa} \in R^{(N+J)\times d}$. These processes can be formulated as:

$$X^{'} = softmax\left(\frac{Q(DWConv(K))^T}{\sqrt{d_k}}\right)V \tag{1}$$

$$X_{lesa} = DWConv(V)X^{'} \tag{2}$$

3.3 Local perception feed-forward network

The feed-forward network module performs dimensional expansion/reduction and non-linear transformation on tokens to improve the nonlinear fitting ability of the model. However, the interaction of spatially contiguous information among tokens is not considered. Complementary to the LESA module, the FFN module can additionally extract local information based on the ability of Transformers to establish long range dependencies. Therefore, we propose a local perception feed-forward network (LPFFN). In each encoder block, after the LESA module, the original feed-forward network layer is replaced with the LPFFN.

As illustrated in Fig. 6, the LPFFN includes the following steps. Firstly, the $X_{lesa} \in R^{(N+J)\times d}$ obtained from the LESA is divided into two parts: keypoint tokens $X_j \in R^{J\times d}$ and visual tokens $X_v \in R^{N\times d}$, where J is the number of keypoints and N is the number of patches. Subsequently, X_v is projected onto a higher dimensional tensor $X_v^{'} \in R^{N\times(r\times d)}$ using a linear projection, where r represents the expansion ratio. Following this, $X_v^{'}$ is restored back to an $X_{sp} \in R^{\frac{H}{P_h}\times\frac{W}{P_w}\times(r\times d)}$ in the spatial dimension, based on the position of the original image. Furthermore, a depth-wise convolution with kernel size of k is performed on X_{sp} to obtain X_{dw}, which enhances its correlation with the $k^2 - 1$ tokens in the surrounding area. Afterwards, the two-dimensional spatial dimension is flattened into a one-dimensional sequence, and then linearly projected back to the initial dimension to obtain the $X_p^{''} \in R^{N\times d}$. Finally, it is added to the input X_{lesa} to obtain $X_{ffn} \in R^{(N+J)\times d}$. The application of residual connections is motivated by the classical residual networks to enhance the propagation ability of gradients across the layers. These processes can be formulated as:

$$X_j, X_v = Split(X_{lesa}), \tag{3}$$

$$X_{sp} = Reshape(GELU(BN(Conv(X_v)))), \tag{4}$$

$$X_{dw} = GELU(BN(dwConv(X_{sp}))), \tag{5}$$

$$X_v^{''} = GELU(BN(Conv(Flatten(X_{dw})))), \tag{6}$$

$$X_{ffn} = Concat\left(X_j, X_v^{''}\right) + X_{lesa}. \tag{7}$$

3.4 Loss function and heatmap estimation

Our model utilizes the MSE loss function to compare the gap between the predicted heatmap and the groundtruth heatmap. MSE is formulated as:

$$L_{MSE} = \frac{1}{J}\sum_{j=1}^{J}\left\|Y_j - \widehat{Y}_j\right\|_2, \tag{8}$$

where Y_j and $\widehat{Y}_j$ represent the groundtruth and predicted heatmap respectively, and J represents the number of keypoints. We use the MLP head to obtain final predicted heatmaps for different keypoints. We employ the last Transformer layer as an aggregator, computing the maximum activation of keypoints in the heatmap to be predicted. It uses the contribution scores gathered from various locations in the image. To create a $h \times w$-sized 2D heatmap, the d-dimensional keypoint tokens output by the Transformer are linearly projected onto the corresponding $h \times w$-dimensional tensor. The resulting one-dimensional tensor is then reshaped into a two-dimensional predicted heatmap.

4 Experiments

4.1 Datasets and evaluation metrics

- **Datasets**
  COCO The COCO2017 [56] dataset contains 200k outdoor images in the wild and 250k person instances. The pose annotations in the dataset are based on 17 keypoints. The train2017 set consists of 57k images and 150k person instances. The val2017 set contains 5k images, and the test2017 set contains 20k images. All the experiments

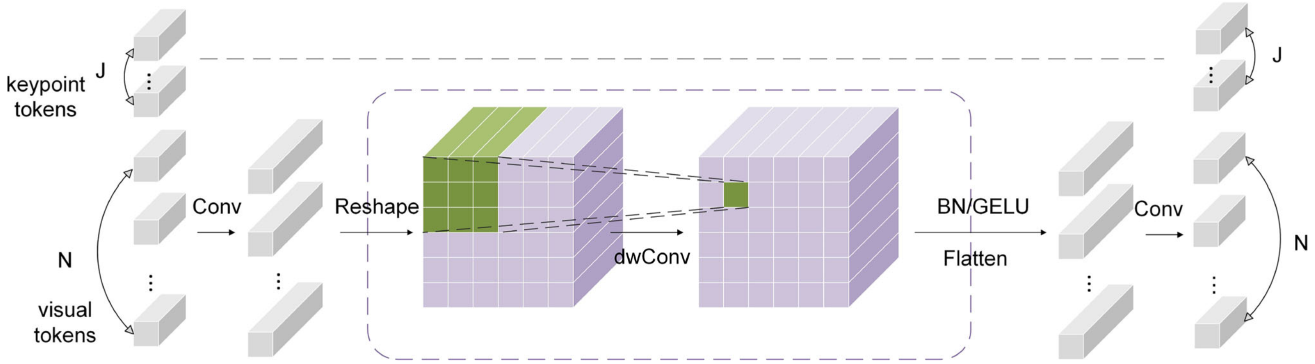

Fig. 6 The structure of the Local Perception Feed-Forward Network. The initial input tokens are divided into keypoint tokens and visual tokens. The visual tokens are then convolved to higher dimensions, followed by reshaping to a two-dimensional tokens. After undergoing depth-wise convolution, tokens are flattened and then restored to its original dimension. Subsequently, These vector are concatenated with the keypoint tokens

reported in this paper are conducted on the train2017 set and evaluated on the val2017 dataset. In Sect. 4.3.1, we present the results of our experiments on the COCO dataset.

