ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20210922053009
Original scientific paper

Establishing a Fusion Model of Attention Mechanism and Generative Adversarial Network to Estimate Students' Attitudes in English Classes

Tong ZHAO*, Tianyi SONG

Abstract: With the rapid development of science and technology, artificial intelligence has been widely used in various fields and a new model of AI-aided education has been developed in the new era. In the education industry, AI-aided education can save teachers' energy, improve teaching efficiency and help to refine teaching methods. In order to estimate students' attitudes towards English teachers' lectures, this paper proposed an AI-aided feedback system. In the constructed system, DG-Net was used to expand the data sets of students, and combined with Attention's Alphapose model to collect students' listening poses. The whole model provided feedback of students' listening postures in English speaking and listening classes, assisting teachers to estimate students' attitudes through data analysis and realizing AI-aided education in English classes.

Keywords: AI-aided education; Alphapose; attention; DG-Net; English class

1 INTRODUCTION

Traditional college English teaching is exam-oriented. Both the teachers and students focus on how to solve the writing and reading tasks in the exams, leaving little attention on the listening and speaking competence. Accordingly, the teaching methods are simple. Teachers explain language points, and students remember them mechanically. As a result, students' interest in English has gradually faded away, and the teaching result is compromised. That is because both the teachers and the students forget that English, as a language, is used to communicate. Students who have the ability to read and write in English do not necessarily possess the ability to communicate with English. Cultivation of English listening and speaking ability is a way to improve students' communicative competence. Only by listening and speaking more, the students' language intuition can be gradually cultivated and their ability to communicate can be acquired. Therefore, listening and speaking practices have been brought into English classrooms.

However, managing an English listening and speaking class is a demanding task for the teachers. It is not possible for the teachers to hold an English listening and speaking activity, while monitoring every students' response and attitude or recording their preference of any particular teaching method. AI-aided education, an emerging classroom model, can help. This study used AI as auxiliary feedback to estimate students' attitudes by analyzing students' poses in an English listening and speaking classroom. With the application of attention mechanism and generative adversarial network, students' attitudes were obtained. The model refined the students' poses and made data analysis, avoiding the subjective factors of manual analysis. The analysis results are intuitive, visual and accurate.

2 RELATED RESEARCH

Researchers have developed some attention mechanism models. Jaderberg et al. developed Spatial Transformer Networks which transform spatial information through attention model to realize the functions of image rotation and zoom transformation [1]. Wang et al. believed that the Non-local method could capture the long-term dependence of one pixel on other pixels by calculating the autocorrelation matrix [2]. Du proposed Interaction-aware Attention model where a new loss function based on PCA was designed on the basis of non-local computation of covariance matrix to help achieve better global interaction between features in the channel dimension [3]. Huang developed CCNet from Non-local, which reduced the calculation of autocorrelation matrix from all pixels in the image to the pixels at the intersection, thus greatly reducing the calculation amount [4]. Similarly, in order to reduce the computation amount, Li et al. developed an EMA model, which combined the attention mechanism with EM algorithm to obtain a set of bases by expectation maximization and ran the attention mechanism on the set of bases [5]. Hu et al. and others put forward a brand-new strategy of "feature recalibration", SENET, which modeled the dependencies among feature channels, giving more weight to effective feature channels and ignoring invalid channels [6]. The attention mechanism proposed by Zhang realized self-adaptive adjustment of channel characteristics by calculating the dependencies among channels [7]. He suggested calculating the channel weights according to the activation values of the channels around the target position [8]. Yu et al. designed a smooth network to select more distinctive features through the channel attention block and the global average pool in order to solve the intra-class inconsistency in semantic segmentation [9]. Wang added channel domain attention to target tracking based on offline training, and through this mechanism, the channels with better tracking effect were given greater weight, and the noisy channels were directly deleted [10]. Zheng et al. give the current frame a weight by spatially measuring the relationship between the previous frame and the current frame, which is an essential channel domain attention mechanism [11]. Woo et al. put forward a convolutional block attention module [12]. In the processing of channel domain, this module was basically similar to SE-Net, obtaining the one-dimensional vector by channel compression first, and then operating the one-dimensional vector, while in space, the feature images obtained by Max

Pooling and Average Pooling were directly spliced together and then convolved. Cao et al. developed a Global Context (GC) block which integrated Non-local and SEnet [13]. After simplifying the foundation of Non-local, it integrated the attention of channel domain and realized a modeling mode without Query dependence. Fu et al. put forward Dual Attention Network, which is mainly a fusion variant of CBAM and non-local [14]. It used the idea of non-local in channel domain and spatial domain respectively, and utilized autocorrelation matrix to capture long-distance dependence. Finally, the attention output in channel domain and spatial domain was fused as the final output. The mixed domain attention mechanism (RANet) proposed by Wang et al. drew on the idea of residual network, constructed a residual attention network by stacking attention modules, constructed an identity map, and realized the training of deep residual attention network, which could be easily extended to hundreds of layers [15]. Huang improved the object detector by using an efficient fine-grained mechanism called Inverted Attention (IA) [16].

