Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080520
ISSN 2371-8838 Vol. 8 Num. 5

Fusing CNN and Transformer Network for Human Pose Estimation

Jiajia Shi[1], Fuchun Zhang[1,*], Zhenni Ma[1]

[1]School of Physics and Electronic Information, Yan'an University, Yan'an, 716000, China
[*]Corresponding author: yadxzfc@yau.edu.cn

Keywords: Keypoint detection, CNN, Transformer, feature fusion

Abstract: Accurate human pose estimation is essential for further human action recognition and behavioral analysis. Existing convolutional networks can extract local feature information but fail to model long-range dependencies, while Transformers excel at capturing global context but lose fine-grained details. To address this, we propose a dual-branch network called the Dual Transformer and CNN Network (DTCNet) that integrates global and local information for human pose estimation. DTCNet is proposed to improve human pose estimation by leveraging both global context and local features. It contains two branches - a Transformer branch that extracts global dependencies and a CNN branch that preserves local details. A fusion module then interacts between these branches, combining their complementary information to enhance representational power. Finally, the heatmap regression decoding unit obtains the pose estimations. Experiments demonstrate that through its dual-branch design, DTCNet effectively balances accuracy and efficiency while addressing limitations of previous methods. It achieves significantly higher average accuracy than the baseline on standard datasets, with 2.9% and 2.1% improvement on MPII and COCO respectively, validating that DTCNet better captures both long-range dependencies and fine-grained aspects needed for accurate pose estimation.

1. Introduction

Human pose estimation, also known as human skeleton detection [1], aims to extract joint points of the human body like shoulders, wrists and knees from images or videos. Those are connected according to rules to form a skeletal structure representing the pose. This process provides rich information on human pose and morphology. As a fundamental computer vision task, pose estimation supports various applications including behavior recognition [2], intelligent monitoring, intention recognition [3] and autonomous driving [4]. With continued research, pose estimation technology shows significant promise across fields such as visual analysis of human actions, video surveillance, understanding intent, and computer-assisted driving. The extraction of skeletal pose structures from images enables deeper human-centered visual understanding with wide-reaching applications.

Starting from the excellent performance of LeNet [5] in handwritten digit classification tasks, CNN has gradually attracted attention in the field of computer vision. Until the work of AlexNet [6], the network architecture based on CNN has really become the mainstream. With the deepening of

CNN research, deeper and more effective networks have been gradually proposed, such as VGG [7], GoogLeNet [8], ResNet [9] and so on. These excellent backbone networks are proposed on classification tasks, and directly applied to human pose estimation tasks often have poor performance. Therefore, Xiao et al. proposed a simple baseline (SBL) [10]. This network gradually shrinks the high-resolution feature map to the low-resolution feature map through the feature extraction network, and then restores the low-resolution feature map to the high-resolution feature map through transposed convolution, which provides the basic idea of human pose estimation network architecture design. In addition, Sun et al. proposed a High Resolution Network (HRNet) [11], which maintains high resolution throughout the feature extraction process by connecting multi-resolution subnets in parallel. Cai et al. proposed a novel network structure (Residual Steps Network, RSN) [12]. RSN uses multi-layer step convolution and step convolution to extract features and perform feature fusion, which improves the understanding and perception of global and local information. Thanks to the local connection and weight sharing mechanism of CNN, these network frameworks can not only effectively extract the rich detail information and complex texture features of the image, but also significantly reduce the computational complexity of the model. However, these methods are limited by the inherent inductive bias of convolution, and it is difficult to capture the correlation between global upper limbs and lower limbs, which affects the overall judgment of posture.