MPII The MPII [57] dataset contains approximately 25k images with over 40k human targets with annotated human keypoint information. Each human target in the MPII dataset is annotated with 16 keypoints. In Sect. 4.3.2, we present the results of our experiments on the MPII dataset.

- **Evaluation metrics**

  OKS We employ the standard Average Precision (AP) as an evaluation metric for the experimental results on the COCO dataset. AP is based on Object Keypoint Similarity (OKS):

$$OKS = \frac{\sum_i \exp\left(-\hat{d}_i^2/2s^2k_i^2\right)\delta\left(v_i > 0\right)}{\sum_i \delta\left(v_i > 0\right)}, \tag{9}$$

where $\hat{d}_i$ is the Euclidean distance between i-th predicted keypoint coordinate and the corresponding ground truth, v_i is the visibility flag of the groundtruth, s is the object scale, $\delta()$ is the indicator function, and k_i is a per-keypoint constant that controls falloff. AP (the mean of AP scores at 10 OKS positions, 0.50, 0.55,…, 0.90, 0.95), AP^{50} (AP at OKS = 0.50); AP^{M} and AP^{L} for medium and large objects respectively, and AR (the mean of AR scores at 10 OKS positions, 0.50, 0.55,…, 0.90, 0.95).

PCKh For the experimental results on the MPII dataset, we evaluate them using the head-normalized Percentage of Correct Keypoints (PCKh) [57]:

$$PCKh_i = \frac{\sum_p \delta\left(d_{p^i} \le T \cdot d_p^{head}\right)}{\sum_p 1}, \tag{10}$$

$$PCKh_{mean} = \frac{\sum_i PCKh_i}{\sum_i 1}, \tag{11}$$

where $PCKh_i$ is the PCK value of the predicted results for the i-th keypoint, d_{p^i} is the Euclidean distance between the predicted value of the i-th keypoint of the p-th person and groundtruth, d_p^{head} is the Euclidean distance between the upper left point and the lower right point of the header rectangle of the p-the person, T is a threshold, and $\delta()$ is the indicator function. In our experiment, T is set to 0.5, which we denote as PCKh@0.5 and use as the evaluation metric to compare with other methods.

4.2 Implementation details and Model architecture configurations

- **Implementation details**

  In our experiments, we adopt the top-down human pose estimation paradigm. The training samples, consisting of cropped images with a single person, are initially detected by a person detector, and then the keypoints are predicted. To carry out this task, we utilize the popular person detector provided by Simple Baseline. Prior to training on COCO2017 dataset, we resize input images to either 256 × 192 resolution. Conversely, when working with MPII dataset, we resize input images to 256 × 256 resolution during training. To minimize quantization errors while decoding from a downscaled heatmap, we employ the coordinate decoding method proposed by Zhang Feng [58]. The Adam optimizer is used to optimize the model and we follow the learning schedule outlined in [14]. Our research adopts the TokenPose setting. We set the base learning rate to 1e-3, which is reduced to 1e-4 and 1e-5 at stage 200 and 260, respectively. We conduct a total of 300 epochs during the training process, as the structure of the Transformer tends to rely on a longer training time to converge. We provide the GFLOPs of the models to compare their complexity. The GFLOPs calculation in this paper includes convolutional layers and linear layers, with the formulas as follows:

$$1GFLOPs = 10^9 FLOPs, \tag{12}$$

$$\begin{aligned} FLOPs &= k \times k \times C_{in} \times C_{out} \times S_{out} \\ &+ S_{in} \times S_{out} \times C, \end{aligned} \tag{13}$$

where C_{in}, C_{out} and C are the number of input and output channels, k is the convolution kernel size, S_{in} and S_{out} represent the size of input and output features, respectively.

- **Model architecture configurations**
  We apply a hybrid variant based on CNNs and Transformers. In this experiment, we use HRNet [14] with various depths as backbone. The architecture configuration is shown in Table 1. CSDNet-s, CSDNet-m, and CSDNet-l use stem-net, HRNet-W32 and HRNet-W48 as backbone, respectively. The stem-net [37, 59] includes several simple convolutions and is often used for quickly downsampling to a quarter of the input resolution.