Generated adversarial network (GAN) was proposed by Goodfellow in 2014 [17]. With the high-definition images generated after the contest, it has attracted extensive attention from researchers. Early generation of GAN has unstable factors. In the iteration controlled by loss function, the loss value often does not decrease. To solve this problem, Radford et al. proposed the DCGAN network, in which the network structure was improved and a stable GAN network was obtained [18]. Arjovsky et al. analyzed the reasons why GAN was prone to collapse based on this theory, and proposed WGAN, which replaced the loss function of GAN with Wsserstein distance, and thus further improved the network performance [19]. DG-Net, different from the previous networks, does not need to use the information outside the data set to generate data, thus greatly improving the level of re-recognition baseline and making stable performance on many open source data sets.

Pose estimation is the basic challenge in computer vision. The initial single-person pose estimation employs the traditional tree model [20], the random forest [21] and the conditional random field model [22]. With the development of deep learning, the traditional method is far less accurate than the deep learning model. Therefore, DeepPose [23], DNN model [24] and CNN model [25] have been widely used. However, in single-person pose estimation, neither the traditional method nor the deep learning model can accurately estimate the pose unless the person is correctly positioned. This problem has been solved by a regional multi-person pose estimation (RMPE) framework [26], which increases the accuracy not only for single-person pose estimation but also for multi-person pose estimation.

3 METHOD

3.1 Attention Mechanism and Modules

Attention mechanism originated from the study of human vision. The research on human eyeballs has proved that only the fovea area of the retina has the greatest sensitivity. Therefore, in order to efficiently use the limited sensitive area of the eyeballs, people tend to choose the area that deserves the most attention and focus on it. This "weight-selective" processing mechanism of human body is called attention mechanism. Relying on this attention mechanism, people can deal with immeasurable information that they receive through vision in an orderly way. Similarly, in the era of information explosion, the field of deep learning has been bombarded by big data. In order to cope with excessive data input, the attention mechanism of human eyes has been applied to the field of deep learning, and has become the hottest data processing module in this field.

Attention mechanism is classified into Item-wise Soft Attention, Item-wise Hard Attention, Location-wise Soft Attention and Location-wise Hard Attention. Soft attention is static, focused on space and channel, while strong attention is dynamic, focused on the process involved. The mechanism of soft attention is subtle. If used in conjunction with an in-depth-learned network model, the weights of the network can be updated in the back propagation of the network model, therefore mainly used in that field of deep learning. According to different attention domains, soft attention can be divided into spatial attention, channel attention and fusion attention. Spatial attention generally involves channel compression on the input feature images at first, which can greatly reduce the number of parameters and simplify the calculation. Channel attention involves pooling the overall input feature image to get a one-dimensional vector, and carrying out feature interaction on this one-dimensional vector to determine the dependencies between channels because different channel attention mechanisms have different ways to deal with feature interaction. Fusion attention involves both spatial attention and channel attention. It conducts serial processing mode of spatial domain first and then channel domain, or channel domain first and then spatial domain, or parallel processing mode of spatial domain and channel domain.

The attention mechanism used in this study is the Convolutional Block Attention Module (CBAM), where spatial attention and channel attention were combined and coordinated in order improve the performance of the module.

3.1.1 Channel Attention Module

As Fig. 1 shows, the feature image with the size of $S \times B \times C$ was input into the channel attention module. The C channels of the feature image were averaged and pooled to obtain $1 \times 1 \times C$ feature vectors with the same number of dimension. These vectors passed through a two-layer neural network with shared weights, and the activation function of each layer was Relu. Two new feature vectors with the same feature dimension were added and activated by Sigmoid activation function to yield the weight coefficient (W_c). Finally, W_c was multiplied with the input to obtain the output of the module.

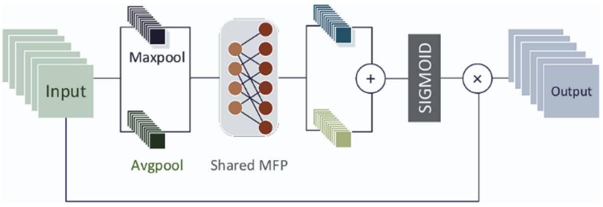


Figure 1 Channel attention module

3.1.2 Spatial Attention Module

As Fig. 2 shows, the feature image with the size of $S \times B \times C$ was input to the spacial attention module, and each pixel of the image was averaged and maximized on each dimension to obtain the feature vector with the same dimension of $S \times B$. The obtained image was convoluted and activated by Sigmoid to obtain W_s. Finally, the input was multiplied by W_s to get the module output.

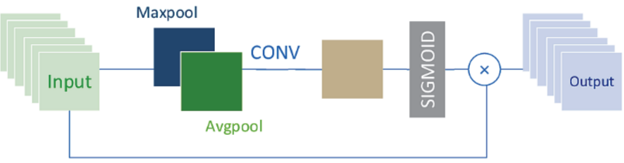


Figure 2 Spatial Attention Module

3.1.3 CBAM

As Fig. 3 shows, CBAM, a mixed domain attention mechanism, was divided into two independent modules: a channel attention module and a spatial attention module. This mixed mechanism saved model parameters and reduced the amount of calculation, making the use of CBAM more convenient. The formula is as follows:

$$W_c = \sigma\left\{\mathrm{MLP}\left[\mathrm{AvgPool}(I)\right] + \mathrm{MLP}\left[\mathrm{MaxPool}(I)\right]\right\} = \\ = \sigma\left\{W_1\left[W_0(I^{c}_{avg})\right] + W_1\left[W_0(I^{c}_{max})\right]\right\} \tag{1}$$

