Adaptive Target Region Attention Network-based Human Pose Estimation in Smart Classroom

Jianwen Mo[1], Guiyun Jiang[2], Hua Yuan[3]*, Zhaoyu Shou[4], Huibing Zhang[5]
School of Information and Communication,Guilin University of Electronic Technology, Guilin 541004, China[1, 2, 3, 4]
School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China[5]

***Abstract*—In smart classroom environments, problems such as occlusion and overlap make the acquisition of student pose information challenging. To address these problems, a lightweight human pose estimation model with Adaptive Target Region Attention based on Lite-HRNet is proposed for smart classroom scenarios. Firstly, the Deformable Convolutional Encoding Network (DCEN) module is designed to reconstruct the encoding of features through an encoder and then a multi-layer deformable convolutional module is used to adaptively focus on the image region to obtain a feature representation that focuses on the target region of interest of the student subject. Secondly, the Channel And Spatial Attention (CASA) module is designed to attenuate or enhance the feature attention in different regions of the feature map to obtain a more accurate representation of the target feature. Finally, extensive experiments were conducted on the COCO dataset and the smart classroom dataset (SC-Data) to compare the proposed model with the current main popular human pose estimation framework. The experimental results show that the performance of the model reaches 67.5(mAP) in the COCO dataset, which is an improvement of 2.7(mAP) compared to the Lite-HRNet model, and 86.6(mAP) in the SC-Data dataset, which is an improvement of 1.6(mAP) compared to the Lite-HRNet model.**

Keywords*—Human pose estimation; smart classroom; Lite-HRNet; deformable convolutional encoding network; target region attention

I. Introduction

In recent years, human pose estimation [1,2,3,4] technology has been widely used in behaviour recognition, action recognition, human-computer interaction and other scenarios along with the rapid development of related technologies in the field of computer vision. With modern and intelligent education being strongly advocated and developed, neural network models based on deep learning are heavily used in classroom detection tasks. In the task of assessing the quality of teaching and learning of student, information about student postures [5,6,7] plays a very important role in assessing the quality of teaching and in teacher understanding of student learning status in the classroom [8]. In the classroom, a student state of learning is demonstrated through a variety of classroom behaviours. Student who are not interested in the content of the classroom will exhibit behaviours such as dawdling, playing with mobile phones and sleeping. Student who are interested in the content of the class show behaviours such as concentration, looking at the board, taking notes, reading, and actively interacting with the teacher. Therefore, how to automatically and accurately collect student pose information in smart classroom [9,10,11] scenarios is an important task that needs to be solved urgently. In smart classroom environments, the acquisition of student pose information is commonly associated with problems such as overlap and occlusion between students, as well as the location of the students leading to large differences in their body sizes, and the presence of small target instances leading to a degradation of the model detection performance. At the same time, the problem of large computational and parametric quantities of the human pose estimation model makes it more difficult to be deployed in smart classroom scenarios. These problems make the acquisition of pose information in smart classrooms a challenging research.

Aiming at the problems of overlapping and occlusion, as well as the large number of model parameters in smart classroom scenarios, this paper proposes a lightweight human pose estimation model framework based on the Lite-HRNet architecture applied to smart classroom scenarios, the Adaptive Target Region Attention Network for Human Pose Estimation. The model is designed with two main modules: (1) The Deformable Convolutional Encoding Network is designed for obtaining a target feature region representation. (2) The Channel And Spatial Attention module is designed to allow the target feature region representation to obtain a more accurate representation of the target region. The model in this paper achieves relatively good performance on two datasets. Extensive ablation experiments are used to validate the effectiveness of each module in the proposed method. The main contributions of this study are summarised as follows:

1) Propose a lightweight pose estimation model for smart classrooms: the Adaptive Target Region Attention Network for Human Pose Estimation. And to construct a student pose estimation dataset suitable for smart classroom environment to provide a database for pose detection in smart classrooms.

2) The Deformable Convolutional Encoding Network (DCEN) is proposed to perform feature extraction on the target region of the feature map to obtain a vector representation with feature regions of interest. The experimental results show that the module designed in this paper can efficiently improve the performance of the model.

3) Proposing an attention mechanism based Channel And Spatial Attention (CASA) module to be used to assist in model training. The module enables the target feature region representation to obtain a better attention effect and fully exploits the spatial and channel feature information in the target feature region.

The rest of the paper is organised as follows. In Section II, the elements involved in the related work are presented. In Section III, the proposed method is described in detail. In Section IV, the experimental results are described and analysed. Finally in Section V, the conclusion of the paper is drawn.

