Lightweight Context-aware Feature Transformer Network for Human Pose Estimation

Yanli Ma , Qingxuan Shi * , Fan Zhang

Posted Date: 11 January 2024

doi: 10.20944/preprints202401.0836.v1

Keywords: Human pose estimation; Expressive power of features; Feature refinement; Global dependencies

Article

Lightweight Context-Aware Feature Transformer Network for Human Pose Estimation

Yanli Ma [1,†,‡], Qingxuan Shi [1,‡] and Fan Zhang [‡]

[1] Hebei Machine Vision Engineering Research Center, Hebei University, Baoding, 071002, China; qingxuanshi@hbu.edu.cn

* Correspondence: qingxuanshi@hbu.edu.cn

† Current address: Hebei Machine Vision Engineering Research Center, Hebei University, Baoding, 071002, China.

‡ These authors contributed equally to this work.

Abstract: We propose Context-aware Feature Transformer Network (CaFTNet), a novel network for human pose estimation. To address the issue of limited modeling of global dependencies in convolutional neural networks, we design Transformerneck to strengthen the expressive power of features. Transformerneck directly substitutes the 3×3 convolution in bottleneck of HRNet with Contextual Transformer (CoT) block, while reducing the complexity of the network. Specifically, CoT first produces keys with static contextual information through 3×3 convolution. Then, relying on the query and contextualization keys, the dynamic contexts are generated through two concatenated 1×1 convolutions. Static and dynamic contexts are eventually fused as an output. Additionally, for the multi-scale networks, in order to further refine the features of the fusion output, we propose an Attention Feature Aggregation Module(AFAM). Technically, given an intermediate input, AFAM successively deduces attention maps along channel and spatial dimensions. Then, Adaptive refinement module(ARM) is exploited to activate the obtained attention maps. Finally, the input undergoes adaptive feature refinement through multiplication with the activated attention maps. Through the above studies, our lightweight network provides a powerful clue for detection of keypoints. Experiments are implemented on the COCO and MPII datasets. The model achieves 76.2 AP on the COCO val2017. Compared to other methods with the CNN as the backbone, CaFTNet reduces the number of parameters by 72.9 %. On the MPII, our method uses only 60.7% of the number of parameters, acquiring semblable results to other methods with the CNN as the backbone.

Keywords: human pose estimation; expressive power of features; feature refinement; global dependencies

1. Introduction

The goal of human pose estimation is to predict keypoints of human anatomy in the images. It has extensive applications in the field of computer vision for instance human action recognition [1–4], human pose tracking [5–9], 3D human pose estimation [10–13] and so on.

CNNs have obtained praiseworthy accomplishments in human pose estimation [14–18]during a recent period of time. However, the convolution's receptive field is confined, which makes the CNNs unable to capture the dependence of remote interaction information. Recently, different methods [19–23] have been presented to remedy the shortcomings of the convolution limitation problem. A typical solution is to expand the receptive field to learn the global dependency information, for example by increasing the network depth [24–26]. However, deepening of the network will lead to a sharp increase in the number of parameters. Recently, Transformer [27] with self-attention has become a novel choice for a variety of visual tasks [28–30] for its capturing interactions between any pairwise positions. For human pose estimation, we expect to leverage global dependencies captured by self-attention to provide contextual clues for occluded keypoints. Because the body keypoints themselves have certain connections, as shown in Figure 1, global dependencies are able to improve

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

the ability to locate difficult keypoints depending on easily detecting keypoints, thereby enhancing the performance of the overall network. There have been some recent works in CNN [31,32] directly model global dependencies with self-attention instead of convolution. For instance, CoT [32] encodes contextual information into the self-attention module, increasing representation ability of features. CoT can substitute 3×3 convolutions in ResNet [33], while owning fewer parameters.

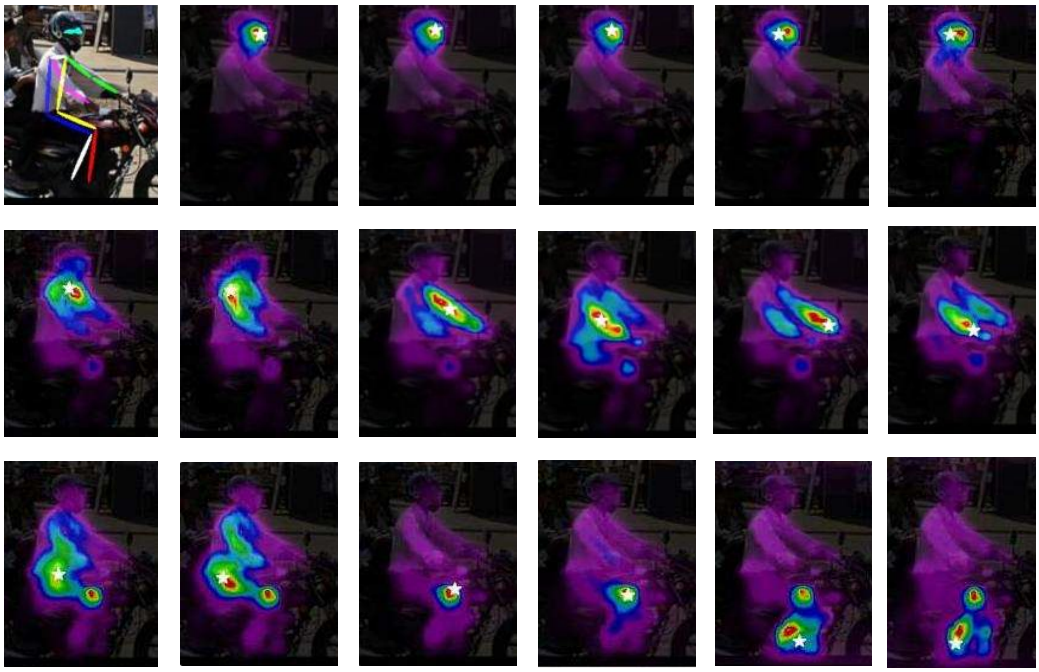


Figure 1. Attention map for position of each predicted keypoint. We can find that the motorcycle coveres the people's left ankle. The left ankle is predicted by relying on contextual information around the knee and the right leg joint.

In order to fully leverage the advantages of CNNs and self-attention mechanisms, some researchers have combined [34,35]them to extract features. However, there are still some drawbacks for multi-scale networks [36–39]. Each subnetwork of multi-scale neural networks has a different resolution in order to exchange information between multiple resolution representations in feature fusion. High resolution features with more attention to detail information can precisely locate the position information of the keypoints. Low resolution features with a larger receptive field can capture global information about the human pose. In feature fusion, the accuracy of keypoints detection will be enhanced if our model can fully exploit the benefits of high and low resolution. However, some existing methods [40,41] ignore the differences between features at different resolutions, resulting in undesired fusion of noise features. To bridge the differences between features at different resolutions, an effective approach is to utilize the attention mechanism. Because attention can make the network to stress or restrain information through learning, so that the network can better grasp the information we need to pay attention to. Recently, some scholars have conducted relevant research [42–44]. For example, CBAM [45]considers channel and spatial relationships and generates spatial attention maps finally. Therefore, we also expect our model to have the ability to learn information in both channel and spatial orientations.