The Transformer [13] structure is not limited to local operations and can model global context information, which has excellent performance in natural language processing tasks. Dosovitskiy et al. proposed ViT(Vision Transformer) [14]. For the first time, the Transformer structure is applied to image classification tasks, which exceeds the classification accuracy based on CNN method. DeiT [15] introduced several training strategies and distillation methods that make data efficient, so that ViT can also be effective on smaller data sets. Swin [16] used local window self-attention instead of global self-attention to reduce the quadratic relationship between network complexity and image size to a linear relationship. At the same time, through local window movement and mask, the unity of local information self-attention and global information communication was completed, and the balance between speed and accuracy was achieved. Although these network architectures have achieved good performance in the field of image classification, it is challenging to apply them directly to pixel-level dense prediction because their output feature mapping is single-scale and low-resolution, even for common input image sizes. Its computational and memory costs are relatively high. Xu et al. proposed ViTPose [17] combined with ViT structure to transform human pose estimation into a sequence-based prediction task, opening up a new paradigm for pose tasks. However, the resolution of the output feature map of the ViT structure is low and single, resulting in the loss of local information. Yang et al. [18] and Li et al. [19] introduced the Transformer structure after extracting the minimum feature map from the CNN network, and converted the features extracted by the convolutional layer into sequences and input them into the Transformer to capture global dependencies, but their parameters are large and the computational complexity is high. Ludwig et al. [20] used readable vectors instead of tokens, and regressed any intermediate s by interpolation, which solved the problem that the previous method could only detect a fixed number of s. Although Transformer can model the global context, it has limitations in extracting fine-grained information. Referring to Swin Transformer [16], Wang et al. [21] proposed a pure Transformer structure based on pyramids, which is divided into multiple stages to generate features of different scales. This structure achieves better performance in dense prediction tasks, but it requires pre-training weights to exert its effect, resulting in its network structure can not be flexibly adjusted.

In order to leverage both the local detail extraction of CNNs and the global information modeling ability of Transformers, this paper proposes the Dual-Transformer and Convolutional

Network (DTCNet), which integrates CNNs and Transformers. The DTCNet introduces CNN branches alongside the Transformer to not only extract local information features but also enhance the network's ability to capture global context. Additionally, the proposed Dual-Fusion CNN-Transformer Block (DCTB) first inputs the encoder branches based on CNN and Transformer respectively for feature extraction and acquisition of long-range dependencies. It then splices the outputs of the CNN and Transformer branches to realize feature interaction and fusion, enhancing the model's representational power. Finally, the fused features are input again into the encoder branches for bidirectional fusion. Extensive experiments on the MPII and COCO datasets demonstrate that compared to other CNN-based or Transformer-based methods, the proposed DTCNet achieves significantly greater improvement over the baseline model and other state-of-the-art algorithms for human pose estimation.

2. Method

2.1. Overall Structure

Human pose estimation is essentially a pixel-level classification task, and obtaining global and local information of human s is the key to improving detection performance. Convolutional CNN is limited by fixed convolution kernels, and limited receptive fields cannot model global information. Transformer self-attention mechanism can obtain global context information by calculation, but it is easy to lose internal information during segmentation and stretching patch. In order to enhance the ability of network context information perception and retain rich detail information, a dual-branch detection network framework based on CNN and Transformer is proposed according to the characteristics of human image data. The overall structure is based on the form of feature extraction network (backbone) and feature reduction network (head).

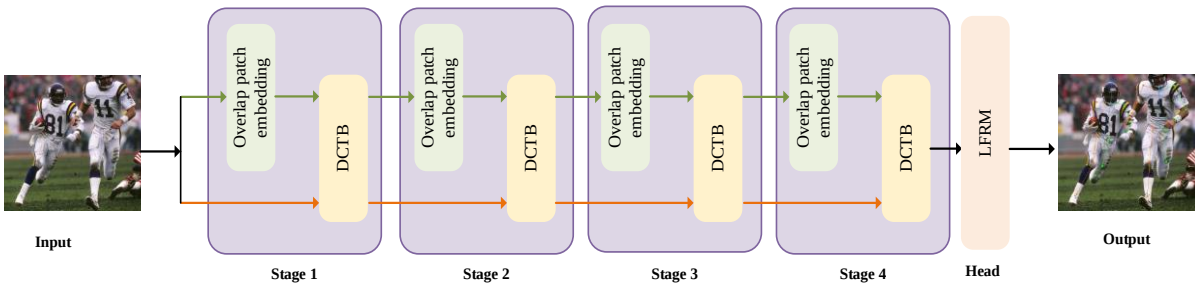

Figure 1: The network architecture of DTCNet

Figure 1 shows the overall structure of the model in this paper. According to the data characteristics of human body images, the model consists of two parallel branches. The CNN branch is directly input into the DCTB to extract features, gradually increasing the receptive field and extracting local information. The Transformer branch first divides the image into several small slices through OPE (Overlapping Patch Embedding), and then flattens the slices into a sequence. After transforming the dimension through the linear mapping layer, it enters the DCTB module to extract global information and downsampling the feature map. Then use the FB module to connect the intermediate features extracted from the independent branches in two directions. Convolutional fusion is performed on the connection features from two branches. Then the supplementary information is transmitted to the original branch in both directions. After repeating 4 stages, put into the LFRM (Last Feature Recovery Module, LFRM).