II. Related Work

Traditional methods for human pose estimation are based on graphical structure solutions, which rely too much on hand-crafted feature, are more influenced by algorithms, and have limited model representation capabilities. Deep learning human pose estimation modelling methods are broadly classified into two types: Bottom-Up and Top-Down. Bottom-Up methods [12,13] first detect individual body parts and then compose these detection gesture points into a whole person. On the other hand, the Top-Down approach [14,15,16] first detects the human body bounding box and then detects the human body pose within each bounding box.

Among them, a high-resolution network (HRNet) [17] with top-down approach, has become a mainstream method for human pose estimation due to its efficient detection performance. However, as the performance of the human pose estimation model improves, it is accompanied by a significant increase in the number of parameters. Wang [18] In order to address the problem of huge computational effort associated with attitude estimation models for high-resolution structures. A fused inverse convolution head module is used to eliminate redundancy in the high-resolution branch and achieve scale feature fusion with low computational effort. As well as the use of large convolution kernels to improve the sensory field of the model and reduce the computational effort of the model. The IGCV3 [19] model decomposes the regular convolution into multiple grouped convolutions to reduce the amount of computation of the convolution function in the model, thus reducing the number of parameters in the model. The MobileNet [20] model reduces the model parameters by decomposing a normal convolution into a deep convolution and a dot convolution, while maintaining the same performance as a normal convolution. The Lite-HRNet [21] model uses the method of performing information exchange across channels to maintain the information exchange between channels, in place of the expensive ordinary convolutional computation.

To address the problems that arise in the task of human pose estimation, Artacho et al [22] utilised a multi-scale feature representation to improve the effectiveness of keypoint feature extraction without significantly increasing the model parameters. Tang et al [23] proposed a new spatio-temporal longitudinal and transversal attention module to reduce the computational effort of the model by decomposing the joints feature matrix in both spatial and temporal dimensions. Zhao et al [24] addressed the problem of increasing the computational burden by increasing the size of the input sequences to enhance the performance of the model by using a compact representation of long skeleton sequences in the frequency domain to efficiently expand the receptive field and improve the robustness to 2D noisy pose detection. Liu et al [25] proposed limb orientation cue-aware networks to prevent overfitting of the depth network leading to uncertain keypoint locations. Yang et al [26] proposed a two-stage pose distillation model for whole-body pose estimation to address the problem of varying body part scales in order to improve the validity and efficiency of the model. Lee et al [27] designed a pose estimation model with self-training loss using pose-aware confidence in semi-supervised and unsupervised pose estimation tasks. In this paper, the lightweight Lite-HRNet is used as the backbone network. Design of deformable convolutional encoding networks and attention mechanism based channel and spatial attention modules to enhance the model ability to extract key point feature. Allow model performance to be efficiently improved without significantly increasing the computational and parametric count of the model. Thereby the model can be more effectively applied to detection tasks in different scenarios.

III. Proposed Method

The Adaptive Target Region Attention Network is designed with two main modules: the DCEN module, the CASA module, and the overall network framework is shown in Fig. 1. Firstly, the visual feature $\{F_{t-2}, F_{t-1}, F_t\}$ of the input sequences are extracted by the backbone network Lite-HRNet-18 network, and they are input into the DCEN module to calculate the information difference between the background feature and the subject feature, and to obtain the target region attention feature M_t. Then, the target region attention feature M_t are mined for channel and spatial information by CASA module to get the target region focus attention feature $M_t^{'}$. Finally, the combination of feature $M_t^{'}$ and visual feature F_t generates enhanced visual feature $F_t^{'}$. The feature $F_t^{'}$ are input into the pose estimation detection head to obtain the keypoint detection heatmap H_t. In the following, each module will be explained in detail.

A. Deformable Convolutional Encoding Network

The Deformable Convolutional Encoding Network is divided into three main steps: (1) Stage Feature Sequence Acquisition, which inputs the image sequence X_t into the Lite-HRNet network to obtain visual feature $\{F_{t-2}, F_{t-1}, F_t\}$. (2) Feature Sequence Fusion. The global visual feature S_t is obtained through a convolutional encoding network on the reconstructive encoding of the feature sequence. (3) Adaptive Target Region Attention Feature. Input the global visual feature S_t into the deformable convolutional network, use the deformable convolution to calculate the region information in the feature map, capture the visual feature of the target region, and reduce the influence of background feature and noise feature on the target region feature. The target region attention feature M_t is obtained through the DCEN module, which is used to pay attention to and capture the feature information of the target region in image. The DCEN module is shown in Fig. 2.