Based on the above studies, in this article, we put forth a Lightweight Context-aware Feature Transformer Network (CaFTNet) built upon HRNet, to improve network efficacy by enhancing the localization accuracy of occluded keypoints. Firstly, to strengthen the semantic features of contextual information, we design a Transformerneck structure. Transformerneck directly replaces the 3×3 convolution in bottleneck with Contextual Transformer (CoT) block, while reducing the complexity of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

the network. Then inspired by the CBAM, to further refine the features of the fusion output, we design an Attention Feature Aggregation Module(AFAM). Due to the diversity of human poses, CBAM is still insufficient for spatial processing as it only employs a 7×7 convolution filter for feature fusion, while spatial attention is decided by the value of each pixel, not the region of 7×7. So, we propose an ARM to activate the obtained features. Therefore, our method further reinforces the feature fusion of multi-scale networks and ameliorates the output features. On the COCO [46], our model achieves better result than other methods with the CNN as the backbone. What's more, notably, the model reduces the number of parameter by 72.9%. On the MPII [47], our method takes advantage of 60.7% of the number of parameters, acquiring semblable results to other methods with the CNN as the backbone.

2. Related Work

2.1. Human Pose Estimation

Currently, CNNs have gained tremendous success in the field of human pose estimation. Hourglass [36] belongs to an hourglass type of network structure, which can perceive more global information. CPN [48] has two stages: GlobalNet and RefineNet, which can alleviate the detection problem of hard keypoints. Simple baseline [24] adds some transpose convolutional layers to restore the resolution. It indicates importance of high resolution feature maps. HRNet [14] proposes a network with high-resolution representations through the whole process, which repeats multi-scale fusions to improve the representation power of feature maps. Accordingly, HRNet achieves impressive results in multiple benchmark datasets. However, HRNet still falls within the category of CNNs, facing the issue of limited receptive fields. Therefore, global information needs to be ameliorated.

2.2. Attention enhanced Convolution

Convolution is dependent on a well-sized convolution kernel to gather information, which leads to the inability of the CNNs to establish global dependencies. The existing multiple approaches to image attention mechanisms suggest that they can compensate for the problem of confining of the convolution receptive field. Therefore, many scholars have explored the apply of attention to improve the capability of CNNs. SENet [43] models the interactions between the channels by using global mean pooling and two fully-connected layers. On the basis of SENet, ECANet [44] is came up with. Local cross-channel interaction strategy without decreasing the dimension is designed, which further improves the performance. CBAM [45] calculates attention maps on the channel and spatial directions to better learn useful information in the feature maps.

Recently, with the rise of the self-attention of Transfomer, the interest is aroused due to the powerful global dependence modeling ability of self-attention. Some works [49–52] have shown that self-attention modules are proposed as individual blocks, which can wholly substitute for the convolutions in HRNet. Although self-attention can effectively capture interactions between any paired positions. However, the pairwise query-key relationships are learned individually on isolated query-key pairs, without taking into account the abundant contextual information between them during the learning process.This seriously restricts the self-attention learning ability of two-dimensional feature maps for visual representation learning. Most recently, [31] displaces 3×3 convolutions with self-attention in the final stage of the network. [32] replaces 3×3 convolution in each bottleneck by leveraging the CoT block, which can take full advantage of the context of the query-key to model global dependencies.

2.3. HRNet

HRNet [14] utilizes a stem to fastly downsample the input feature. As shown in Figure 2, the HRNet network can be segmented into four stages. The first stage mainly has a high-resolution subnetwork. And from the second stage, a low-resolution subnetwork is added to each stage. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

resolution of the new subnetwork is half of the lowest resolution of the previous stage. Each stage will interact with information through multi-resolution blocks.

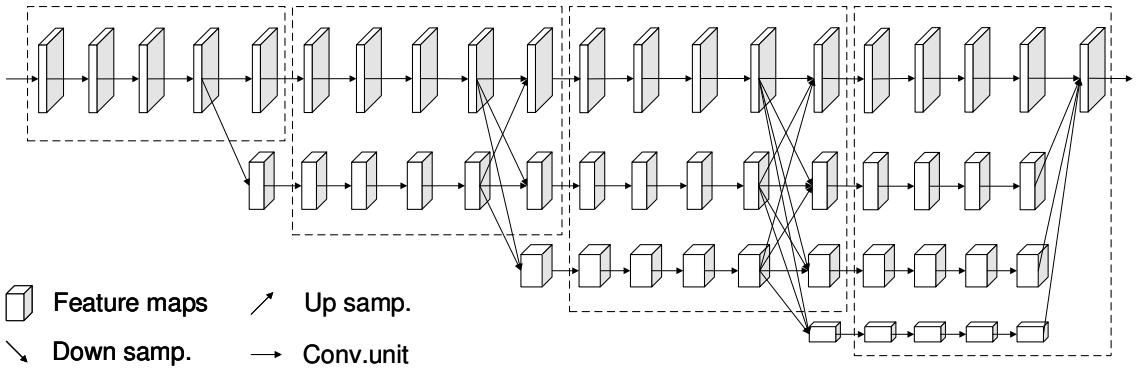


Figure 2. The architecture of HRNet.

HRNet has achieved remarkable success as a feature extractor. The problem of the limited receptive field inherent to the convolution operation needs to be improved. HRNet is unable to establish long-term dependencies, resulting in incorrect estimation of some human poses. For this reason, this paper proposes a Lightweight Context-aware Feature Transformer Network(CaFTNet). CaFTNet firstly capitalizes on the CoT block to enhance the expressiveness of the features. Then in feature fusion, the CaFTNet exploits AFAM to enhance the representative power of the output feature maps. And our final results are also better.

3. Methods

In this section, we put forward a CaFTNet to better perform the feature extraction. Figure 3 depicts the framework of our presented model. To begin with, we briefly review framework of our CaFTNet. Then, we introduce Transformerneck and AFAM in detail.