2.2. Dual-Fusion CNN-Transformer Block (DCTB)

Considering that the target scale of pose estimation is different, the shape is changeable and complex, and some s are blurred and accompanied by occlusion, it is necessary to fully combine the global and local feature information from Transformer branch and CNN branch. This paper proposes a DCTB module that integrates CNN and Transformer structure. It can use the respective advantages of CNN and Transformer to extract local and global features respectively. Through interactive fusion, it not only constructs context dependencies, but also enriches local detail information and enhances the ability of the network to extract features. DCTB is the core module of the feature extraction stage. The detailed structure of DCTB is shown in Figure 2, which consists of two branches, one is Transformer branch and the other is CNN branch.

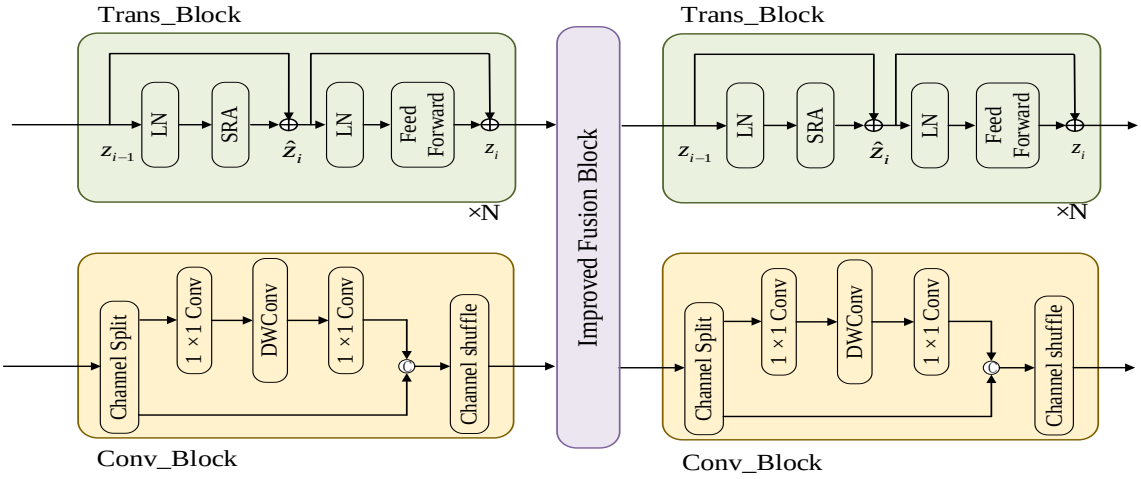

Figure 2: Dual-Fusion CNN-Transformer Block (DCTB)

2.2.1. Transformer branch

The Transformer branch is mainly composed of N continuous spatial reduction self-attention modules. In order to reduce the computational complexity, the multi-head self-attention mechanism of Trans_Block in this paper uses SRA [21]. Each module consists of LN, SRA, residual connection and feed forward network (FFN). Suppose that at stage i, the Transformer branch input is x_{i1}. The variables output by SRA and residual connection is $\hat{x}_{i1}$. The output of the Transformer branch is x'_{i1}. In the i+1 stage, the input of the Transformer branch is x_{i2}. The output connected by SRA and residual is $\hat{x}_{i2}$. The output of the Transformer branch is x'_{i2}. The output of the convolution module in the CNN branch after Flatten is $\tilde{y}_{i1}$. The overall calculation process is as follows:

$$\hat{x}_{i1} = SRA\left(LN\left(x_{i1}\right)\right) + x_{i1} \tag{1}$$

$$x'_{i1} = FFN\left(LN\left(\hat{x}_{i1}\right)\right) + \hat{x}_{i1} \tag{2}$$

$$\hat{x}_{i2} = SRA\left(LN\left(x_{i2}\right)\right) + x_{i2} \tag{3}$$

$$x'_{i2} = FFN\left(LN\left(\hat{x}_{i2}\right)\right) + \hat{x}_{i2} \tag{4}$$

In order to supplement the location information and enhance the local feature expression, the

output of the Transformer branch is fused with the output of the CNN branch. Among them, the feature size of the input sequence of the Transformer branch is $L_i \times D_i$, L_i is the characteristic length of the sequence in the i-th stage, D_i is the number of sequence feature dimensions of the i-th stage. The output feature map size of Conv_Block is $H_i \times W_i \times C_i$, H_i and W_i is the height and width of the feature map in the i stage, C_i is the number of channels of the i-stage feature map, and $L_i = H_i \times W_i$. Firstly, the output of the Transformer branch is rearranged into the form of sequence features, the size of which is $H_i \times W_i \times C_i$, and then concatenates with the output of Conv_Block on the feature dimension, the size is $H_i \times W_i \times 2C_i$. Then the spliced features are input into the FB module for feature fusion. Finally, the fused features are split and input into the T_fusion module for processing. The final output sequence feature size is $L_i \times D_i$.

2.2.2. CNN branch

When constructing our CNN branch, the common practice is to use ordinary CNN coding blocks to create a feature extraction network to achieve the conversion of high-resolution images to low-resolution images. For our network model, this approach requires no small computational overhead. Therefore, this paper draws on the lightweight idea of ShuffleNetV2 [22], and uses deep separable convolution [23] and channel shuffle instead of ordinary convolution to realize the lightweight of convolution operation. The former will greatly reduce the network size and computational overhead of the convolution operation, and the latter will supplement the information loss caused by the convolution grouping through channel shuffle.

As shown in Figure 2, the Conv_Block module uses the channel spilt module to divide the number of channels of the input image into two parts on average, one for residual connection and one for feature extraction. The channel shuffle module reorders the channels of the stacked feature maps to achieve feature fusion between groups. In the basic module, the shape of the feature map is unchanged, and the number of channels is unchanged. In the down-sampling module, the length and width of the feature map are halved, and the number of channels is doubled.

2.2.3. Improved Fusion Block

Figure 3 depicts our Improved Fusion Block (IFB). IFB consists of FB module, T-Fusion module and C-Fusion module. The FB module is mainly composed of four continuous convolution residual bottleneck modules. The processing process is similar to the CNN branch. Each convolution residual bottleneck module includes 1×1 convolution, ReLU and another 1×1 convolution. The residual connection between the input and the output accelerates the model convergence. Specifically, the intermediate feature sum T_i and M_i is given from the i-th CNN_Block and Transformer_Block， and the FB module is used to fuse the feature maps of two branches:

$$M_i = FB(Concat(rearrange(T_i), F_i)) \tag{5}$$

Among them, $M_i \in R^{H_i \times W_i \times 2C_i}$ represents the fusion feature. After preliminary fusion, the fusion feature M_i is divided into two features along the channel dimension, which is $M_i^T \in R^{H_i \times W_i \times C_i}$ and $M_i^F \in R^{H_i \times W_i \times C_i}$, and then input into the T-Fusion module and the C-Fusion module, which are composed of MLP blocks and convolution blocks. Then, each fused feature flows back to each branch and added to the original input features T_i and F_i.

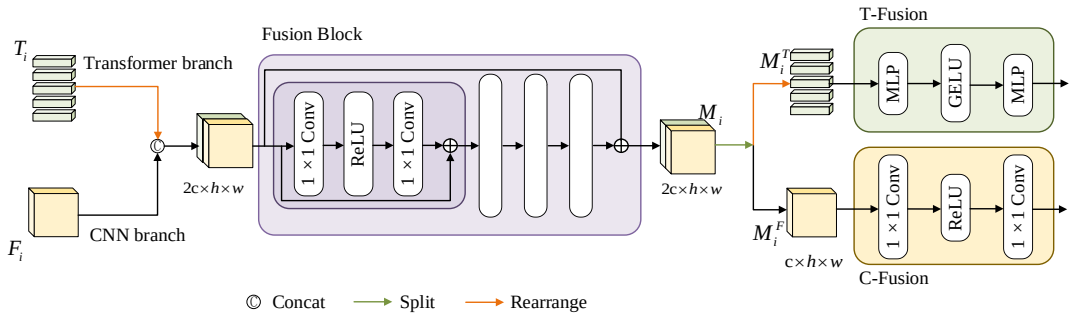

Figure 3: Structure of Improved Fusion Block (IFB)