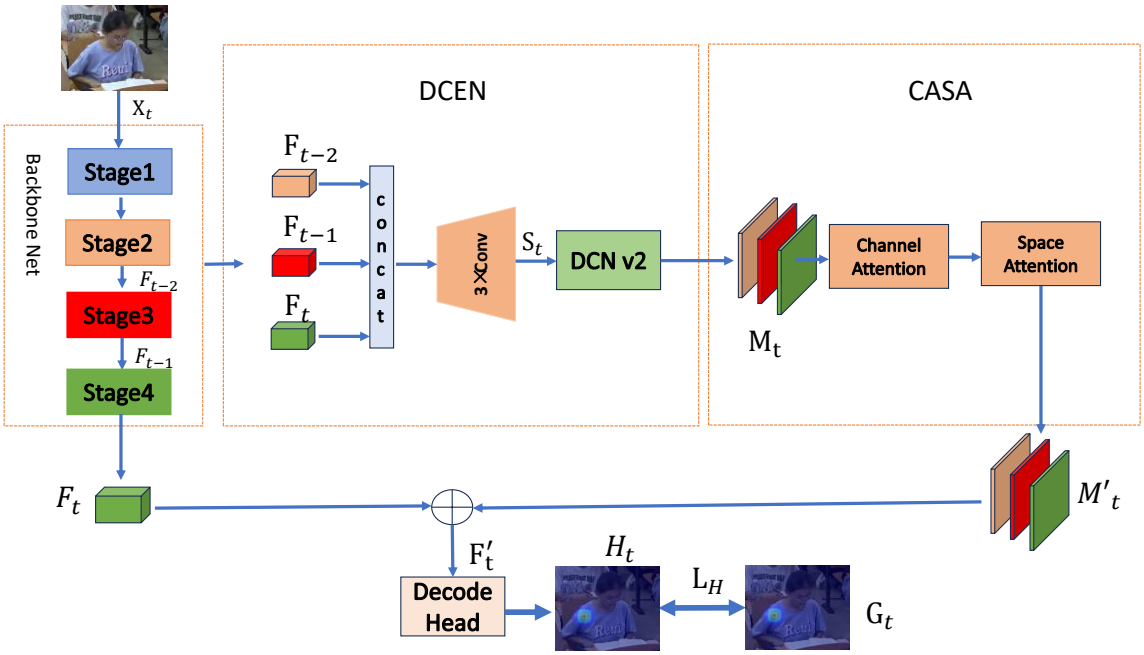

Fig. 1. Overall network framework.

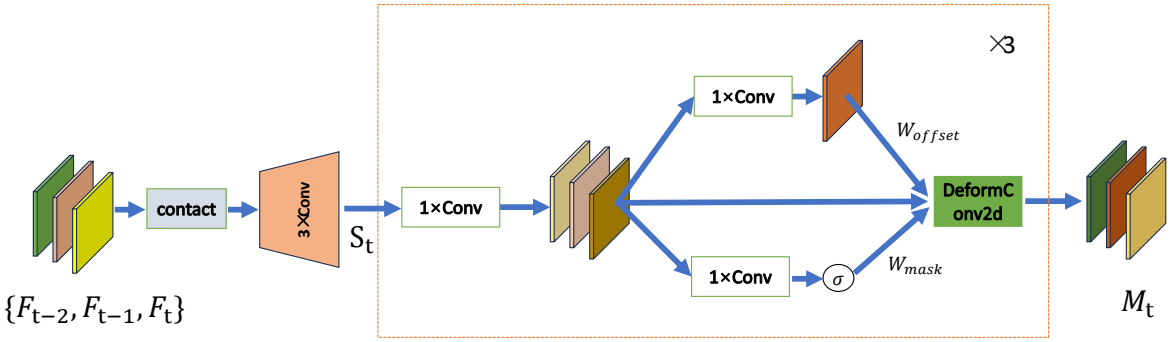

Fig. 2. DCEN module.

Stage Feature Sequence Acquisition: visual feature is extracted by Lite-HRNet network. Lite-HRNet replaces the expensive 1×1 convolution in the shuffle block [28] with a lightweight conditional channel weighting module, allowing the model to maintain efficient performance while reducing the computational effort of the network. Since the computational complexity is linear, it is lower than the quadratic time complexity of point-by-point convolution. Input the image sequence X_t into the Lite-HRNet network and acquire the visual feature $\{F_{t-2}, F_{t-1}, F_t\}$ of the three stages in the Lite-HRNet network. Where F_{t-2}, F_{t-1}, F_t is the output visual feature of the second, third, and fourth stages of the Lite-HRNet network.