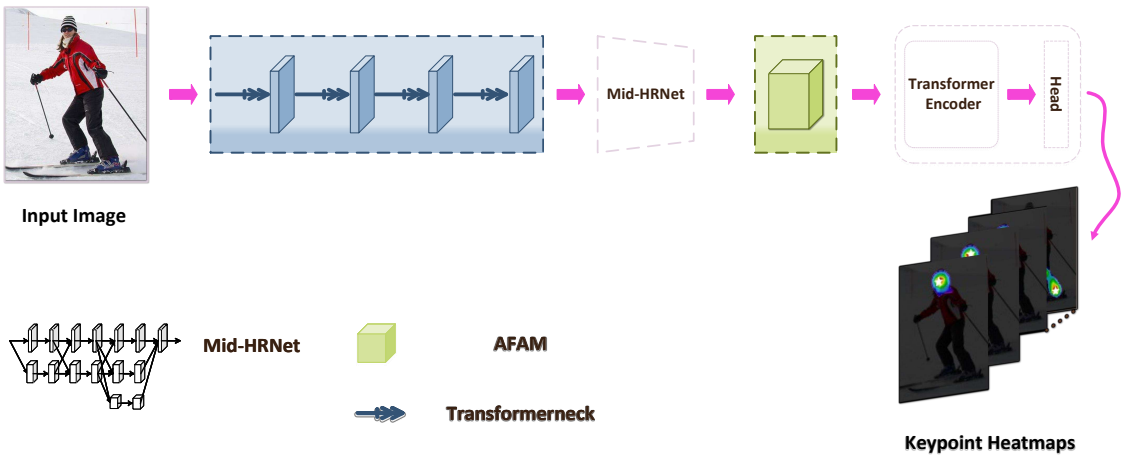


Figure 3. The overview of CaFTNet. Firstly, Transformerneck is used to extract preliminary input features with contextual information. Secondly, the input features continue to encode the feature information through the Mid-HRNet. Then AFAM further refines the contextual features. Next, Transformer Encoder Layer encodes position representation of keypoints. Finally, a head predicts keypoints heatmaps. Mid-HRNet refers to the second and third stages of the HRNet.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

3.1. *Context-aware Feature TransformerNetwork(CaFTNet)*

The aim of this paper is to enhance the representational ability of the feature maps and lighten network model in pose estimation. The overall architecture of CaFTNet is revealed in Figure 3 . The CaFTNet puts HRNet as the backbone and enhances it with the presented Transformerneck and Attention Feature Aggregation Module(AFAM).

First, our proposed Transformerneck is used to extract preliminary input features with contextual information. It is represented by a blue dashed line box. Transformerneck is to replace the 3×3 convolution with CoT while keeping the bottleneck framework unchanged. Secondly, this input features continue to encode the feature information through the Mid-HRNet. Then, we place the AFAM on the head of the neural networks to further refine the enriched contextual features. AFAM is represented by a green dashed line box. AFAM successively deduces attention maps along channel and spatial directions. Adaptive refinement module(ARM) is exploited to activate the obtained attention maps. The input undergoes adaptive feature refinement through multiplication with the activated attention maps. Next, output of the AFAM goes through Transformer Encoder Layer to encode position representation of keypoints. Finally, a head is attached to Transformer Encoder output to predict keypoints heatmaps.

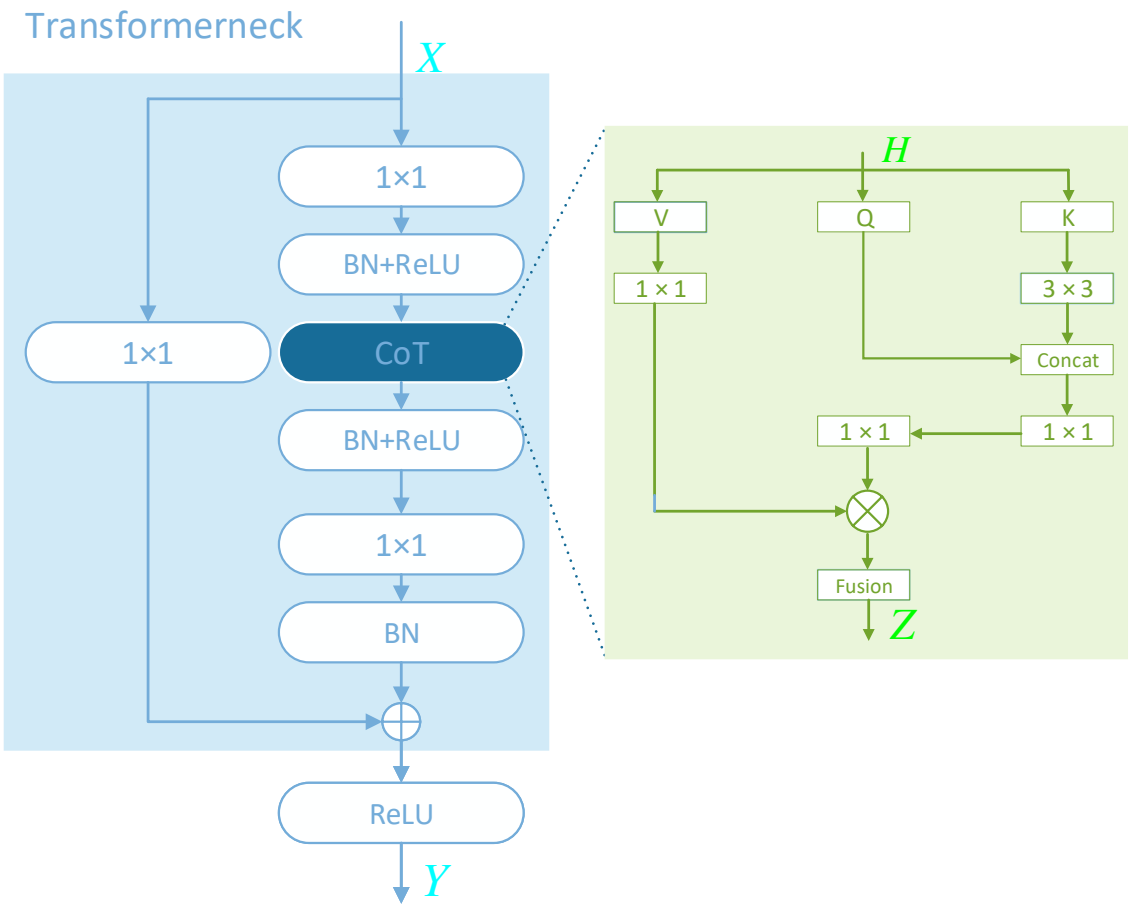


Figure 4. The overall structure of Transformerneck.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

3.2. Transfomerneck

For a middle input $X \in R^{H\times W\times C}$, an output H is first obtained through a 1×1 convolution and a nonlinear activation layer. The output H is sent into the CoT (As shown in the green rectangle enclosed in Figure 4). H is represented as:

$$H = ReLU(BN(Conv_{1\times1}(X))). \tag{1}$$

H will then be defined the K, Q and V along three different paths. The K first produces contextualized K_1 through 3×3 convolutions. The formula of K_1 is described as follows:

$$K_1 = Conv_{3\times3}(K). \tag{2}$$

Then, K_1 and Q are conducted the operation of concatenation. And the result of the operation generates the attention map M by two series of 1×1 convolutions. The formula of M is calculated :

$$M = Conv_{1\times1}(ReLU(Conv_{1\times1}(Concat(K_1, Q)))). \tag{3}$$

Next, V first passes through a 1×1 convolution to obtain V_1, the feature map K_2 can be computed as follows:

$$V_1 = Conv_{1\times1}(V), \tag{4}$$

$$K_2 = F(V_1 \otimes M), \tag{5}$$

where $F(\otimes)$ denotes the matrix multiplication operation. The final output Z of CoT is thus calculated as the fusion of K_1 and K_2 . Z continues to produce T through a nonlinear activation layer and a 1×1 convolution layer. The T and a shortcut connection are added element-wise to produce Y with context relations. Finally, Y is sent to the next module via the Relu activation function.