$$W_s = \sigma\left\{\mathrm{CONV}\left[\mathrm{AvgPool}(I),\mathrm{MaxPool}(I)\right]\right\} = \\ = \sigma\left\{\mathrm{CONV}\left[(I^{s}_{avg}; I^{s}_{max})\right]\right\} \tag{2}$$

$$O = A_s\left[A_c(I) \otimes I\right] \otimes I \tag{3}$$

$I \in R^{C \times S \times B}$ is the input of the module, $A_c(I) \in R^{C}$ is channel attention, $A_s(I) \in R^{S \times B}$ is spacial attention, W_c is weight coefficient of channel attention, W_s is weight coefficient of spacial attention, output is $O \in R^{C \times S \times B}$, W_0, W_1 is the weight coefficient of layer 1 and layer 2 of the shared network. $\otimes$ is the multiplication of matrix.

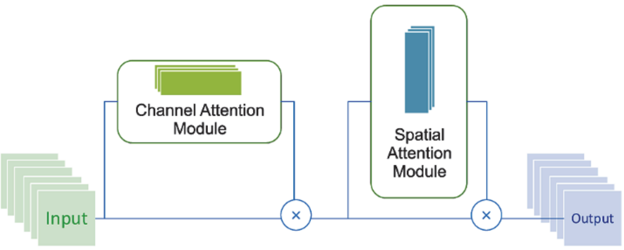


Figure 3 CBAM

3.2 Generative Adversarial Networks (GAN)

GAN consisted of a generator and a discriminator. By learning the features of real images in the data set, the generator generated "pseudo-images" with high similarity with the real ones, aiming to generate pictures that cannot be discriminated by the discriminator. The discriminator tried to distinguish the real images from the one generated by the generator. The two networks contested with each other to improve the realism of the generated data. Nowadays, DG-Net network, a particular kind of GAN, is widely used in various fields.

3.2.1 The Generator

The generation module of DG-Net consisted of self-identity generation and cross-identity generation. Self-identify generation occurred when two sample images of the same ID (the same person) passed through the generator (Fig. 4a). The two images were slightly different in characteristics like clothes, posture, and position. In that case, the generated images retained the same ID. Cross-identity generation occurred when two sample images of different ID (different people) passed through the generator (Fig. 4b). The generator exchanged the features of the samples, and provided the original ID with different features.

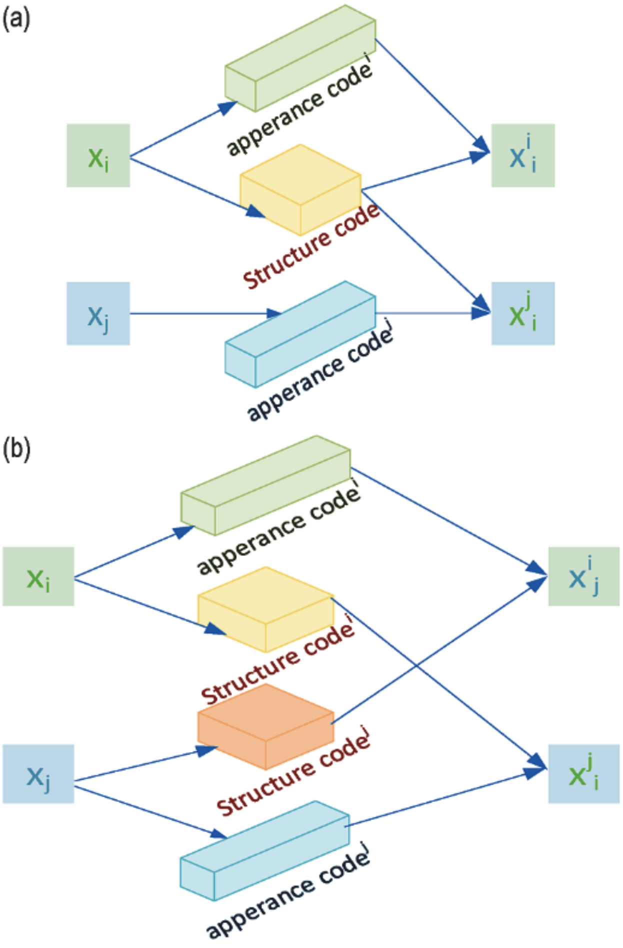


Figure 4 (a) self-identity generation; (b) cross-identity generation

3.2.2 The Discriminator

As Fig. 5 shows, the sample images of the discriminative learning module first passed through the appearance coder, and the primary feature was separated from the find-grained feature. The appearance coder then mapped the decomposed features to better predict the ID of the sample. The discrimination module used three losses to control the final result, namely ID loss, appearance loss and fine feature loss, which were calculated according to the following formula:

$$L_{\text{id}}^{S}=E\left\{-\log\left[p(y_i \mid x_i)\right]\right\} \tag{4}$$

$$L_{\text{prim}}=E\left[-\sum_{k=1}^{K} q(k \mid x_j^i)\log\left(\frac{p(k \mid x_j^i)}{q(k \mid x_j^i)}\right)\right] \tag{5}$$

$$L_{\text{fine}}=E\left\{-\log\left[p(y_{\text{j}} \mid x_j^i)\right]\right\} \tag{6}$$

where k is the digital code of the sample ID.