2.3. OPE Block

This paper adds an OPE (Overlapping Patch Embedding) module, which uses overlapping to mark images and model local continuity information. The commonly used Patch Embedding operation ignores the correlation between patch blocks, potentially leading to loss of features around patch boundaries. In this paper, the OPE module is used to downsampling the feature map. When cutting the patch, there is an overlapping between each patch. The operation is similar to the movement of the convolution kernel on the feature map. The resolution of the feature map is controlled by modifying the values of the three parameters K, S and P. K represents the size of the patch, S represents the distance between adjacent patches, and P represents the size of the filling. We expanded the patch window to make the adjacent windows overlap half of the area, and filled zeros on the feature map to maintain the resolution.

2.4. Feature Recovery Module

The final feature reduction module LFRM is used to fuse the final output of the Transformer branch and the CNN branch, and three sets of Deconvolution (DeConv) are used to realize the upsampling operation of the low-resolution feature map. Finally, the 1×1 convolution is used to adjust the channel to the heatmap regression, and its structure is shown in Figure 4.

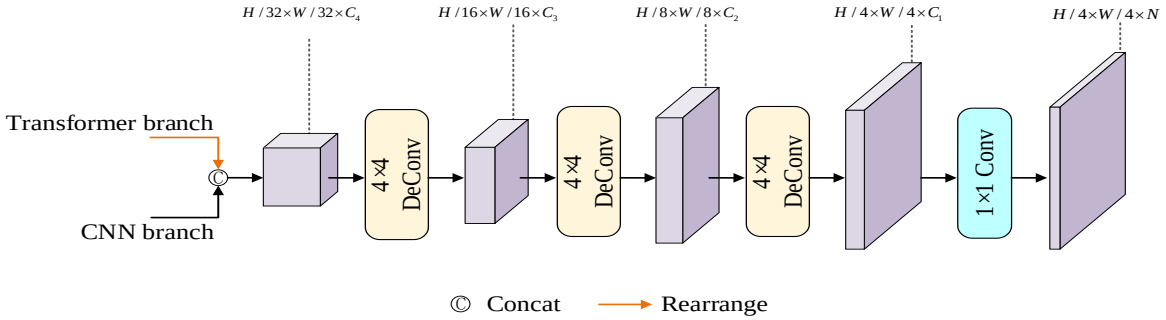

Figure 4: Last Feature Recovery Module (LFRM)

The output of the two branches is fused by the fusion module after concating. The output of the Transformer branch first undergoes a Rearrange operation, while the output of the CNN branch is spliced along the channel dimension. After convolution in the Fusion Block, the final fused feature map is obtained.

3. Experimental Analysis

3.1. Datasets and Evaluation Criteria

This method follows a two-stage top-down human pose estimation paradigm, similar to CPN [24], and is verified by two benchmark datasets (MPII and COCO).

The MPII [25] human posture dataset consists of about 25,000 images from a wide range of real-world activities, including 40,000 individual instances and body posture annotations marked with 16 s. About 28,000 individual instances are used as training samples, and about 12,000 individual instances are used as test samples. For the MPII dataset, the standard evaluation index of the top-down paradigm of human pose estimation is the head normalization probability of the correct. In this paper, PCKh with a threshold of 0.5 is used as the evaluation criterion.

The COCO [26] keypoint detection dataset contains more than 200,000 images and 250,000 human instances with 17 keypoints. The dataset is divided into training set train2017, validation set val2017 and test set test-dev2017, with 57,000, 5,000 and 20,000 images respectively. For the COCO dataset, the standard evaluation index of the top-down paradigm of human pose estimation is based on the similarity of targets (OKS). This paper uses AP (the average precision of key points at OKS = 0.50, 0.55, ..., 0.90, 0.95), AP75 (precision at OKS = 0.75), AP50, APM (precision for detecting medium objects), APL (precision for detecting large objects) and AR (the average recall of keypoints at OKS=0.50, 0.55, ..., 0.90, 0.95) as evaluation metrics.