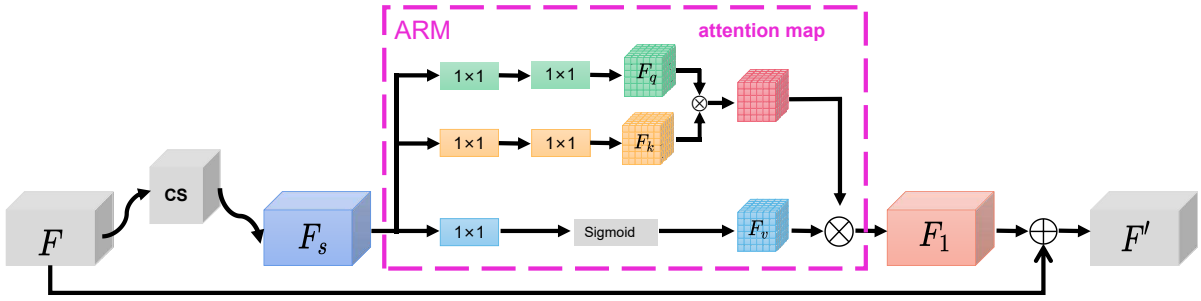


Figure 5. The overall structure of Attention Feature Aggregation Module. CS: Channel Attention Module, Spatial Attention Module.

3.3. Attention Feature Aggregation Module(AFAM)

To begin with, we consider feature map $F \in R^{H\times W\times C}$ as input in Figure 5. F through a cs module, generating the spatial attention map F_s that we need. This process can be described as two steps in Figure 6. The first step, F_c can be described as:

$$F_c = Sigmoid(MLP(AvgPool(F)) + MLP(MaxPool(F))) \otimes F. \tag{6}$$

The second step, the F_c was fed to the spatial attention model to gain the F_s. The F_s is adopted as:

$$F_s = Sigmoid(Conv_{7\times7}([AvgPool(F_c); MaxPool(F_c)])) \otimes F_c. \tag{7}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

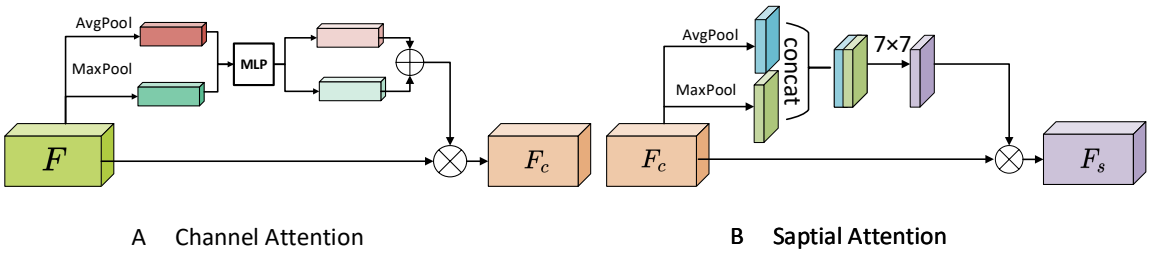


Figure 6. The overall structure of Convolution Block Attention Module.

Next, F_s is reshaped to feature sequences F_q, F_k and F_v in order to model the spatial context relations of the corresponding features. Detailed description of this process is as follows: (1)F_s obtains the spatial context feature F_v through a 1×1 convolution and a sigmoid layer in the last row. F_v is represented as:

$$F_v = Sigmoid\left(Conv_{1\times1}\left(F_s\right)\right). \tag{8}$$

(2)F_s rearranges the spatially related context features together respectively through two 1×1 convolutions and a non-linear activation layer to obtain F_q and F_k. F_q and F_k are represented as:

$$F_q = Con_{1\times1}\left(ReLU\left(Conv_{1\times1}\left(F_s\right)\right)\right), \tag{9}$$

$$F_k = Con_{1\times1}\left(ReLU\left(Conv_{1\times1}\left(F_s\right)\right)\right). \tag{10}$$

(3)The F_q and F_k are multiplied element-wise to obtain an attention map with contextual relationships, which is subsequently applied to feature for recalibrating the output feature F_1. F_1 is represented as:

$$F_1 = F_v \otimes Sigmoid\left(F_q \otimes F_k\right). \tag{11}$$

Finally, F_1 and F are added element-wise to achieve F'. The F' is adopted as:

$$F' = F_1 \oplus F. \tag{12}$$

4. Experiments

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

4.1. Model Variants

Based on HRNet, we present Lightweight Context-aware Feature Transformer Network. In our structure, there are two different depths of CNNs to extract the input features. Detailed setup information are presented in Table 1. The network utilized by CaFTNet-R is ResNet. The backbone network utilized by CaFTNet-H4 is the HRNet-W48. From Table 2, we can find that the model reaches best result when the network employs CaFTNet-H4.

4.2. Technical details

Our model takes advantage of the top-down [53–55] approach. The experimental environment configuration is shown as follows: Two RTX 2080s are deployed. Python Version is 3.7. Framework is PyTorch 1.10.0. The network model is optimized utilizing the Adam [56] optimizer during training, with an initial learning rate of 0.001 and 0.00001 at 220 rounds. The network is trained 230 rounds with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

a batchsize of 16 for each GPU. Because the size of the pictures in the dataset is different, the pictures are modified by image pre-processing. Here, images are cropped to 256 × 192 on the COCO and 256 × 256 on the MPII.

Table 1. Parameter configuration information for the different CaFTNet models.

Model	Backbone	Layers	Heads	Flops	Params
CaFTNet-R	ResNet	4	8	5.29G	5.55M
CaFTNet-H3	HRNet-W32	4	1	8.46G	17.03M
CaFTNet-H4	HRNet-W48	4	1	8.73G	17.30M

Table 2. Ablation study on different Backbone.