3.2.3 DG-Net

As Fig. 6 shows, DG-Net innovatively integrated the discriminator into the generator, shared the appearance encoder with the generation module, enabled online learning, and benefited the generator and the discriminator, thus forming a complete adversarial network framework. The inputs to the system are three images represented as X_j, X_i, and X_t, X_i and X_t have the same ID but different features, while X_j has a different ID. X_j passes the structure module only, X_t passes the application module only, and X_i passes both the structure and application modules. The IDs and the features are encoded and reconstructed to generate new picture data, which is feedback for appearance encoding online. After that, the improved appearance encoder feeds the reconstructed image and the real image together to discriminator for min-max contest. There were several losses in the whole network, seven in the generator and two in the discriminator. The whole network loss was calculated as follows:

$$\text{Lotol}(E_{\text{f}}, E_{\text{s}}, G, D)=\lambda_{\text{img}}L_{\text{img}}+L_{\text{code}}+L_{\text{id}}^{\text{s}}+\lambda_{\text{id}}L_{\text{id}}^{\text{c}}+ \\ +L_{\text{adv}}+\lambda_{\text{prinm}}L_{\text{prim}}+\lambda_{\text{fine}}L_{\text{fine}} \tag{7}$$

$L_{\text{img}}=L_1+L_2$ is the loss of the same ID image. $L_{\text{code}}=L_{\text{code1}}+L_{\text{code2}}$ is the loss of cross-identify generation. $\lambda_{\text{img}}, \lambda_{\text{id}}, \lambda_{\text{prim}}, \lambda_{\text{fine}}$ is the weight coefficient which controls the relation between different losses.

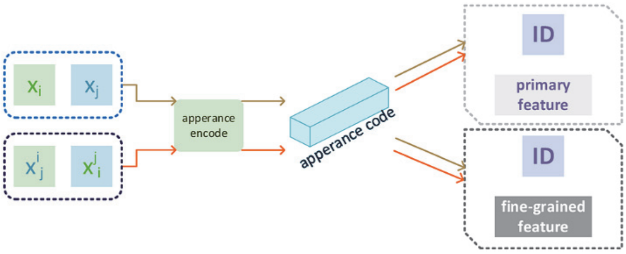


Figure 5 Discriminative RE-ID learning

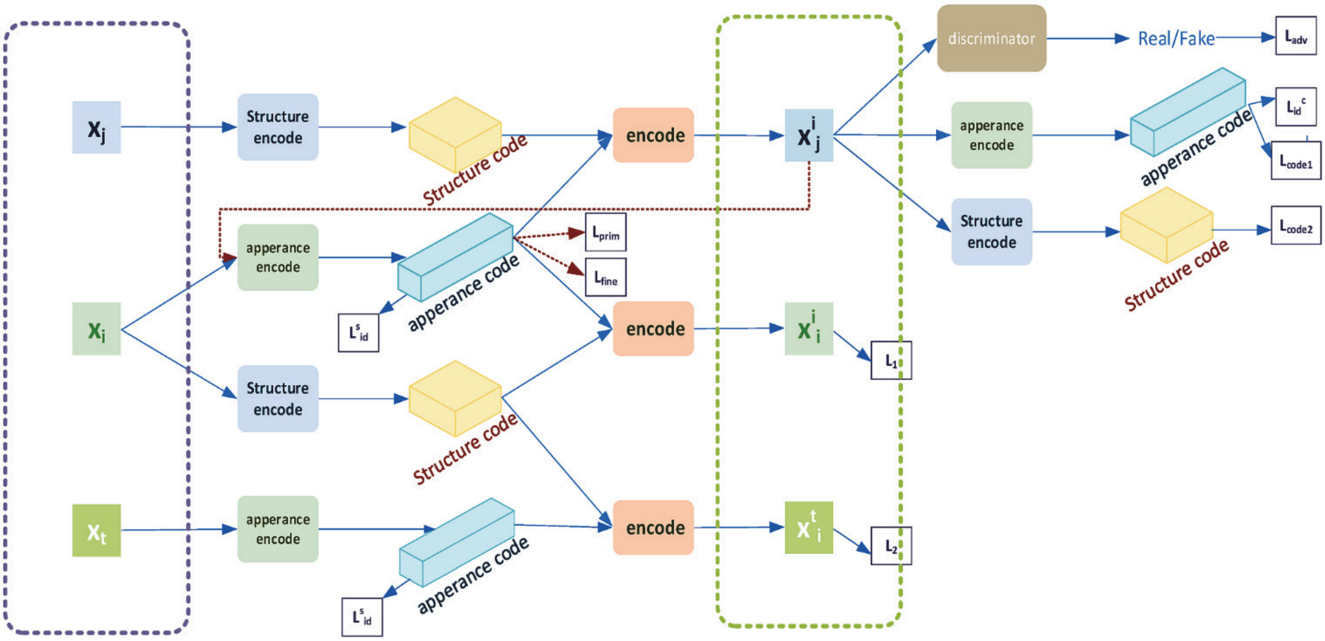


Figure 6 Infrastructure of DG-Net

3.3 Pose Analysis

There are two methods of human posture analysis nowadays: top-down method and bottom-up method. The former is to locate the positions of all the people first, frame the people then, and finally estimate the posture of the people in each frame one by one. The latter begins from locating all the joint points, connecting all the joint points to form the human body then, and finally estimating the posture. Every method has its own drawbacks. Results of the former depend exclusively on the location of the target human body, while the latter tend to confuse the joint points of the densely distributed human targets.