Feature Sequence Fusion: multiple stages of visual feature of different coarseness were obtained from the Lite-HRNet network. They possess semantic information about visual feature in different depths. In order to better utilise the semantic information of these visual feature, an up-sampling approach is used to reconstruct and encode the different stages of the visual feature into a fusion that increases the resolution of the feature sequence and enhances the retention of edge information. And combining their shallow and deep visual feature to generate the global visual feature S_t with global visual information and more fine-grained. The operation is shown in Eq. (1):

$$S_t = Conv(F_t \oplus F_{t-1} \oplus F_{t-2}) \quad (1)$$

Where is a network of 3 convolutional layers.

Adaptive Target Region Attention Feature: Input the global visual feature S_t into the deformable convolutional network [29], and use the deformable convolution to adaptively capture the regional information of the feature map to obtain the target region attention feature M_t. Firstly, the global visual feature S_t are used to compute a trainable parameter W_{offset}, which is used to supervise the adaptive region orientation of the deformable convolution. The operations are shown in Eq. (2), (3), and (4):

$$S_t^{'} = \frac{W_i * S_t - E(W_i * S_t)}{\sqrt{Var(W_i * S_t) + \varepsilon}} \quad (2)$$

$$M_{t-1} = \frac{S_t^{'}}{1+e^{-S_t^{'}}} \quad (3)$$

$$W_{offset} = M_{t-1} * W_i \quad (4)$$

Where W_{offset} is a feature map with directional shifts, which serves to compute the shifts in the x and y directions of the input feature, W_i convolutional weights, $Var()$ is the averaging function, $E()$ is the expectation function, ε is an offset constant.

Then, the penalty weight parameter W_{mask} is added for guiding the training of the network and speeding up the convergence of the deformable convolutional network. As shown in Eq. (5):

$$W_{mask} = (1+e^{-W_i * M_{t-1}})^{-1} \quad (5)$$

Finally, the input feature vector M_{t-1} with an offset parameter W_{offset} with a penalty weight W_{mask} is input into the deformable convolution function for deformable convolution operation to compute the target region attention feature M_t. The operation is shown in Eq. (6), (7), and (8):

$$M_{t1} = f(W_{mask}, M_{t-1}, W_{offset}) \quad (6)$$

$$M_{t2} = \frac{M_{t1} - E(M_{t1})}{\sqrt{Var(M_{t1}) + \varepsilon}} \quad (7)$$

$$M_t = (1+e^{M_{t2}})^{-1} \quad (8)$$

Where $f()$ is a deformable convolution function, W_{mask} is a penalty weight parameter, W_{offset} is an offset parameter. $Var()$ is the averaging function, $E()$ is the desired function, ε is an offset constant.

B. Channel And Spatial Attention Module

The target region attention feature M_t obtained through the DCEN module is still susceptible to mission-independent noise signals such as background, occlusion, and overlap. By using the CASA module based on the attention mechanism [30] to mine the feature information of the channel and space of the target region attention feature M_t. Allow the feature to gain supervision and attention in channel and space. The designed CASA module has two sub-modules which are used to focus on the valid information of the feature vectors in channel and space respectively and to enhance the characteristics of the valid information as well as to suppress the noisy information. The architecture of the CASA module is shown in Fig. 3.

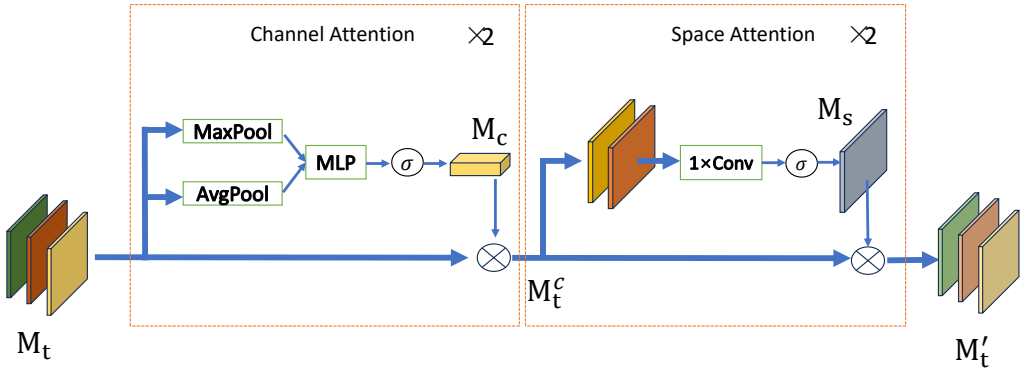

Fig. 3. CASA module.