Model	Backbone	AP	AR	Flops	Params
CaFTNet-R	ResNet	73.7	79.0	5.29G	5.55M
CaFTNet-H3	HRNet-W32	75.6	80.9	8.46G	17.03M
CaFTNet-H4	HRNet-W48	76.2	81.2	8.73G	17.30M

4.3. Results on COCO

4.3.1. Dataset and Evaluation Metrics

The COCO [46] has more than 200,000 images and 250,000 instances, each containing up to 17 human keypoints. The network model is trained on the train2017 dataset, and the network model is verified and tested on val2017 (including 5,000 images) and test-dev2017 (including 20,000 images). Our model is measured by Object Keypoint Similarity (*OKS*) on the COCO dataset. *OKS* defines the similarity between different human keypoints, AP^{50} indicates the accuracy of the keypoints at OKS = 0.5, and AP^{75} is the accuracy of the keypoints at *OKS* = 0.75. The *mAP* is defined as the mean accuracy value of the predicted keypoints at the 10 thresholds of *OKS* = 0.50, 0.55 ... 0.90, 0.95. AP^M is utilized to describe the accuracy of the detection of medium size keypoints, and AP^L represents the accuracy of large size keypoints. The formula of OKS is described as:

$$OKS = \frac{\sum_i exp\left(-d_i^2/2s^2k_i^2\right)\delta\left(v_i > 0\right)}{\sum_i \delta\left(v_i > 0\right)}, \tag{13}$$

where d_i is the Euclidean distance between the i-th predicted keypoint coordinate and the corresponding groundtruth, v_i is the visibility flag of the keypoint, s is the object scale, and k_i is a keypoint-specific constant.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

Table 3. Comparison results with different other methods on the COCO val2017. CaFTNet-R and CaFTNet-H * reach good results in terms of parameter number and calculation speed.

Model	Input Size	AP	AR	Flops	Params
ResNet-50[33]	256×192	70.4	76.3	8.9G	34.0M
ResNet-101[33]	256×192	71.4	76.3	12.4G	53.0M
ResNet-152[33]	256×192	72	77.8	35.3G	68.6M
TransPose-R-A3[57]	256×192	71.7	77.1	8.0G	5.2M
TransPose-R-A4[57]	256×192	72.6	78.0	8.9G	6.0M
CaFTNet-R	256×192	**73.7**	**79.0**	**5.29G**	**5.55M**
HRNet-W32[14]	256×192	74.7	79.8	7.2G	28.5M
HRNet-W48[14]	256×192	75.1	80.4	14.6G	63.6M
TransPose-H-A4[57]	256×192	75.3	80.3	17.5G	17.3M
TransPose-H-A6[57]	256×192	75.8	80.8	21.8G	17.5M
TokenPose-L/D6[58]	256×192	75.4	80.4	9.1G	20.8M
TokenPose-L/D24[58]	256×192	75.8	80.9	11.0G	27.5M
CaFTNet-H3	256×192	**75.6**	**80.9**	**8.46G**	**17.03M**
CaFTNet-H4	256×192	**76.2**	**81.2**	**8.73G**	**17.30M**

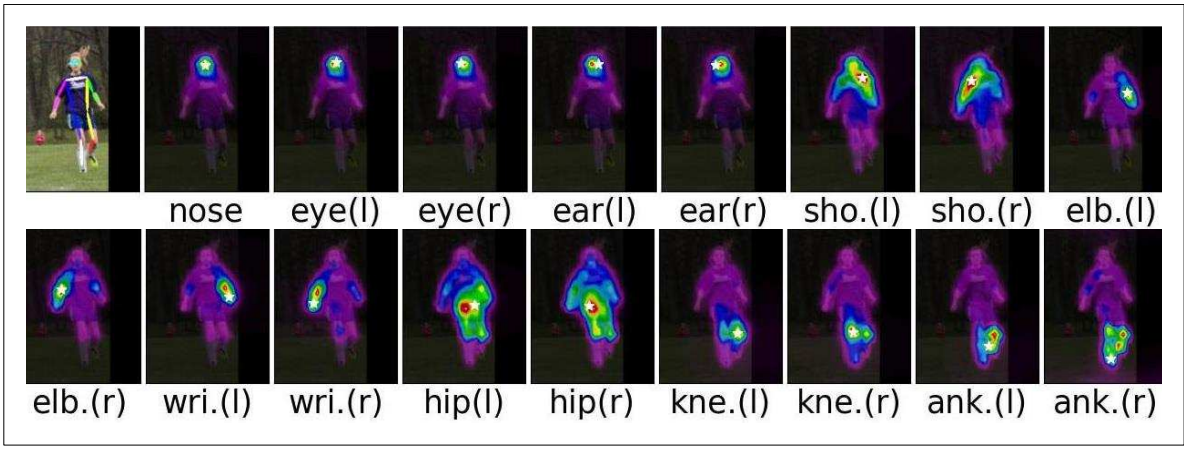
4.3.2. Quantitative Results

The models are compared for their performance on the COCO val2017, and the results are shown in Table 3. Where the number of parameters and GFLOPs are calculated from the human pose estimation network model. The experimental results show that the CaFTNet model gains good performance with less number of parameters and GFLOPs. CaFTNet-H4 acquires 76.2 AP scores with an input size of 256×192, better than other models with the same input size. In contrast to the TransPose-R-A4 [57], CaFTNet-R has an 8.3% drop in the number of parameters, but increasing AP scores by 1.1. In contrast to the ResNet-152 [33], our CaFTNet-R obtains better performance, utilizing only 7.2% of the model parameters. With the complex network model of HRNet-W48 [14] contrast, CaFTNet acquires a good AP score with much lower complexity. Table 4 exhibitions the results of our approach and the other approaches on the COCO test-dev2017. Our CaFTNet-H4 reaches 75.5 AP. Due to the effective perceptual context semantic information, CaFTNet achieves a good balance between accuracy and complexity.

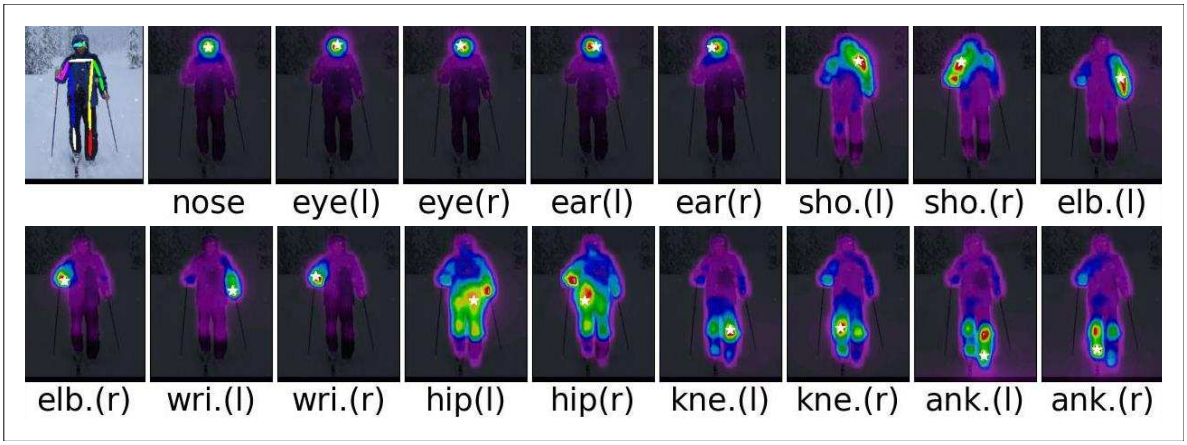
4.3.3. Qualitative Comparisons

Different keypoints rely on different regions. We can find that for the keypoints of the head like the nose, the eyes, etc. Their positioning depends mainly on the interdependencies between the keypoints, and it is worth noting that the prediction of the wrist or knees depend on the favorable cues around them. For instance, the prediction of the right knee depends on the left knee and the right lower limb. A closer look shows that our network has the ability to derive useful information from its relevant parts for keypoints to predict targets. In this way, we can understand why the model can predict the occluded keypoints(e.g. the occluded right knee in Figure 7 (**a**)).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1



(a) Visualization of image 1.