AlphaPose, with both speed and accuracy, is a top-down method, which mainly includes three modules of Symmetric STN+SPPE, Parametric pose Non-Maximum-Suppression (NMS) and pose-guided poses generator (PGPG). In this study, three technologies, namely SYMMETRIC SPATIAL TRANSFORMER NETWORK (SSTN), deep proposals generator (DPG) and PARAMETRIC POSE NON MAXIMUM SUPPRESSION (p-NMS) were used to solve the problem of multi-person posture estimation. SSTN was added to the structure of single person pose estimation (SPPE) to optimize the structure and extract high-quality human body regions despite the inaccurate location of human body frames. P-NMS was used to solve the redundancy of detection. The structure had its own pose estimation scheme to compare the similarity between the poses, optimizing the pose distance parameters with data-driven method. PGPG was used to enhance the training data and simulate the generation process of human body region frame by learning the description information of different poses in the output results.

3.3.1 Symmetric STN + SPPE

As Fig. 7 show, this module consisted of Spatial Transform Network (STN) [1], Spatial De-transform network and SPPE. This module estimated the pose of the incorrect input, and the pose estimation result was then mapped to the original image. In this way, the position of the input box was constantly adjusted until it became the precise input.

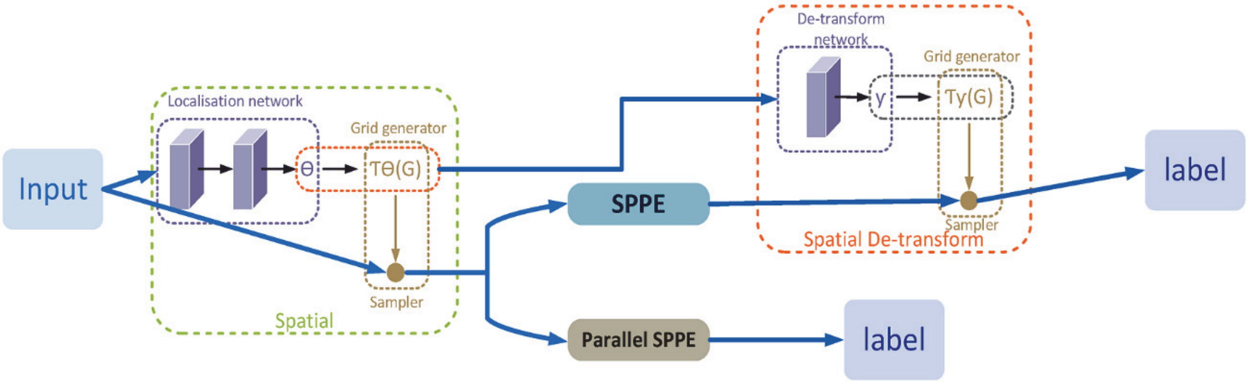


Figure 7 SSTN+SPPE module

3.3.2 NMS

When the people images were identified by the discriminator, a lot of potential windows of the pedestrian were obtained. Every window was scored. Since each person got multiple windows and most of windows had a high degree of overlap, Non-Maximum Suppression (NMS) was needed to select those windows with the highest scores and stop the windows with low scores. This module was used to stop the redundant boxes and accurately locate the human position.

3.3.3 AlphaPose

As Fig. 8 shows, PMPE module got the raw frame of the human position using the target detection algorithm, such as YOLO. Then the human posture was detected through STN+SPPE+SDTN module and removed of redundancy through NMS. The PGPG part of the module was finally used to enhance the trained data set.

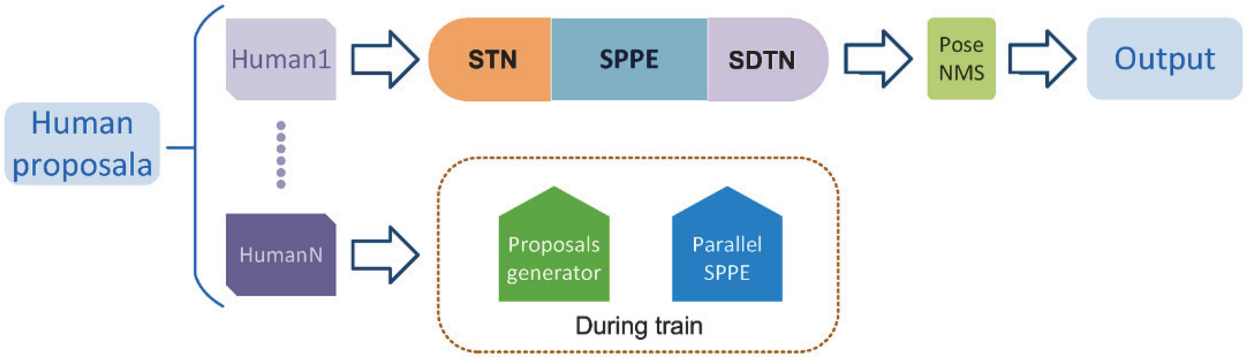


Figure 8 AlphaPose network

4 EXPERIMENT AND THE RESULTS

4.1 Experimental Environment and Data Set

Experimental data were obtained from the real pictures of students in some English classes. DG-Net was constructed on OMEN by HP Desktop PC 880-p1xx. The computer's processor was Intel(R) Core(TM) i7-9700 CPU, RAM is 16GB, equipped with NVIDIA GeForce RTX2080Ti GPU, and the running environment was python+pytorch. The alphapose was completed on linux server, a Ubuntu16.04 system, whose processor was Intel(R) Xeon(R) Silver 4110 CPU, equipped with four NVIDIA GeForce RTX 2080Ti GPU, and the running environment was docker+anaconda+python+pytorch.