4.3 Quantitative analysis

4.3.1 Results on COCO dataset

We compare with several methods based on CNNs and Transformers on the COCO dataset. The model configurations are presented in Table 1. The experimental results are summarized in Table 2. The result of CSDNet-m/12 shows an AP of 75.0%. Notably, our AP result is comparable to HRNet [14], but with significantly fewer parameters and GFLOPs. Similarly, both CSDNet-m/12 and UniFormer-B [55] reach an AP of 75.0%, but CSDNet-m/12 requires fewer parameters and GFLOPs compared to UniFormer-B [55]. CSDNet-l achieves an experimental result of 75.8%, while using about 3 million fewer parameters compared to TokenPose-L/D24 [54]. The results of CSDNet-l and ViTPose [50] are both 75.8% but CSDNet-l have fewer parameters than ViTPose, only about one-third. Compared to TransPose-H-A6 [51], the results are comparable. Although there are many parameters, GFLOPs is still less than half of it. Furthermore, our approach is slightly better than HRFormer-B [49] in terms of results, while utilizing fewer parameters and GFLOPs. These findings reflect the effectiveness of our proposed improvements, which are not inferior, and in some cases even superior to other CNN+Transformer-based architectures. It indicates that the internal structure of CNNs and Transformers is indeed worth further analysis and exploration, and there is still room for improvement.

4.3.2 Results on MPII dataset

We adhere to the experimental setup and testing procedure of TokenPose L/D6, as described in TokenPose [54]. Specifically, we employ HRNet-W48 as the backbone and add 6 Transformer layers, while keeping the experimental settings such as learning rate and epochs unchanged. All experiments are conducted using an input image size of 256 × 256. In Table 3, we compare our experimental results with those of other experiments. Additionally, we provide the PCK for different keypoints, namely: Head, Shoulder, Elbow, Wrist, Hip, Knee, and Ankle. The amount of parameters in our experiment is 25.3M, 1.8M more than TokenPose-L/D12 and 2.8M less than TokenPose-L/D24. Our method achieves an overall accuracy improvement of 0.5 percentage points compared to D6, and achieves better performance compared to D12 and D24. Notably, our approach also exhibits substantial improvements in detecting various body parts, including elbow, wrists and ankles. Our initial results on the MPII dataset demonstrate that training a Transformer-based model on large-scale pose correlation data would be effective for robust representation of pose estimation.

4.4 Ablation study

We explore the internal structures of CNNs and Transformers to further improve the model performance. To evaluate the effectiveness of our proposed modules, we conduct ablation experiments. We also study the individual modules separately.

Effectiveness of Components To evaluate the effectiveness of our proposed modules, we conduct ablation experiments on PSA, LESA, and LPFFN. Specifically, we conduct experiments using CSDNet-m/6, which incorporates 6 Transformer layers. The model with no modifications or added modules serves as the baseline. By adding or replacing our modules, we assess their impact on the overall model performance. As illustrated in Table 4, we make some modifications to the model architecture, specifically replacing BasicBlock with AtBlock, the FFN module in Transformer layers with LPFFN module, and the MHA module in Transformer layers with LESA module. The first three rows in the table represent the model results using AtBlock, LPFFN, and LESA modules, respectively. It can be observed that the individual usage of AtBlock and LPFFN increases by 0.2 percentage points compared to the baseline. It is evident that the performance of the AtBlock and LPFFN modules is relatively better than LESA. Subsequently, pairwise combinations are conducted. It can be observed that the combination of AtBlock and LPFFN performs slightly better than the combination with LESA individually. Finally, the result using all modules reaches 90.7%, surpassing the baseline by 0.5 percentage points. Moreover, our optimized model achieves superior detection

Table 1 Model architecture configurations

Model	CNN backbone	Layers	Heads	Patch size
CSDNet-s	Stem-net	12	8	4 × 3
CSDNet-m/6	HRNet-W32-stage3	6	8	4 × 3
CSDNet-m/12	HRNet-W32-stage3	12	8	4 × 3
CSDNet-l	HRNet-W48-stage3	6	8	4 × 3

'Layers' represents the number of Transformer layers in the model

Table 2 Comparison with other advanced top-down 2D pose estimation approaches on the COCO validation set

Method	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
SimpleBaseline-Res50 [40]	34.0M	8.9	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline-Res101 [40]	53.0M	12.4	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline-Res152 [40]	68.6M	15.7	72.0	89.3	79.8	68.7	78.9	77.8
HRNet-W32 [14]	28.5M	7.1	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W48 [14]	63.6M	14.6	75.1	90.6	82.2	71.5	81.8	80.4
TokenPose-B [54]	13.5M	5.7	74.7	89.8	81.4	71.3	81.4	80.0
TokenPose-L/D6 [54]	20.8M	9.1	75.4	90.0	81.8	71.8	82.4	80.4
TokenPose-L/D24 [54]	27.5M	11.0	**75.8**	90.3	82.5	72.3	**82.7**	**80.9**
TransPose-H-S [51]	8 M	10.2	74.2	–	–	–	–	78.0
TransPose-H-A4 [51]	17.3M	17.5	75.3	–	–	–	–	80.3
TransPose-H-A6 [51]	17.5M	21.8	**75.8**	–	–	–	–	80.8
ViTPose-B [50]	86 M		**75.8**	–	–	–	–	81.1
HRFormer-S [49]	7.8M	2.8	74.0	90.2	81.2	70.4	80.7	79.4
HRFormer-B [49]	43.8M	14.1	75.6	**90.8**	82.8	71.7	82.6	80.8
UniFormer-S [55]	25.2M	4.7	74.0	90.3	82.2	66.8	76.7	79.5
UniFormer-B [55]	53.5M	9.2	75.0	90.6	**83.0**	67.8	77.7	80.4
CSDNet-m/6	11.4M	5.3	74.5	89.6	81.5	71.2	81.6	79.9
CSDNet-m/12	17.4M	6.9	75.0	89.9	81.7	71.4	81.9	80.1
CSDNet-l	24.7M	10.0	**75.8**	90.1	82.5	**72.4**	82.5	**80.9**