(b) Visualization of image 2.

Figure 7. Visualization of heatmaps predicting keypoints locations and their dependent regions for different input pictures according to the CaFTNet-R model.

Visualization. Contrast the results for more visualizations are revealed on COCO in Figure . The source image is displayed at the top of the picture, and the middle row displays the HRNet results, and our results are displayed in the last row. The objects (enclosed in red circles) that were not detected by HRNet in the first and second columns of images are likely due to occlusion by other objects. HRNet may have treated the undetected objects as background during the detection process. In comparison, the proposed AFAM in this paper can weight the features during information fusion output, allowing for better prediction of occluded objects. Images of in the third column, our approach can accurately detect occluded keypoints. This is because our model introduce CoT, which allows for better capture of contextual information, providing beneficial cues for detecting occluded keypoints. As a result, our approach achieves superior results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

Table 4. Comparison results with different other methods on the COCO test-dev2017. CaFTNet-R and CaFTNet-H * reach good results in terms of parameter number and calculation speed.

Model	Input Size	AP	AP^{50}	AP^{75}	AP^{m}	AP^{l}	Params
G-RMI[18]	357×257	64.9	85.5	71.3	62.3	70.0	42.6M
Integral[59]	256×256	67.8	88.2	74.8	63.9	74.0	45.0M
CPN [48]	384×288	72.1	91.4	80.0	68.7	77.2	58.8M
RMPE[16]	320×256	72.3	89.2	79.1	68.0	78.6	28.1M
SimpleBaseline[24]	384×288	73.7	91.9	81.8	70.3	80.0	68.6M
HRNet-W32[14]	384×288	74.9	92.5	82.8	71.3	80.9	28.5M
HRNet-W48[14]	256×192	74.2	92.4	82.4	70.9	79.7	63.6M
TransPose-H-A4[57]	256×192	74.7	91.6	82.2	71.4	80.7	17.3M
TransPose-H-A6[57]	256×192	75.0	92.2	82.3	71.3	81.1	17.5M
TokenPose-L/D6[58]	256×192	74.9	90.0	81.8	71.8	82.4	20.8M
TokenPose-L/D24[58]	256×192	75.1	90.3	82.5	72.3	82.7	27.5M
CaFTNet-H3	256×192	**75.0**	**90.0**	**82.0**	**71.5**	**82.5**	**17.03M**
CaFTNet-H4	256×192	**75.5**	**90.4**	**82.8**	**72.5**	**83.3**	**17.30M**

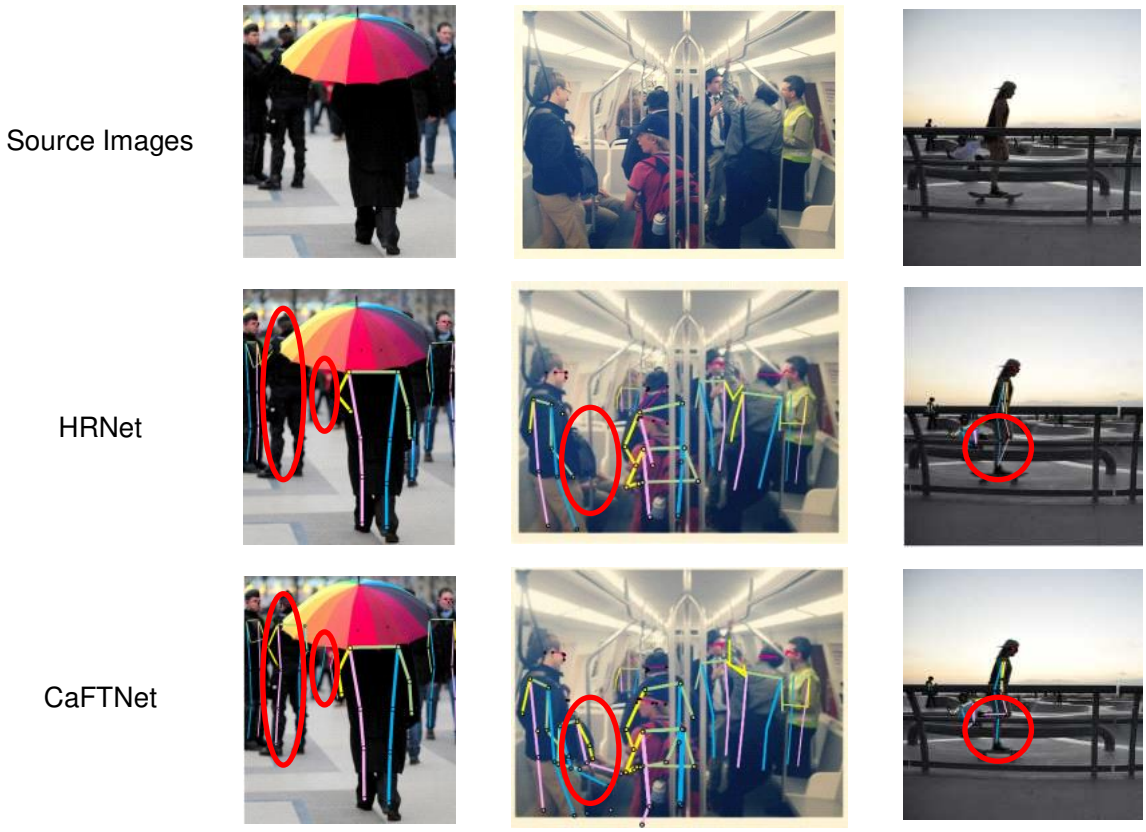


Figure 8. Qualitative comparisons on COCO.

4.4. Results on MPII

4.4.1. Dataset and Evaluation metric

MPII[47] is a single-person pose estimation dataset that captures the whole-body pose of people in real scenes, including 28,821 training images, 11,701 test images, a benchmark dataset in single-person pose estimation. The division of the training and validation sets are 22,246 and 2958 images,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

respectively. The standard evaluation index of MPII is *PCKh* (head-normalized percentage of correct keypoint), using the head segment length as the normalization reference. *PCKh* is described as:

$$PCKh_i = \frac{\sum_p \delta\left(\frac{d_{pi}}{L_p^{head}} \leq 0.5\right)}{\sum_p 1}, \tag{14}$$

$$PCKh_{mean} = \frac{\sum_p \sum_i \delta\left(\frac{d_{pi}}{L_p^{head}} \leq 0.5\right)}{\sum_p \sum_i 1}, \tag{15}$$

where i represents the i-th keypoint, p represents the p-th pedestrian, d_{pi} is the Euclidean distance between the p-th individual's i-th predicted keypoint coordinate and the corresponding groundtruth, $\delta(\cdot)$ represents the indicator function, L_p^{head} indicates the p-th head segment length. We report *PCKh@0.5* ($\alpha = 0.5$) score for fair comparison with other methods.