4.2 Attention Mechanism and DG-Net Fusion Pose Estimation Model

As Fig. 9 shows, the module used alphapose to estimate students' pose. In order to better cope with the excessive number of students in the English class, attention module was added in the process of pose estimation. The attention module was embedded in Alphapose network, which reduced network parameters and time complexity. In the input data set, DG-Net network was applied to enhance the image data of English class, and the problem of small data set was solved.

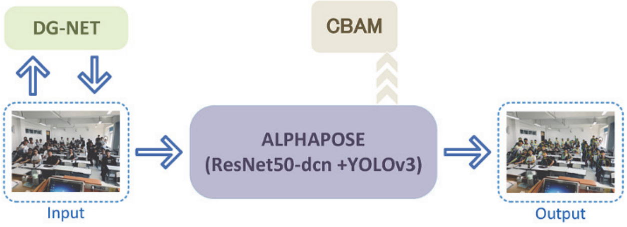
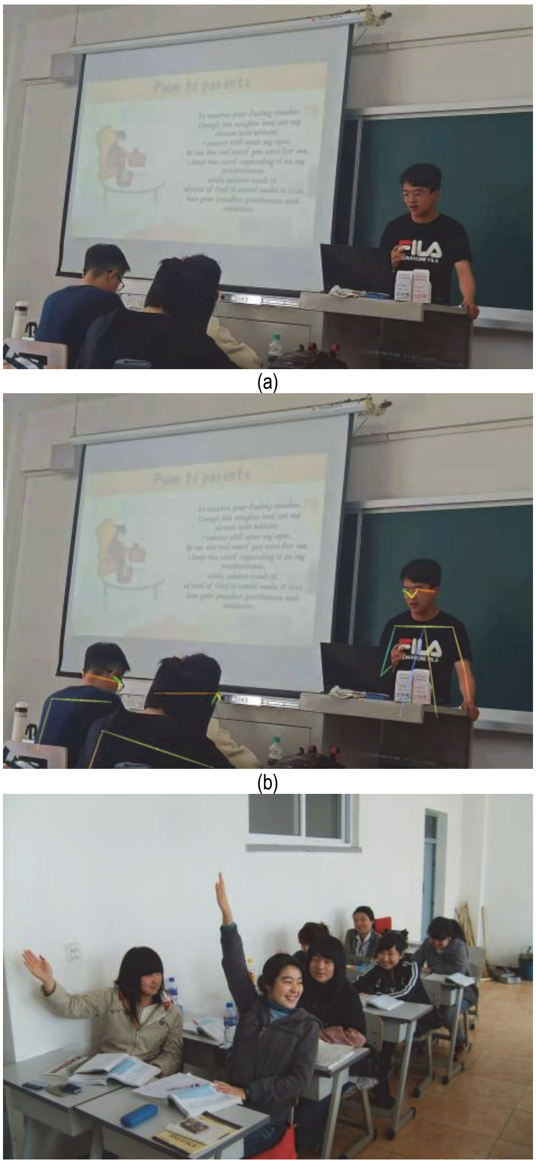