The input size is 256 × 192

Table 3 Comparison with other advanced top-down 2D pose estimation approaches on the MPII validation set (PCKh@0.5)

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Mean	#Params
SimpleBaseline-Res50 [40]	96.4	95.3	89.0	83.2	88.4	84.0	79.6	88.5	34.0M
SimpleBaseline-Res101 [40]	96.9	95.9	89.5	84.4	88.4	84.5	80.7	89.1	53.0M
SimpleBaseline-Res152 [40]	97.0	95.9	90.0	85.0	89.2	85.3	81.3	89.6	68.6M
HRNet-W32 [14]	96.9	96.0	90.6	85.8	88.7	86.6	82.6	90.1	28.5M
TokenPose-L/D6 [54]	97.1	95.9	91.0	85.8	89.5	86.1	82.7	90.2	21.4M
TokenPose-L/D12 [54]	**97.2**	95.8	90.7	85.9	89.2	86.2	82.3	90.1	23.5M
TokenPose-L/D24 [54]	97.1	95.9	90.4	86.0	89.3	**87.1**	82.5	90.2	28.1M
CSDNet-l	97.1	**96.1**	**91.2**	**86.8**	**89.5**	87.0	**83.1**	**90.7**	25.3M

The input size is 256 × 256

accuracy at multiple keypoints, with significant improvements observed at the elbow, wrist and ankle joints.

The position of PSA in the AtBlock We discuss the optimal location of adding PSA modules and conduct ablation experiments. We utilize TokenPose-L/D6 [54] as the baseline. Specifically, we add PSA to the first and second 3 × 3 convolution respectively, and evaluate the experimental results. As indicated in Table 5, adding PSA to the first 3 × 3 convolution yields significantly better results. However, adding PSA to the second 3 × 3 convolution results in worse performance than the original experiment. These results indicate that the location of attention mechanisms usage also greatly affects the performance of them.

Table 4 The ablation study of three proposed modules on the MPII dataset

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Mean
BasicBlock(baseline)	96.9	96.1	90.9	85.5	89.4	86.9	82.5	90.2
AtBlock	**97.3**	**96.1**	91.1	86.6	89.4	86.8	82.6	90.4
LPFFN	97.1	95.9	90.9	86.9	89.1	86.8	82.7	90.4
LESA	97.1	95.8	90.7	85.8	89.2	86.4	82.8	90.3
AtBlock+LPFFN	97.1	96.0	90.9	86.2	**89.6**	**87.0**	82.6	90.6
AtBlock+LESA	97.0	95.9	91.1	87.0	89.1	86.9	82.9	90.5
LPFFN+LESA	90.0	95.8	90.9	86.8	89.2	86.8	82.9	90.5
AtBlock+LPFFN+LESA	97.1	**96.1**	**91.2**	**86.8**	89.5	**87.0**	**83.1**	**90.7**

The first column represents the utilized modules

Table 5 The ablation experiment on the placement position of PSA

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Mean
Baseline	97.1	95.9	91.0	85.8	89.5	86.1	82.7	90.2
After 1st	97.3	96.1	91.1	86.6	89.4	86.8	82.6	90.4(+0.2)
After 2nd	96.8	95.6	90.3	84.8	89.4	84.7	81.6	90.2(+0)

Table 6 Ablation study results on the type of LPFFN

Kernel Size	Mean	#Params
×	72.5	6.6M
1 × 1	71.5(−1.0)	7.54M
3 × 3	73.5(+1.0)	7.61M
5 × 5	73.2(+0.7)	7.76M

The type of LPFFN For depth-wise convolution, the kernel size determines the size of the region in which local correlation is established. Therefore, we compare the effect of different kernels on the experiment. We employ TokenPose-Small-v1 [54] as the baseline, which does not include any depth-wise convolutions. Both kernel sizes of 3 × 3 and 5 × 5 yield good results. Based on the trade-off between number of parameters and accuracy, we choose the depth-wise convolution with the kernel of 3 × 3. As shown in Table 6, when the kernel size is 3 × 3, the experimental result is the best, and the parameter consumption is not significant.

4.5 Qualitative analysis

Examples of the final predicted heatmap of 17 keypoints in the COCO dataset are visualized in Fig. 7. We use the keypoint tokens output from the last Transformer layer and process them through the MLP head to predict the keypoint heatmaps. The MLP output generates 17 channels, each corresponding to one of the keypoints. As illustrated in Fig. 7, each channel is dedicated to a fixed keypoint. If a keypoint exists, the corresponding heatmap estimation is outputted. Conversely, if a key point does not exist, the corresponding channel output is set to none.