Feature Sequence Channel Attention: Firstly, the information of each channel of target region attention feature M_t is aggregated by global maximum pooling and global average pooling. They are used to preserve the most significant feature in each channel and the overall average feature. The two feature information is then optimised using an MLP fully connected network trained to obtain parameter M_c with feature information for each channel, Finally, the multiplication operation of M_c with M_t makes the feature M_t aggregate the useful feature information of each channel. The implementation of the channel attention module is shown in Eq. (9) and (10):

$$M_c = \sigma(F(Avg(M_t)) + F(Max(M_t))) \quad (9)$$

$$M_c^t = M_t \otimes M_c \quad (10)$$

Where σ is the *sigmoid* activation function. $Avg()$, $Max()$ are the global average pooling and global maximum pooling functions, $F()$ is a three-layer fully connected network.

Feature Sequences Spatial Attention: The approach to spatial attention is similar to that of channel attention, where global maximum pooling and global average pooling are used for different dimensions to aggregate the feature information of feature M_t^c Firstly, instead of aggregating the feature information of each channel, spatial attention aggregates the feature information of all channels together, making each channel spatially connected to each other. The two features information is then optimised using a convolutional network trained to obtain parameter M_s with spatial correlation information for each channel feature. Finally, the

multiplication operation of M_s and M_t^c is performed to obtain the target region focus attention feature $M_t^{'}$ with channel and spatial attention information. The spatial attention module is implemented as in Eq. (11) and (12):

$$M_s = \sigma(W_i * (Avg(M_t^c), Max(M_t^c))) \quad (11)$$

$$M_t^{'} = M_s \otimes M_t^c \quad (12)$$

Where σ is the *sigmoid* activation function, W_i is the convolution weight, and the convolution kernel size is 3×3.

C. Loss Function

Combine the target region focus attention feature $M_t^{'}$ with the visual feature F_t to obtain the enhanced visual feature $F_t^{'}$. Input to the detection head generates a pose point heatmap H_t. The detection head module is implemented by a convolutional network. The loss function uses the L_H loss of the heat map of standard attitude estimation to supervise the attitude estimation model. The operation is shown in Eq. (13), (14), and (15):

$$F_t^{'} = F_t + M_t^{'} \quad (13)$$

$$H_t = W_i * F_t^{'} \quad (14)$$

$$L_H = \| H_t - G_t \|^2 \quad (15)$$

Where H_t and G_t denote the predicted and real attitude thermograms, W_i is the convolution weight.

IV. Experiments

Two pose estimation datasets: the COCO dataset and the Self-Constructed Smart Classroom dataset (SC-Data) are used in the experiments to evaluate the effectiveness of the models, and the results of comparisons with other mainstream human pose estimation models in both datasets are reported. Also, extensive ablation experiments are conducted to validate the effectiveness of the module proposed in this paper.

A. Introduction to the Dataset

SC-Data dataset: SC-Data dataset is a dataset made based on real classroom teaching data, which has 6,000 images and 16,800 instances of student pose data. There are 14000 instances in the training set and 2800 instances in the test set. The SC-Data dataset is made from one semester's worth of student classroom data and contains information about the student's classroom postures over the course of a semester. This will provide data to understand the complete pose information of students in a particular subject and provide a more accurate source of dataset for obtaining student pose information on teaching.

COCO 2017 dataset: COCO has over 200000 images and 250000 person instances with 17 keypoints, train2017 dataset (includes 57000 images 150000 person instances), val2017 (includes 5000 images).

B. Experimental Setup

In this paper, the network is trained using 1 NVIDIA A100 GPU, the optimisation algorithm is Adam with an initial learning rate of 0.0002 and a batch size of 64, the input to the network is an image with a fixed 4:3 aspect ratio, cropped from the original and resized to 256×192, the model is implemented in the PyTorch framework. In the pose evaluation metrics, the model evaluated using mean accuracy (mAP), the AP is first calculated for each joint and then the final performance (mAP) is obtained by averaging over all joints. The criterion is based on the metrics of the COCO dataset pose estimation.