4.4.2. Quantitative Results

Table 7 presents the contrast results of the different approaches on the MPII, which can find that the approaches in this paper exceed the HRNet. In more detail, we can find from the Table 7 that although our final results are only higher than 0.1 of the baseline method. Especially for ankles detection, our method is 0.3 higher than the TokenPose-L/D24. And each test result of our method outperforms the baseline method. And thus proves that our method achieves better performance on this dataset. More importantly, our method uses only 61 percent of the number of baseline method parameters. Our results are better compared to SimpleBaseline-Res152, and number of our parameter decreases by 74.8%.

4.4.3. Qualitative Comparisons

We reveal some contrasting results on the MPII dataset in Figure 9. The source image is displayed at the top of the picture, and the middle row displays the HRNet results, and our results are displayed in the last row. From the visualization results of the third line, our method can correctly detect the keypoints of the occlusion not detected by HRNet. Mainly because our model can better capture the contextual information and provide favorable clues for blocking keypoints. Thus, our approach achieves better results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

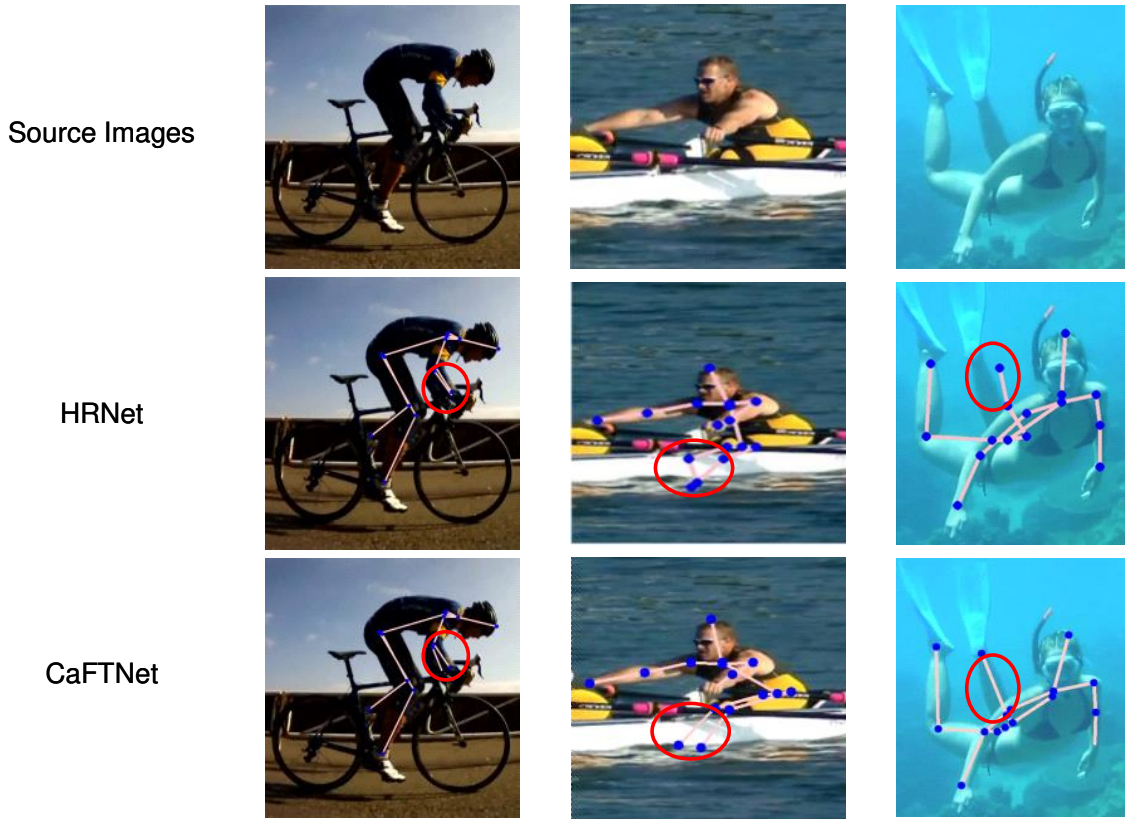


Figure 9. Qualitative comparisons on MPII.

4.5. Ablation experiments

The ablation experiments are chosen for training validation on the COCO dataset, considering the role of Transformerneck and AFAM in the network model.

4.5.1. Transformerneck

In this paper, two sets of ablation experiments are devised to verify the effect of employing bottleneck or Transformerneck separately based on our different network models. When implementing our network model with CaFTNet-H, we replace the bottleneck structure with our proposed Transformerneck, while keeping the other structures unchanged. The structures using Transformerneck reached 75.7 AP in Table 5. We also report the results of replacing bottleneck with Transformerneck when employing CaFTNet-R. Results with Transformerneck yield 0.6 more values than results employing bottleneck. The results illuminate the fundamentality of excavating the contextual information for decoding the subsequent features.

4.5.2. Attention Feature Aggregation Module(AFAM)

We investigate the effects of different attention on the experimental results, for example (i)SENet; (ii)ECANet; (iii)CBAM; (iv)AFAM. Due to their different operation of the feature map, the influence on the results of the experiment is also different. SENet [43] models the interactions between the channels by using global mean pooling and two fully-connected layers. On the basis of SENet, [44] proposes ECANet. Local cross-channel interaction strategy without decreasing the dimension is designed, which further improves the performance. [45] puts forward CBAM to focus more on spatial attention maps. AFAM compensates for the lack of spatial processing of CBAM, and it keeps the network focused at our more desired features. Table 6 presents the contrasting results from our different settings. Although the difference between them is small, we can become conscious of our proposed AFAM at 0.5 higher

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2024 doi:10.20944/preprints202401.0836.v1

than using SE. The results expose that more spatial information is needed when solving the feature fusion problems.

Table 5. The effects of CoT for different models on the COCO dataset.

Model	bottleneck	Transformerneck	AP
CaFTNet-R	✓		72.6
CaFTNet-R		✓	73.2
CaFTNet-H	✓		75.3
CaFTNet-H		✓	75.7

Table 6. Contrasting results for the COCO under different attentions.

Model	Baseline	SE	ECA	CBAM	AFAM	AP
CaFTNet-R	✓					72.6
CaFTNet-R	✓	✓				72.7
CaFTNet-R	✓		✓			72.8
CaFTNet-R	✓			✓		73.0
CaFTNet-R	✓				✓	73.2

Table 7. Results on the MPII validation set (PCKh@0.5).