As illustrated in Fig. 8, we compare the estimation of our method with HRNet for complex poses. It is obvious that our method outperforms HRNet in accurately estimating occluded or undetected limb parts in complex poses. According to our analysis, it is because our method takes into account both local and global information. By considering nearby keypoints in conjunction with distant keypoints from a global perspective, our method is able to effectively analyze and localize keypoints that are challenging to discernible or occluded in complex poses. This comprehensive approach of considering both local and global information contributes to its promising performance in estimating complex poses. In the first and third columns of the pose, the arms are obstructed, and local information alone cannot effectively connect the wrist to the human body over long distances in HRNet. In contrast, our approach leverages global information, enabling more accurate association of keypoints that are farther away from the human body with the body structure.

As demonstrated in Fig. 9, we conduct a comparative analysis between our proposed method and HRNet for estimating limbs that are challenging to recognize under poor lighting conditions. It can be seen that in such conditions, keypoints and limbs of the human body are difficult to discern and even blend with the environment. However, our method effectively addresses this challenge by preserving high-resolution information and mitigating the loss of information due to reduced resolution. By retaining detailed information of blurred objects and lower noise levels, our method facilitates accurate identification of blurred objects. In particular, high-resolution images of blurred objects may reveal more details at the pixel level, enabling us to better capture the shape, contour, and texture features of the objects. In the final column of Fig. 9, under similar low-light conditions,

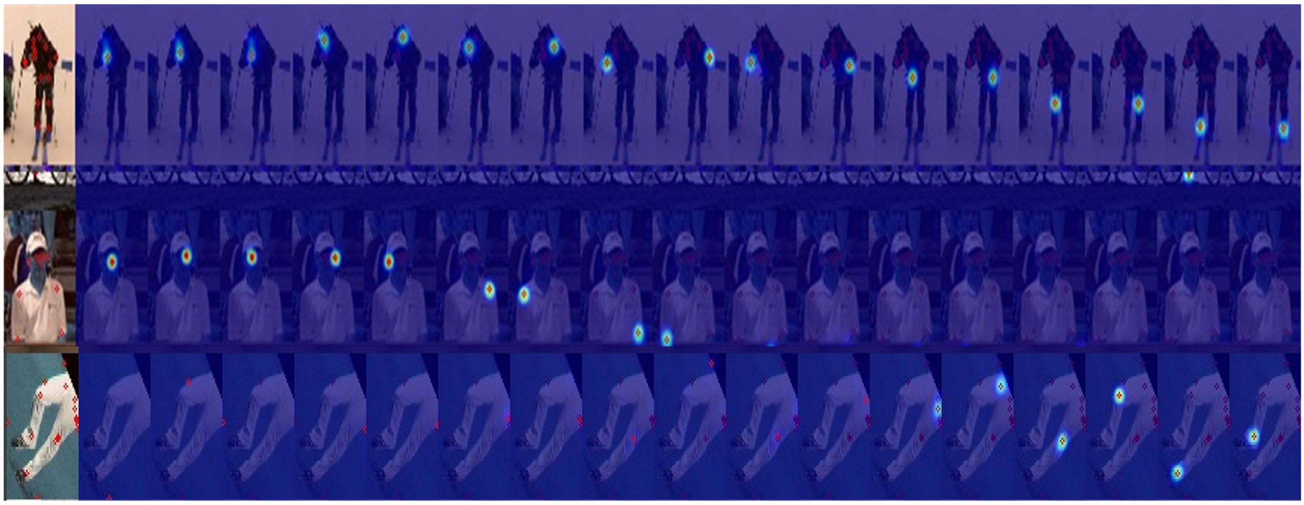

Fig. 7 Examples of heatmap visualization of 17 human keypoints on the COCO dataset. The keypoint tokens output from the Transformer layers are processed through the Multilayer Perceptron (MLP) module to obtain 17 channels, corresponding to 17 keypoints