C. Experiment Results and Analyses on the COCO Dataset

The models in this paper were evaluated on the COCO dataset and the performance of the comparison models on the COCO test set is shown in Table I. The human pose estimation performance of this paper model reaches 67.5(mAP). Compared to Small HRNet-W16 and Lite-HRNet-18 the gain is improved by 12.3(mAP) and 2.7(mAP). Compared to Lite-HRNet-30 the performance is improved by 0.3(mAP), but the model parameters decreases by 0.3(M). Compared to Integral Pose Regression [31] and G-RMI [32], which are computationally and parameter intensive, the model in this paper achieves quite good performance, but there is a substantial reduction in model complexity and number of parameters. The results of comparing the computational complexity of this paper model with other models are shown in Fig. 4, where the GFLOPs decrease by 0.41(GFLOPs) compared to ShuffleNetV2, while at the same time, the performance has a 7.6(mAP) improvement. Compared to Lite-HRNet-18 the GFLOPs increase by 0.52(GFLOPs), but have a 2.7(mAP) performance improvement.

TABLE I. Comparison Models on the COCO Test Set

Model	AP(mAP)	AP50	AR	Params
Our	**67.5**	88.2	67.0	1.4M
Lite-HRNet-18[21]	64.8	87.3	65.6	1.1M
Lite-HRNet-30	67.2	88.0	73.3	1.8M
Small HRNet-W16	55.2	83.7	62.1	1.3M
G-RMI[32]	64.9	85.5	69.7	57M
MobileNetV2[20]	64.6	87.4	70.7	9.6M
ShuffleNetV2[28]	59.9	85.4	66.4	7.6M
Integral Pose Regression[31]	67.8	88.2	-	45.0M
DY-MobileNetV2[33]	68.2	88.4	74.7	16.1M
DY-ReLU[34]	68.1	88.5	-	9.0M
LitePose-XS[18]	49.5	74.5	-	1.7M

TABLE II. COMPARISON OF SC-DATA DATASET

Model	AP(mAP)	AP50	AR	Params
Our	**86.6**	98.9	89.9	1.4M
Lite-HRNet-18	85.0	96.6	88.7	1.1M
Naive Lite-HRNet-18	85.3	96.7	88.9	1.4M
Lite-HRNet-30	85.4	98.5	88.9	1.8M

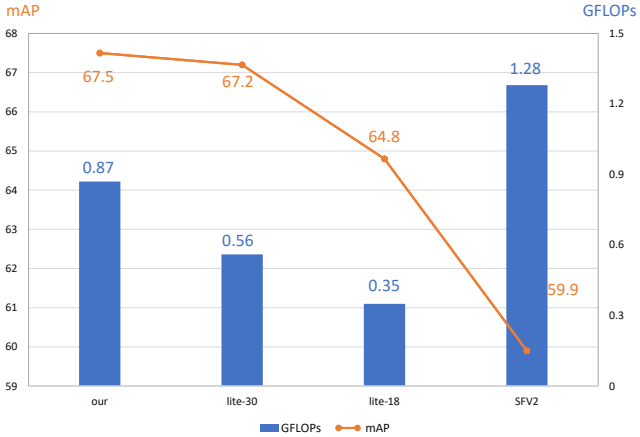

Fig. 4. Comparison of computational complexity and accuracy of the COCO dataset.

D. Experimental Results and Analyses on the SC-Data Dataset

Evaluating the method of this paper on the SC-Data dataset of this paper and comparing the performance of the model algorithm on the validation set is shown in Table II, and the experimental inference results of the model of this paper are shown in Fig. 5. The performance of the model in this paper on the SC-Data dataset, reaches 86.6(mAP). Compared to Lite-HRNet-18 and Lite-HRNet-30 with a gain of 1.6(mAP) and 1.2(mAP) points respectively. The model in this paper maintains the efficient performance while the model complexity is also reduced by 0.3(M) relative to Lite-HRNet-30. By comparing the experiments on the two datasets, this makes the model application scenarios of this paper richer, and at the same time, the lightweight human pose estimation model in this paper is easy to deploy to smart classroom scenarios.

In the scenario of occlusion and overlap in the smart classroom, the inference performance comparison is carried out by comparing with other models, and the comparison results are shown in Fig. 6, Fig. 6(1) shows the model of this paper, and Fig. 6(2) shows the Lite-HRNet18. By comparing with other models, it can be concluded that in the estimation of the gesture of the students of the target, the model of this paper can reduce the problems caused by problems such as the occlusion or overlap between the student The detection errors of the pose points are shown in the red circles marked in Fig. 6(2). The experimental results show that the model in this paper can effectively reduce the error detection of not the same target in smart classroom scenarios, and has better performance for scenarios with occlusion and overlap.