3.2. Experimental Environment and Settings

The experimental environment of this paper is Ubuntu18.04, the CPU is Intel (R) Xeon (R) Gold6230, the GPU is TeslaT4 (16G), the Python version is 3.8.18, the Pytorch version is 1.8.1, the optimizer is Adam, the learning plan follows the setting of SBL [10], the basic learning rate is set to 0.0005, and it is reduced to 0.00005 and 0.000005 respectively in the 170th and 200th times, and the training process ends in 210 cycles. The training batch size of the model in this paper is set to 64, and the model is trained from scratch without using pre-training weights. Based on the MMPose code library, this paper adopts common training strategies, that is, different data pipelines are set for the training set and the validation set. For the training set, the clipping strategy is to first extend each human detection box to a fixed aspect ratio (height: width = 1.25), and then a translation factor (0.16) is set for random translation. Then, a data augmentation strategy is designed, including random flipping, random rotation ([-40, 40]) and random scaling ([0.5, 1.5]). The main purpose is to improve the scale invariance and rotation invariance. As shown in Figure 5, the network effect is visualized as follows.

Figure 5: Visualization

3.3. Experimental Results and Analysis

This paper follows the top-down human pose estimation paradigm and compares the detection accuracy with other representative advanced methods on MPII and COCO datasets. Table 1 reports the comparison of the proposed method with other methods on the MPII verification set. The results show that the proposed method can maintain good detection accuracy and is competitive in the model.

From Table 1, it can be seen that on the MPII test set, the DTCNet has excellent detection performance. It has better performance in hand, shoulder and other parts average PCK index, reaching 96.52%, 95.58% and 89.01%, respectively. Models such as Hourglass-52 and ShuffleNetV2 are constructed based on CNN structure, which can capture local context information and locate pixels, but lack of global information. Models such as PVT-S and Swin-S take advantage of Transformer ability to establish long-distance feature dependencies to obtain rich global information, but lack of fine-grained local information.

Compared with the human detection baseline model PVT, the DTCNet has improved in all indicators, and the average PCK index has increased by 4.61%. Compared with the CNN network ShuffleNetV2, the average PCK index is increased by 6.61%. Compared with other advanced ratio models, the network DTCNet in this paper reaches the best value in the average PCK index, indicating that the model can accurately detect the human body, and its prediction results are highly similar to the real values. The two-way fusion of CNN and Transformer enables the model in this paper to better perceive detailed information, reduce the probability of regional false detection, and improve the detection accuracy.

Table 1: Comparison of experimental results of MPII validation set (PCKh@0.5)

Method	Input Size	Params/10^6	GFLOPs	Hea/%	Sho/%	Elb/%	Wri/%	Hip/%	Kne/%	Ank/%	Mean/%
ResNest-50	256×256	35.93	8.97	96.3	95.6	89.1	84	87.8	84.9	80.4	88.8
Simple Baseline	256×256	34	7.28	96.1	95.1	88.2	81.9	88.1	83	77.5	87.7
Swin-S	256×256	54.1	15.4	96.1	94.8	87.2	80.8	87.9	82.3	77.8	87.3
EfficientViT	256×256	3.04	1.89	95.4	94	85.6	78.9	86.3	79.8	73.7	85.8
LiteHRNet-30	256×256	1.76	0.56	95.2	93.5	84.7	78.1	86.2	78.9	73.8	85.1
MobileNetV2	256×256	9.57	2.12	95.6	93.8	84.8	77.8	85.7	79.4	73	85
PVT-S	256×256	28.17	5.47	95	93.3	84	77.4	85	78.1	72.8	84.4
ShuffleNetV2	256×256	7.55	1.83	94.2	92.3	81.9	74.1	84	75.2	68.8	82.4
Hourglass-52	256×256	94.85	28.67	96.5	95.5	88.8	83.8	88.2	85	80.8	88.9
DTCNet(ours)	256×256	44.12	8.29	96.52	95.58	88.97	83.85	89.16	85.16	80.32	89.1