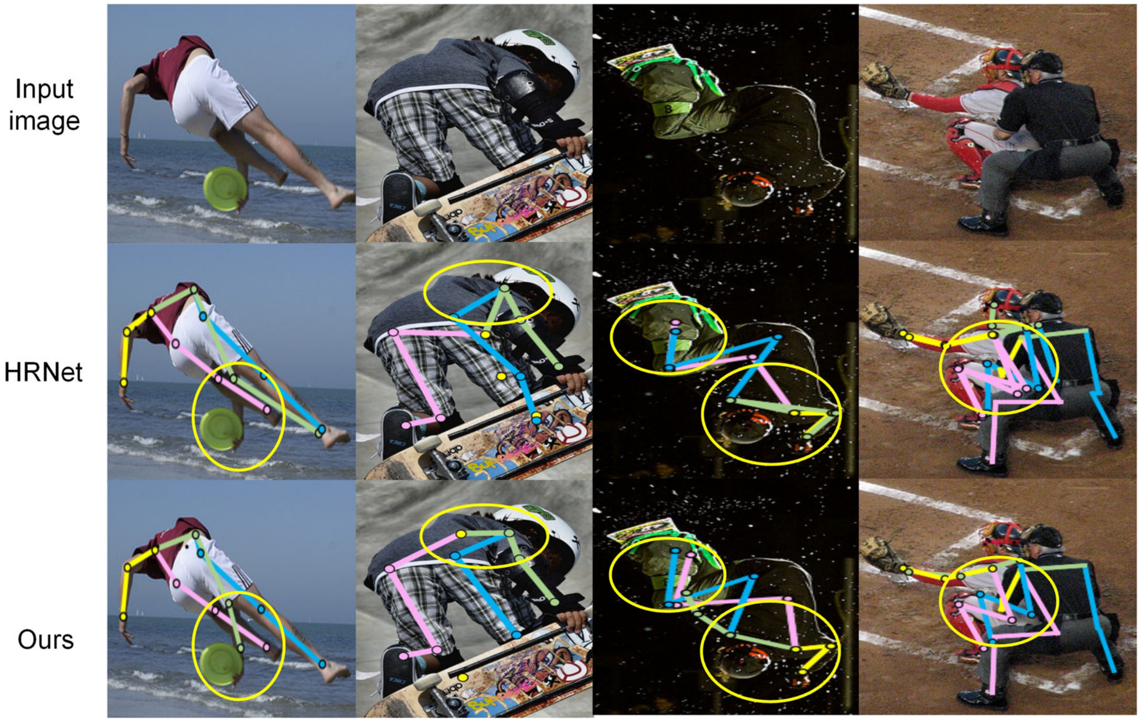

Fig. 8 Comparison of our method with HRNet method in terms of complex poses

HRNet exhibits errors in connecting keypoints and fails to recognize the correct arm posture. In contrast, our method successfully identifies the accurate arm posture. As a result, our method exhibits certain advantages in recognizing limbs that are blurry due to poor lighting conditions.

Figure 10 illustrates the comparison between our proposed method and HRNet on some examples of large-scale human pose estimation. It is evident that our method achieves superior performance in estimating the posture of the entire body at larger scales. We leverage the LESA module to capture long-range dependencies, enabling us to better recognize the complete posture and eliminate erroneous poses. Additionally, as shown in the last column, we enhance the connectivity of local information, resulting in better distinguishing poses that have similar content in certain regions.

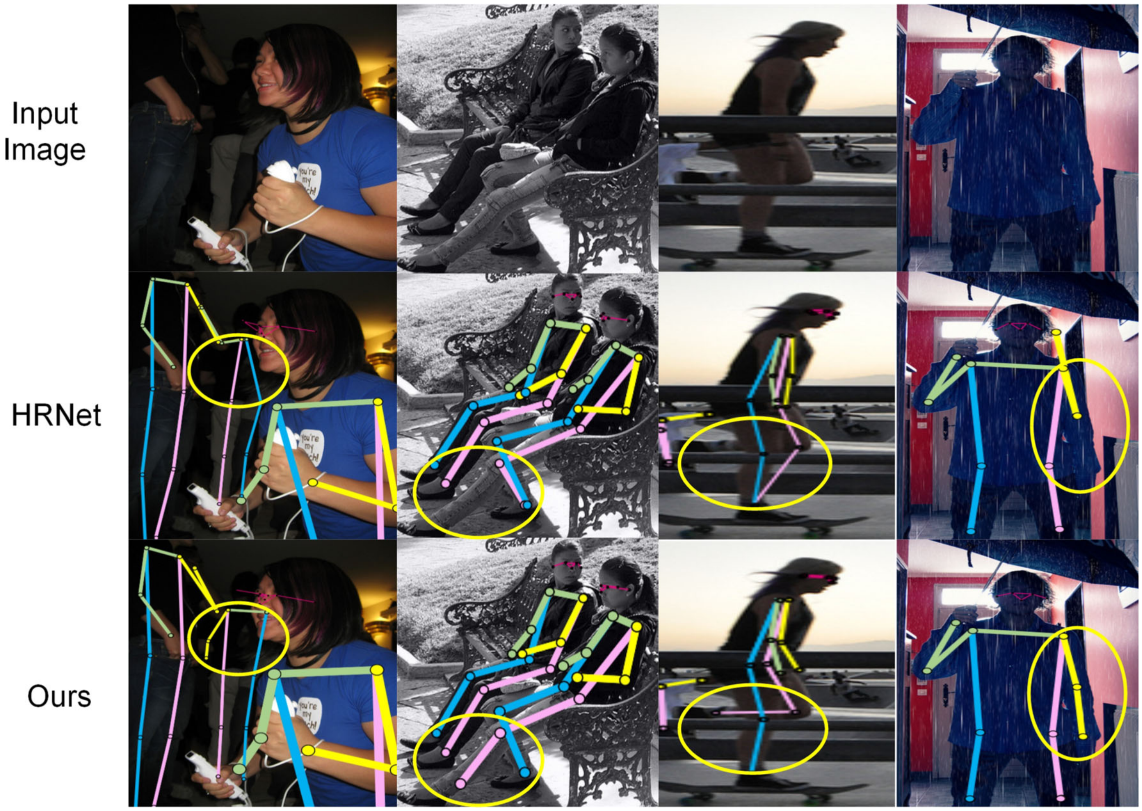

Fig. 9 Comparison of our method with HRNet method for pose estimation under adverse lighting conditions