Fig. 5. Smart classroom pose estimation results.

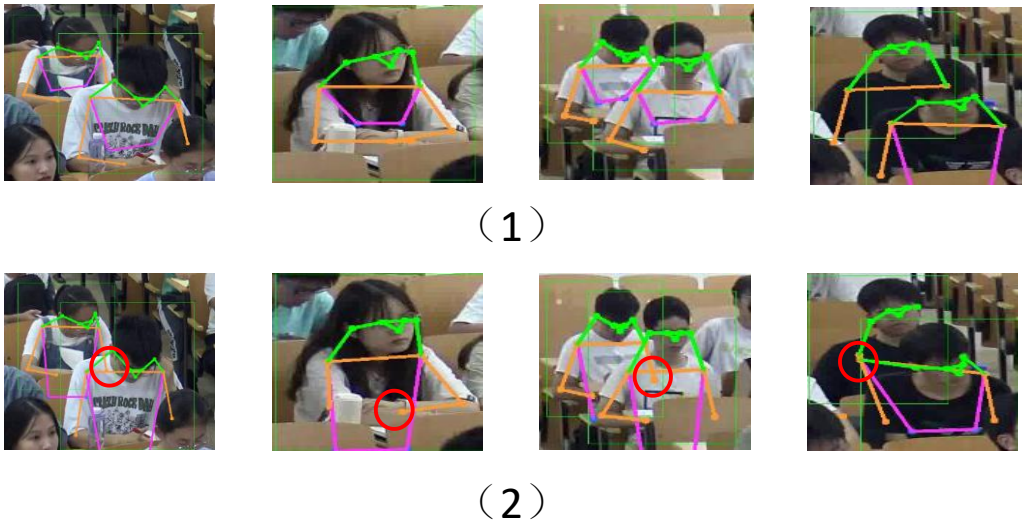

Fig. 6. Estimation results of student pose for occluded scenes in smart classroom.

TABLE III. ABLATION EXPERIMENTS WITH ADDED MODULES

Lite-HRNet	DCEN	CASA	COCO(mAP)	SC-Data (mAP)
√		√	65.7	85.6
√	√		66.6	85.7
√	√	√	67.5	86.6

E. Ablation Study

An ablation experiment was designed to add the DCEN module and the CASA module so as to verify the contribution of each module in the network, as shown in Table III, where "√" represents the addition of this module in the network. In the SC-Data dataset, the DCEN module provided a performance gain of 0.7(mAP)(85.0 → 85.7) compared to the Lite-HRNet network. The CASA module provided a performance gain of 0.6(mAP)(85.0 → 85.6). The performance enhancement of the model in the network where the DCEN module is added together with the CASA module is better compared to one added module alone, which improves the model with a performance gain of 1.6(mAP)(85.0 → 86.6). The significant performance gain indicates that the modules proposed in this paper play an important role in extracting feature information from the target region. In the COCO dataset, the DCEN module can improve the performance gain of the model by 1.8(mAP)(64.8 → 66.6), and the CASA module can improve the performance gain of the model by 0.9(mAP)(64.8 → 65.7). The ablation experiments verify that the modules in this paper can be applied to different scenarios and effectiveness.

V. CONCLUSION

This paper addresses the task of student pose estimation in smart classrooms by proposing a lightweight human pose estimation model with Adaptive Target Region Attention Network. Firstly, this paper proposes the deformable convolution-based target region attention module (DCEN) to capture student subject region representations. Secondly, in order to further obtain more precise attention to the target region, the channel and spatial attention module (CASA) is proposed to attend to the information about the relevant tasks on the space and channels of the feature map. Finally, a large number of experiments show that the model has excellent performance on both the COCO dataset and the homemade smart classroom dataset (SC-Data), while the number of parameters in this paper model has been greatly reduced and the detection speed has been greatly improved compared to the human pose estimation model with a large amount of computation and parameters. In future work, the paper will focus on deploying the pose estimation model to the classroom and applying the acquired student pose information to the task of assessing teaching quality.

ACKNOWLEDGMENT

This research was funded by The National Natural Science Foundation of China (62001133, 62177012, 61967005). The Fund of Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing, Guilin University of Electronic Technology, (No. GXKL06200114). This research was supported by Guangxi Natural Science Foundation under Grant No. 2024GXNSFDA999015.