In addition, this paper also compares and analyzes the parameters and computational complexity of each pose estimation segmentation model, and the results are shown in Table 1. In order to extract more abundant and effective feature context information, DTCNet adopts parallel dual-branch structure to obtain global and local complex information respectively. The feature fusion module complements and fuses missing information. The module structure is relatively complex, resulting in a slightly higher parameter number and computational complexity of the model. Compared with ResNest-50 and HRNet-W32 models, the parameters in this paper are higher, but the computational complexity is reduced by 0.68G and 1.56G. Compared with Hourglass-52, Swin-S and other models based on CNN and Transformer, the DTCNet parameters and computational complexity of this network are greatly reduced. However, compared with the lightweight models such as EfficientViT-M0 and MobileNetV2, the network DTCNet parameter number and calculation index in this paper are relatively high, which is a trade-off process. However, for the field of pose estimation, detection accuracy and accuracy are crucial goals, and the added parameters are conducive to improving the model segmentation effect. Combining PCK and other

performance indicators, parameter quantity and computational complexity, DTCNet has appropriate parameters and computational complexity while maintaining high detection accuracy.

Table 2: Comparison of experimental results of COCO val2017 set

Method	Input Size	Params/106	GFLOPs	AP/%	AP50/%	AP75/%	APM/%	APL /%	AR/%
LiteHRNet-30	256×192	1.76	0.42	67.6	88	75.6	-	-	73.6
MobileNetV2	256×192	9.57	1.59	64.8	87.4	72.5	-	-	70.9
ShuffleNetV2	256×192	7.55	1.37	60.2	85.7	67.2	-	-	66.8
LDMNet	256×192	4.8	2.1	70.1	91.5	78.2	67.5	74.1	73.4
PVT-S	256×256	28.17	5.47	70.9	91.5	78.4	68.3	75.2	74.1
LEViTPose-S	256×256	2.16	1.45	71	91.6	78.5	68.2	75.1	74.1
ViTPose-B	256×256	90.04	23.8	73.2	92.5	81.5	71.2	76.6	76.5
DTCNet(ours)	256×192	44.12	8.29	73.4	90.1	81.1	70.1	79.9	78.9

Table 2 reports the comparison results of DTCNet with other methods on COCO val2017. The average accuracy of the DTCNet network in the COCO validation set reaches 73.4%, which has a great advantage over the mainstream human pose estimation algorithms. As shown in Table 2, compared with the Transformer-based model ViTPose-B, DTCNet improves the average accuracy by 0.2%, but the number of parameters decreases by 50.9% and the amount of computation decreases by 65.1%. Compared with the baseline network PVT-S, the average accuracy of the network DTCNet in this paper is increased by 2.5%, indicating that the addition of CNN branch has achieved better results. Compared with other networks in the table, our method has great advantages. By comparing the performance of the DTCNet on the val set and test set of MPII and COCO, it is found that the performance growth of the network is relatively balanced, and there is no obvious over-fitting phenomenon, which shows that the model designed in this paper can effectively learn the generalization features in the data. On the whole, the algorithm proposed in this paper has great competitiveness on MPII data set and COCO data set. In the case of strict detection accuracy requirements, the advantage of DTCNet is more obvious, and the design goal of the human pose estimation algorithm is achieved. In addition, the model in this paper has good computational efficiency and competitive in real-time reasoning.

3.4. Ablation Experiment

The ablation experiment in this paper is carried out on the MPII validation set. All configuration strategies follow the settings in the MPII comparison experiment, including input size of 256×256, batch size of 64, etc. In this paper, the Only-trans algorithm framework proposed by PVT is used as the basis for model design. This model is used as a baseline for model improvement, called DTCNet (ours), as shown in Table 3.

Table 3: Data analysis for ablation experiments

Method	DCTB				OPE	Mean/%
	Trans	Conv	FB	IFB		
PVT	√	×	×	×	×	85.9
DTCNet(ours)	√	√	√	×	×	88.7
	√	√	×	√	×	88.8
	√	√	√	×	√	88.9
	√	√	×	√	√	89.1

The ablation experiment mainly includes the CNN branch, the feature fusion block (FB), the improved feature fusion module (IFB), and the overlapping embedding block (OPE).