Figure 9 AAD network model

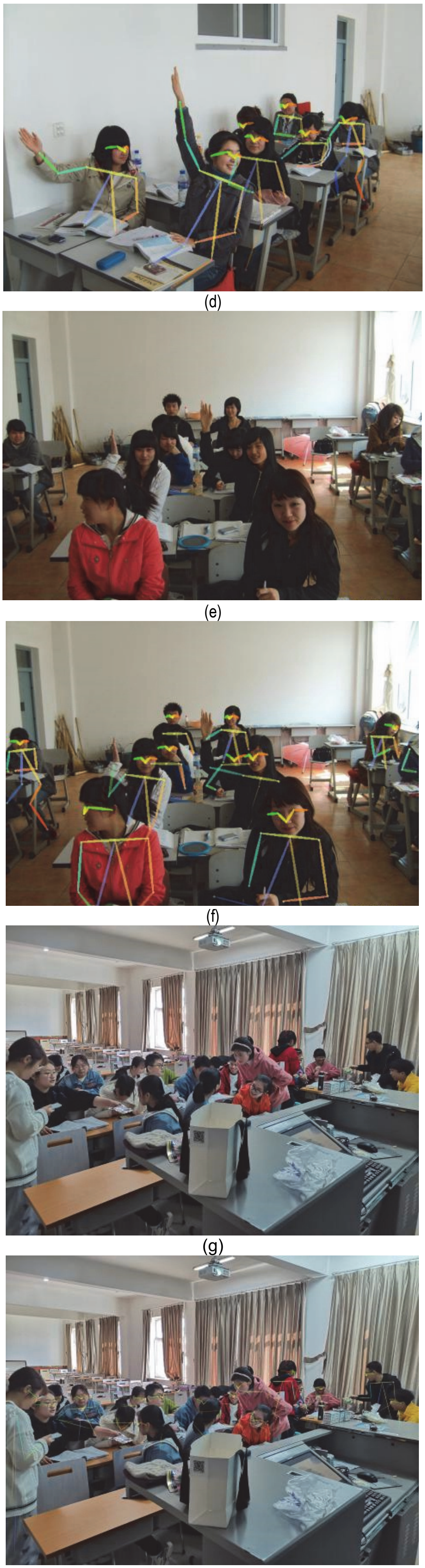
4.3 Results

As Fig. 10 shows, the network worked well. Although the input image data was a small sample and easy to cause an under-fitting situation, the addition of DG-Net expanded the input data set and enhanced the input image data, which makes the whole model fit better. Even when students' bodies overlapped or appeared sideways, the attitude information of each student could still be accurately obtained. CBAM module, integrated in Alphapose network, solved the problem of too large network and low speed, which provided speed support for the real-time monitoring system of the subsequent research.


(a)

(b)

(c)

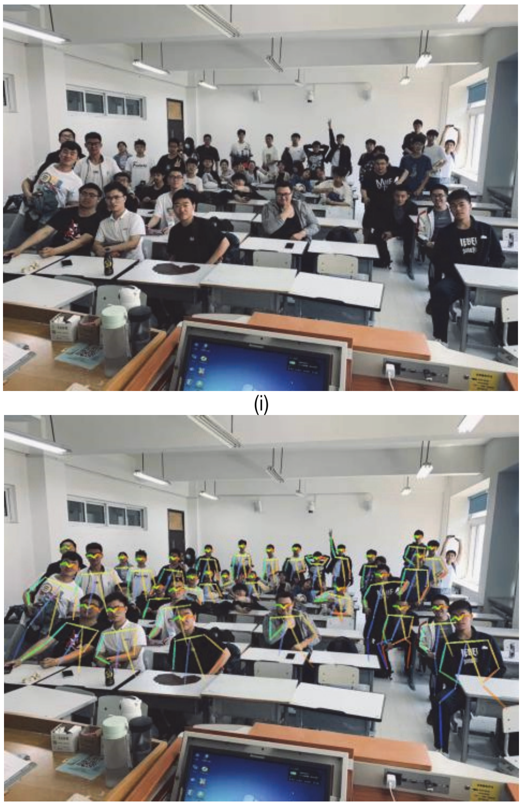

(d)

(e)

(f)

(g)

(h)

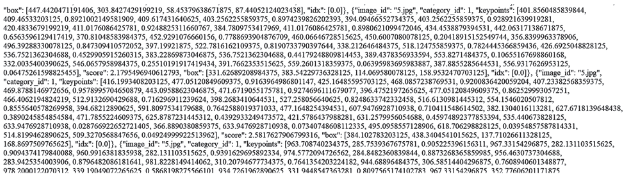


(i)

(j)

Figure 10 Students' Pose estimation (a), (c), (e), (g) and (i) are real images of the students; (b), (d), (f), (h) and (j) are pose estimation images

In addition to displaying the pose skeleton on the original image, the network web also generated the relevant json file (Fig. 11). File data, such as the ID of each student, the ID of every image, and the coordinates of each joint are listed in Tab. 1.



```
"box": [447.4420471191406, 303.8427429199219, 58.45379638671875, 87.44052124023438], "idx": [0.0]}, {"image_id": "5.jpg", "category_id": 1, "keypoints": [401.8560485839844,
409.46533203125, 0.8921002149581909, 409.617431640625, 403.2562255859375, 0.8974239826202393, 394.09466552734375, 403.2562255859375, 0.92892163991928l,
420.4833679199219, 411.0176086425781, 0.9248825311660767, 384.7809753417969, 411.0176086425781, 0.8980621099472046, 434.4538879394531, 442.06317138671875,
0.6563596129417419, 370.81048583984375, 452.9291076660156, 0.7788693904876709, 460.06646728515625, 450.6007080078125, 0.20418915152549744, 356.8399963378906,
496.39288330078125, 0.847309410572052, 397.19921875, 522.7816162109375, 0.8190733790397644, 338.212646484375, 518.124755859375, 0.7824445366859436, 426.6925048828125,
536.7521362304688, 0.4529905915260315, 383.22869873046875, 536.7521362304688, 0.44179248809814453, 389.4378356933594, 553.8271484375, 0.10655167698860168,
332.0035400390625, 546.0657958984375, 0.25510191917419434, 391.7662353515625, 559.2601318359375, 0.06395983695983887, 387.8855285644531, 556.9317626953125,
0.06475261598825455], "score": 2.1795496940612793, "box": [331.62689208984375, 383.54229736328125, 114.069580078125, 158.9532470703125], "idx": [0.0]}, {"image_id": "5.jpg",
"category_id": 1, "keypoints": [416.1993408203125, 477.05120849609375, 0.9163964986801147, 425.16485595703125, 468.0857238769531, 0.9200836420059204, 407.23382568359375,
469.8788146972656, 0.9578995704650879, 443.09588623046875, 471.6719055175781, 0.9274696111679077, 396.4752197265625, 477.05120849609375, 0.86252993057251,
466.4062194824219, 512.9132690429688, 0.716296911239624, 398.2683410644531, 527.258056640625, 0.8248633742332458, 516.6130981445312, 554.1546020507812,
0.8555640578269958, 394.68212890625, 591.8097534179688, 0.7642588019371033, 477.1648254394531, 607.9476928710938, 0.710411548614502, 382.1304016113281, 627.6718139648438,
0.3890245854854584, 471.7855224609375, 625.8787231445312, 0.4392933249473572, 421.5786437988281, 631.2579956054688, 0.45974892377853394, 535.440673828125,
633.9476928710938, 0.028766922652721405, 366.8890380859375, 633.9476928710938, 0.07340748608112335, 495.0958557128906, 618.706298828125, 0.03954857587814331,
514.8199462890625, 509.3270568847656, 0.04924999922513962], "score": 2.5817627906799316, "box": [384.102783203125, 438.3404541015625, 137.7102661132812 5,
168.8697509765625], "idx": [0.0]}, {"image_id": "5.jpg", "category_id": 1, "keypoints": [963.708740234375, 285.7539367675781, 0.905225396156311, 967.33154296875, 282.131103515625,
0.909437417984008 8, 960.9916381835938, 282.131103515625, 0.9391629695892334, 974.5772094726562, 284.8482360839844, 0.8873268365859985, 956.4630737304688,
283.9425354003906, 0.8796482086181641, 981.8228149414062, 310.20794677734375, 0.7641354203224182, 944.68896484375, 306.58514404296875, 0.7608940601348877,
978.2000122070312, 339.19049072265625, 0.5868198275566101, 934.7261962890625, 331.9448547363281, 0.8097565174102783, 967.33154296875, 352.77606201171875
```