REFERENCES

[1] Wang, Jinbao, et al. "Deep 3D human pose estimation: A review." Computer Vision and Image Understanding 210 (2021): 103225.

[2] Zheng, Ce, et al. "Deep learning-based human pose estimation: A survey." ACM Computing Surveys 56.1 (2023): 1-37.

[3] Liu, Wu, et al. "Recent advances of monocular 2d and 3d human pose estimation: a deep learning perspective." ACM Computing Surveys 55.4 (2022): 1-41.

[4] Song, Liangchen, et al. "Human pose estimation and its application to action recognition: A survey." Journal of Visual Communication and Image Representation 76 (2021): 103055.

[5] Lin, Feng-Cheng, et al. "Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection." Sensors 21.16 (2021): 5314.

[6] Liu, Hai, et al. "Arhpe: Asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction." IEEE Transactions on Industrial Informatics 18.10 (2022): 7107-7117.

[7] Liu, Tingting, et al. "GMDL: Toward precise head pose estimation via Gaussian mixed distribution learning for students' attention understanding." Infrared Physics & Technology 122 (2022): 104099.

[8] Deng, Chao, Jiao Peng, and ShuFei Li. "Research on the state of blended learning among college students–A mixed-method approach." Frontiers in Psychology 13 (2022): 1054137.

[9] Alfoudari, Aisha M., Christopher M. Durugbo, and Fairouz M. Aldhmour. "Understanding socio-technological challenges of smart classrooms using a systematic review." Computers & Education 173 (2021): 104282.

[10] Kaur, Avneet, Munish Bhatia, and Giovanni Stea. "A survey of smart classroom literature." Education Sciences 12.2 (2022): 86.

[11] Wang, Jingxian, et al. "Teacher beliefs, classroom process quality, and student engagement in the smart classroom learning environment: A multilevel analysis." Computers & Education 183 (2022): 104501.

[12] Du, Congju, Han Yu, and Li Yu. "A scale-sensitive heatmap representation for multi-person pose estimation." IET Image Processing 16.4 (2022): 1194-1207.

[13] Luo, Zhengxiong, et al. "Rethinking the heatmap regression for bottom-up human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[14] Xu, Xixia, et al. "Location-free human pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[15] Feng, Runyang, et al. "Mutual information-based temporal difference learning for human pose estimation in video." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[16] Khirodkar, Rawal, et al. "Multi-instance pose networks: Rethinking top-down pose estimation." Proceedings of the IEEE/CVF International conference on computer vision. 2021.

[17] Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[18] Wang, Yihan, et al. "Lite pose: Efficient architecture design for 2d human pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[19] Sun, Ke, et al. "Igcv3: Interleaved low-rank group convolutions for efficient deep neural networks." arXiv preprint arXiv:1806.00178 (2018).

[20] Howard, Andrew, et al. "Searching for mobilenetv3." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[21] Yu, Changqian, et al. "Lite-hrnet: A lightweight high-resolution network." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[22] Artacho, Bruno, and Andreas Savakis. "Unipose+: A unified framework for 2d and 3d human pose estimation in images and videos." IEEE transactions on pattern analysis and machine intelligence 44.12 (2021): 9641-9653.

[23] Tang, Zhenhua, et al. "3D human pose estimation with spatio-temporal criss-cross attention." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[24] Zhao, Qitao, et al. "Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[25] Liu, Tingting, et al. "LDCNet: limb direction cues-aware network for flexible human pose estimation in industrial behavioral biometrics systems." IEEE Transactions on Industrial Informatics (2023).

[26] Yang, Zhendong, et al. "Effective whole-body pose estimation with two-stages distillation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[27] Lee, Taeyeop, et al. "Tta-cope: Test-time adaptation for category-level object pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[28] Ma, Ningning, et al. "Shufflenet v2: Practical guidelines for efficient cnn architecture design." Proceedings of the European conference on computer vision (ECCV). 2018.

[29] Zhu, Xizhou, et al. "Deformable convnets v2: More deformable, better results." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[30] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV). 2018.

[31] Sun, Xiao, et al. "Integral human pose regression." Proceedings of the European conference on computer vision (ECCV). 2018.

[32] Papandreou, George, et al. "Towards accurate multi-person pose estimation in the wild." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[33] Chen, Yinpeng, et al. "Dynamic convolution: Attention over convolution kernels." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[34] Chen, Yinpeng, et al. "Dynamic relu." European Conference on Computer Vision. Cham: Springer International Publishing, 2020.