According to the results of Table 3, only adding CNN branch and feature fusion module can

greatly improve the accuracy of the model, but the loss of parameters and GFLOPs is large. Using the improved feature fusion block module, only a small amount of model parameters are increased while the accuracy is improved. The addition of overlapping feature embedding module can further improve the accuracy of the model without increasing the amount of parameter calculation. The final experimental results show that by adding the CNN branch and overlapping feature embedding module, the DTCNet (ours) designed using the improved Fusion Block module can achieve the best results.

4. Conclusion

The main contribution of this paper is to propose a dual-branch network for human pose estimation with global and local information interaction. This structure extracts local features through CNN branches and Transformer branches model global context information, which can better identify details and suppress background interference, thus effectively dealing with the difficulty of human detection. A fusion module is designed that effectively combines the local features extracted by CNN and the global features extracted by Transformer, improving the detection ability of the network. The Transformer structure in this model does not need to be pre-trained on large-scale data. The effectiveness of the method is verified on public datasets like COCO and MPII. The model balances accuracy and efficiency well with its parameter quantity and computation, achieving good accuracy. While the proposed model can accurately detect the target areas of human bodies, it is a two-dimensional network, which is more suitable for two-dimensional images. In follow-up studies, the model will be further improved and optimized to more effectively apply to three-dimensional human pose estimation with richer image information.

References

[1] S. Salisu, A. S. A. Mohamed, M. H. Jaafar, A. S. B. Pauzi, H. A. Younis, A Survey on Deep Learning-Based 2D Human Pose Estimation Models,Computers, Materials Continua, 2023.

[2] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting Skeleton-based Action Recognition, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 2959-2968.

[3] Z. Fang, A.M. López, Intention Recognition of Pedestrians and Cyclists by 2D Pose Estimation,IEEE Transactions on Intelligent Transportation Systems, 2020.

[4] M. Lu, Y. Hu, X. Lu, Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals,Applied Intelligence, 2020.

[5] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition,Proceedings of the IEEE, 1998.

[6] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks,Commun. ACM, 2017.

[7] S. Liu, W. Deng, Very deep convolutional neural network based image classification using small training sample size, 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 2015, 730-734.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S.E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions,2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition,2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[10] B. Xiao, H. Wu, Y. Wei, Simple Baselines for Human Pose Estimation and Tracking, V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.) Computer Vision – ECCV 2018, 2018, 472-487.

[11] K. Sun, B. Xiao, D. Liu, J. Wang, Deep High-Resolution Representation Learning for Human Pose Estimation, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 5686-5696.

[12] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhou, E. Zhou, X. Zhang, J. Sun, Learning Delicate Local Representations for Multi-Person Pose Estimation, European Conference on Computer Vision, 2020.

[13] A. Vaswani, N.M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is All you Need, Neural Information Processing Systems, 2017.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G.

Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,ArXiv, 2020.
[15] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H.e. J'egou, Training data-efficient image Transformers & distillation through attention, International Conference on Machine Learning, 2020.
[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
[17] Y. Xu, J. Zhang, Q. Zhang, D. Tao, ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation, ArXiv, 2022.
[18] S. Yang, Z. Quan, M. Nie, W. Yang, TransPose: Towards Explainable Human Pose Estimation by Transformer, ArXiv, 2020.
[19] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S. Xia, E. Zhou, TokenPose: Learning Keypoint Tokens for Human Pose Estimation,2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
[20] K. Ludwig, P. Harzig, R. Lienhart, Detecting Arbitrary Intermediate Keypoints for Human Pose Estimation with Vision Transformers, 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), 2022, 663-671.
[21] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions,2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
[22] N. Ma, X. Zhang, H. Zheng, J. Sun, ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design, ArXiv, 2018.
[23] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,ArXiv, 2017.
[24] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded Pyramid Network for Multi-person Pose Estimation, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 7103-7112.
[25] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, 3686-3693.
[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common Objects in Context, D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.) Computer Vision – ECCV 2014, 740-755.