Model	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Mean	Params
SimpleBaseline-Res50[24]	96.4	95.3	89.0	83.2	88.4	84.0	79.6	88.5	34.0M
SimpleBaseline-Res101[24]	96.9	95.9	89.5	84.4	88.4	84.5	80.7	89.1	53.0M
SimpleBaseline-Res152[24]	97.0	95.9	90.0	85.0	89.2	85.3	81.3	89.6	68.6M
HRNet-W32[14]	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3	28.5M
TokenPose-L/D24[58]	97.1	95.9	90.4	86.0	89.3	87.1	82.5	90.2	28.1M
CaFTNet-H4	**97.2**	**96.1**	**90.5**	**86.5**	**89.3**	**86.9**	**82.8**	**90.4**	**17.3M**

5. Conclusions

In this article, we put forth the Lightweight Context-aware Feature Transformer Network (CaFTNet) for enhancing the efficacy of human pose estimation models. Since CNNs cannot capture long-range dependencies between global regions, we devise Transformerneck. Furthermore, to bolster the representation power of the fusion output feature maps, we design Attention Feature Aggregation Module(AFAM). Extensive experiments carried out on the COCO and MPII corroborate the availability of the proposed approach.

6. Patents

Funding: This work is supported by The Natural Science Foundation of Hebei Province (F2019201451).

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	Linear dichroism

 doi:10.20944/preprints202401.0836.v1

References

1. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In Proceedings of the proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1227–1236.
2. Yang, C.; Xu, Y.; Shi, J.; Dai, B.; Zhou, B. Temporal pyramid network for action recognition. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 591–600.
3. Rahnama, A.; Esfahani, A.; Mansouri, A. Adaptive Frame Selection In Two Dimensional Convolutional Neural Network Action Recognition. In Proceedings of the 2022 8th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS). IEEE, 2022, pp. 1–4.
4. Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; Liu, J. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence* **2022**.
5. Snower, M.; Kadav, A.; Lai, F.; Graf, H.P. 15 keypoints is all you need. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6738–6748.
6. Ning, G.; Pei, J.; Huang, H. Lighttrack: A generic framework for online top-down human pose tracking. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 1034–1035.
7. Wang, M.; Tighe, J.; Modolo, D. Combining detection and tracking for human pose estimation in videos. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11088–11096.
8. Rafi, U.; Doering, A.; Leibe, B.; Gall, J. Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. Springer, 2020, pp. 36–52.
9. Kwon, O.H.; Tanke, J.; Gall, J. Recursive bayesian filtering for multiple human pose tracking from multiple cameras. In Proceedings of the Proceedings of the Asian Conference on Computer Vision, 2020.
10. Kocabas, M.; Athanasiou, N.; Black, M.J. Vibe: Video inference for human body pose and shape estimation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5253–5263.
11. Chen, H.; Guo, P.; Li, P.; Lee, G.H.; Chirikjian, G. Multi-person 3d pose estimation in crowded scenes based on multi-view geometry. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. Springer, 2020, pp. 541–557.
12. Kolotouros, N.; Pavlakos, G.; Black, M.J.; Daniilidis, K. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 2252–2261.
13. Qiu, H.; Wang, C.; Wang, J.; Wang, N.; Zeng, W. Cross view fusion for 3d human pose estimation. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4342–4351.
14. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. *arXiv e-prints* **2019**.
15. 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. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. Springer, 2020, pp. 455–472.
16. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. Rmpe: Regional multi-person pose estimation. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 2334–2343.
17. Newell, A.; Huang, Z.; Deng, J. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems* **2017**, *30*.
18. Papandreou, G.; Zhu, T.; Kanazawa, N.; Toshev, A.; Tompson, J.; Bregler, C.; Murphy, K. Towards accurate multi-person pose estimation in the wild. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4903–4911.
19. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.

20. Yang, W.; Li, S.; Ouyang, W.; Li, H.; Wang, X. Learning feature pyramids for human pose estimation. In Proceedings of the proceedings of the IEEE international conference on computer vision, 2017, pp. 1281–1290.
21. Jiang, W.; Jin, S.; Liu, W.; Qian, C.; Luo, P.; Liu, S. PoseTrans: A Simple Yet Effective Pose Transformation Augmentation for Human Pose Estimation. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V. Springer, 2022, pp. 643–659.
22. Tang, W.; Yu, P.; Wu, Y. Deeply learned compositional models for human pose estimation. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 190–206.
23. Ren, F. Distilling Token-Pruned Pose Transformer for 2D Human Pose Estimation. *arXiv preprint arXiv:2304.05548* **2023**.
24. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 466–481.
25. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
26. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**.
28. Raaj, Y.; Idrees, H.; Hidalgo, G.; Sheikh, Y. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4620–4628.
29. Luvizon, D.C.; Picard, D.; Tabia, H. Multi-task deep learning for real-time 3D human pose estimation and action recognition. *IEEE transactions on pattern analysis and machine intelligence* **2020**, *43*, 2752–2764.
30. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence* **2021**, *44*, 2872–2893.
31. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Vaswani, A. Bottleneck Transformers for Visual Recognition **2021**.
32. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual Transformer Networks for Visual Recognition **2021**.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
34. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
35. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the integration of self-attention and convolution. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 815–825.
36. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer, 2016, pp. 483–499.
37. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International conference on machine learning. PMLR, 2019, pp. 6105–6114.
38. Pfister, T.; Charles, J.; Zisserman, A. Flowing convnets for human pose estimation in videos. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 1913–1921.
39. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context attention for human pose estimation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1831–1840.
40. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5386–5395.
41. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

42. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-excite: Exploiting feature context in convolutional neural networks. *Advances in neural information processing systems* **2018**, *31*.
43. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
44. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534–11542.
45. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
46. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, 2014, pp. 740–755.
47. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, 2014, pp. 3686–3693.
48. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7103–7112.
49. Gao, P.; Lu, J.; Li, H.; Mottaghi, R.; Kembhavi, A. Container: Context aggregation network. *arXiv preprint arXiv:2106.01401* **2021**.
50. Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; Le, Q.V. Attention augmented convolutional networks. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 3286–3295.
51. Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Stand-alone self-attention in vision models. *Advances in neural information processing systems* **2019**, *32*.
52. Zhao, H.; Jia, J.; Koltun, V. Exploring self-attention for image recognition. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10076–10085.
53. Huang, J.; Zhu, Z.; Guo, F.; Huang, G. The devil is in the details: Delving into unbiased data processing for human pose estimation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5700–5709.
54. Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C. Distribution-aware coordinate representation for human pose estimation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 7093–7102.
55. Li, W.; Wang, Z.; Yin, B.; Peng, Q.; Du, Y.; Xiao, T.; Yu, G.; Lu, H.; Wei, Y.; Sun, J. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148* **2019**.
56. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
57. Yang, S.; Quan, Z.; Nie, M.; Yang, W. Transpose: Keypoint localization via transformer. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11802–11812.
58. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.T.; Zhou, E. Tokenpose: Learning keypoint tokens for human pose estimation. In Proceedings of the Proceedings of the IEEE/CVF International conference on computer vision, 2021, pp. 11313–11322.
59. Sun, X.; Xiao, B.; Wei, F.; Liang, S.; Wei, Y. Integral human pose regression. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 529–545.