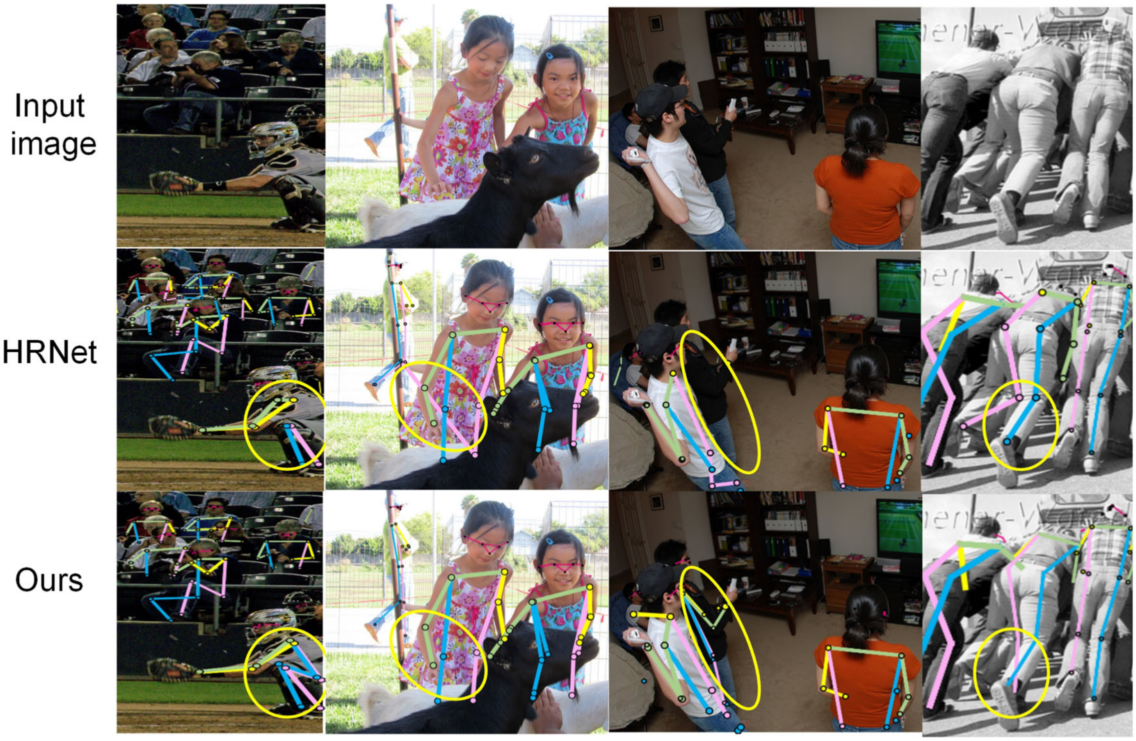

Fig. 10 Comparison of our method with HRNet method in terms of large scale human body

5 Conclusion

In this paper, we employ a hybrid architecture that synergistically leverages the strengths of CNNs and Transformers to capture both local and global information for enhanced feature extraction in human pose estimation. To alleviate the issue of reduced resolution in CNNs, we propose the AtBlock to preserve internal high resolution, thereby enabling more precise pixel-level regression for the estimation of human details. Furthermore, we replace the Multi-Head Attention (MHA) layers and the feed-forward network layers in the Transformer with local enhancement self-attention (LESA) and local perception feed-forward network (LPFFN), compensating for the Transformer's limitations in extracting local information in 2D space and alleviating the limitations of utilizing Transformers in the field of computer vision. Our modifications enable more accurate estimation of poses in complex environments. Overall, our approach enhances the network's feature extraction ability, paving the way for further advances in human pose estimation to some extent.

Funding Partial financial support was received from the Natural Science Foundation of Hebei Province (F2019201451).

Data availability The data used in this study were sourced from publicly accessible websites (https://cocodataset.org/, http://human-pose.mpi-inf.mpg.de/). Additional data can be obtained from the corresponding authors; please contact qingxuanshi@hbu.edu.cn for further information.

Declarations

Conflict of interest All authors have no Conflict of interest to declare that are relevant to the content of this article.

Ethical approval Ethical approval is not applicable to this article, as it does not involve research with humans or animals.

References

1. Li, B., Dai, Y., Cheng, X., Chen, H., Lin, Y., He, M.: Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN, pp. 601–604 (2017)
2. Li, B., He, M., Dai, Y., Cheng, X., Chen, Y.: 3D skeleton based action recognition by video-domain translation-scale invariant mapping and multi-scale dilated CNN. Multimedia Tools Appl. **77**, 22901–22921 (2018)
3. Luvizon, D.C., Picard, D., Tabia, H.: 2D/3D pose estimation and action recognition using multitask deep learning, pp. 5137–5146 (2018)
4. Li, B., Chen, H., Chen, Y., Dai, Y., He, M.: Skeleton boxes: solving skeleton based action detection with a single deep convolutional neural network, pp. 613–616 (2017)
5. Zhou, T., Wang, W., Qi, S., Ling, H., Shen, J.: Cascaded human-object interaction recognition, pp. 4263–4272 (2020)
6. Wan, B., Zhou, D., Liu, Y., Li, R., He, X.: Pose-aware multi-level feature network for human object interaction detection, pp. 9469–9478 (2019)
7. Yang, J., Zhang, J., Yu, F., Jiang, X., Zhang, M., Sun, X., Chen, Y.-C., Zheng, W.-S.: Learning to know where to see: a visibility-aware approach for occluded person re-identification, pp. 11885–11894 (2021)
8. Chen, H., Lagadec, B., Bremond, F.: Ice: Inter-instance contrastive encoding for unsupervised person re-identification, pp. 14960–14969 (2021)
9. Luo, Z., Wang, Z., Huang, Y., Wang, L., Tan, T., Zhou, E.: Rethinking the heatmap regression for bottom-up human pose estimation, pp. 13264–13273 (2021)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks, pp. 7132–7141 (2018)
11. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module, pp. 3–19 (2018)
12. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation, pp. 3146–3154 (2019)
13. Liu, H., Liu, F., Fan, X., Huang, D.: Polarized self-attention: Towards high-quality pixel-wise regression (2021). arXiv preprint arXiv:2107.00782
14. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation, pp. 5693–5703 (2019)
15. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
16. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. **3**(3), 201–215 (2002)
17. Rensink, R.A.: The dynamic representation of scenes. Vis. Cogn. **7**(1–3), 17–42 (2000)
18. Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., Huang, G.: On the integration of self-attention and convolution, pp. 815–825 (2022)
19. Zhu, L., Wang, X., Ke, Z., Zhang, W., Lau, R.W.: Biformer: Vision transformer with bi-level routing attention, pp. 10323–10333 (2023)
20. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018). arXiv preprint arXiv:1810.04805
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł, Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
22. Li, W., Chen, H., Guo, J., Zhang, Z., Wang, Y.: Brain-inspired multilayer perceptron with spiking neurons, pp. 783–793 (2022)
23. Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., Shen, C.: Conditional positional encodings for vision transformers (2021). arXiv preprint arXiv:2102.10882
24. Guo, J., Tang, Y., Han, K., Chen, X., Wu, H., Xu, C., Xu, C., Wang, Y.: Hire-MLP: Vision MLP via hierarchical rearrangement, pp. 826–836 (2022)
25. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: transformers for image recognition at scale (2020). arXiv preprint arXiv:2010.11929
26. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention, pp. 10347–10357 (2021)
27. Li, Y., Zhang, K., Cao, J., Timofte, R., Van Gool, L.: Localvit: Bringing locality to vision transformers (2021). arXiv preprint arXiv:2104.05707
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows, pp. 10012–10022 (2021)
29. Liu, H., Jiang, X., Li, X., Bao, Z., Jiang, D., Ren, B.: Nommer: Nominate synergistic context in vision transformer for visual recognition, pp. 12073–12082 (2022)

30. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention, pp. 4794–4803 (2022)
31. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: CVT: introducing convolutions to vision transformers, pp. 22–31 (2021)
32. Lin, W., Wu, Z., Chen, J., Huang, J., Jin, L.: Scale-aware modulation meet transformer, pp. 6015–6026 (2023)
33. Tang, W., Yu, P., Wu, Y.: Deeply learned compositional models for human pose estimation, pp. 190–206 (2018)
34. Gkioxari, G., Toshev, A., Jaitly, N.: Chained predictions using convolutional neural networks, pp. 728–743 (2016)
35. Lifshitz, I., Fetaya, E., Ullman, S.: Human pose estimation using deep consensus voting, pp. 246–260 (2016)
36. Cai, Y., Wang, Z., Luo, Z., Yin, B., Du, A., Wang, H., Zhang, X., Zhou, X., Zhou, E., Sun, J.: Learning delicate local representations for multi-person pose estimation, pp. 455–472 (2020)
37. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation, pp. 5386–5395 (2020)
38. Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos, pp. 1913–1921 (2015)
39. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation, pp. 1831–1840 (2017)
40. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking, pp. 466–481 (2018)
41. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation, pp. 483–499 (2016)
42. Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines, pp. 4724–4732 (2016)
43. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation, pp. 1281–1290 (2017)
44. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation, pp. 7103–7112 (2018)
45. Ramakrishna, V., Munoz, D., Hebert, M., Andrew Bagnell, J., Sheikh, Y.: Pose machines: Articulated pose estimation via inference machines, pp. 33–47 (2014)
46. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in neural information processing systems, vol. 27 (2014)
47. Papandreou, G., Zhu, T., Chen, L.-C., Gidaris, S., Tompson, J., Murphy, K.: Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model, pp. 269–286 (2018)
48. Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., Tu, Z.: Pose recognition with cascade transformers, pp. 1944–1953 (2021)
49. Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: Hrformer: High-resolution transformer for dense prediction (2021). arXiv preprint arXiv:2110.09408
50. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Adv. Neural. Inf. Process. Syst. **35**, 38571–38584 (2022)
51. Yang, S., Quan, Z., Nie, M., Yang, W.: Transpose: Keypoint localization via transformer, pp. 11802–11812 (2021)
52. Huang, L., Tan, J., Liu, J., Yuan, J.: Hand-transformer: nonautoregressive structured modeling for 3D hand pose estimation, pp. 17–33 (2020)
53. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers, pp. 1954–1963 (2021)
54. Li, Y., Zhang, S., Wang, Z., Yang, S., Yang, W., Xia, S.-T., Zhou, E.: Tokenpose: Learning keypoint tokens for human pose estimation, pp. 11313–11322 (2021)
55. Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unifying convolution and self-attention for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
56. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: common objects in context, pp. 740–755 (2014)
57. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis, pp. 3686–3693 (2014)
58. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation, pp. 7093–7102 (2020)
59. Stoffl, L., Vidal, M., Mathis, A.: End-to-end trainable multi-instance pose estimation with transformers (2021). arXiv preprint arXiv:2103.12115

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.