Figure 11 json file

Table 1 Interpretation of json file elements

Element	Format	Interpretation
Image_id	int	id of the image
category_id	int	id of the person
keypoints	$[x, y, v] \times k$	The abscissa x, the ordinate y and the point mark v of the key point; k is the number of keypoints.
score	int	score
box	$[x, y, w, h]$	The abscissa x and ordinate y of the upper left corner of the frame, and the width w and height h of the frame.
idx	[0.0]	————————

There are 17 joint points generated by json. In the json file, all the students' posture skeleton information was digitally output. This enabled the manual check of the accuracy of students' posture estimation in unlabeled data sets and the alarm of the abnormal poses of the students.

5 CONCLUSION

This study proposes a fusion network to estimate the students' attitude in an English listening and speaking class. It is a new AI-aided language teaching method. The overall accuracy of the attitude estimation model is high even in the complex environment where the number of students is large and the body overlap is serious. This proposal is a feasible scheme for the future AI-aided education system. In this study, the deep learning method is innovatively added to the traditional education, so that teachers can capture students' attitude in class in a convenient, quick and accurate way, thus improving the educational methods. However, this model does not work well for the students at the edge of the picture or standing sideways. In that case, the students will be missed or the pose will be lost due to the low score in the NMS process. Therefore, the follow-up research will focus on solving these problems.

Acknowledgment

This study was supported by a grant from the Scientific Research Fund of the Educational Department of Liaoning Province (grant No: 2019WZD03).

6 REFERENCES

[1] Jaderberg, M., Simonyan, K., Zisserman, A. et al. (2015). Spatial Transformer Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017-2025.

[2] Wang, X., Girshick, R., Gupta, A. et al. (2018). Non-local Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7794-7803.

[3] Du, Y., Yuan, C., Li, B. et al. (2018). Interaction-aware Spatio-temporal Pyramid Attention Networks for Action Classification. *Proceedings of the European Conference on Computer Vision (ECCV)*, 373-389. https://doi.org/10.1007/978-3-030-01270-0_23

[4] Huang, Z., Wang, X., Huang, L. et al. (2019). CCNet: Criss-Cross Attention for Semantic Segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 603-612. https://doi.org/10.1109/ICCV.2019.00069

[5] Li, X., Zhong, Z., Wu, J. et al. (2019). Expectation-Maximization Attention Networks for Semantic Segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9167-9176. https://doi.org/10.1109/ICCV.2019.00926

[6] Hu, J., Shen, L., Sun, G., et al. (2018). Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132-7141. https://doi.org/10.1109/TPAMI.2019.2913372

[7] Zhang, Y., Li, K., Li, K. et al. (2018). Image Super-Resolution Using Very Deep Residual Channel Attention Networks. *Proceedings of the European Conference on Computer Vision (ECCV)*, 286-301. https://doi.org/10.1007/978-3-030-01234-2_18

[8] He, A., Luo, C., Tian, X. et al. (2018). A Twofold Siamese Network for Real-Time Object Tracking. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* https://doi.org/10.1109/cvpr.2018.00508

[9] Yu, C., Wang, J., Peng, C. et al. (2018). Learning a Discriminative Feature Network for Semantic Segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 325-341. https://doi.org/10.1109/CVPR.2018.00199

[10] Wang, Q., Teng, Z., Xing, J. et al. (2018). Learning Attentions: Residual Attentional Siamese Network for High Performance Online Visual Tracking. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* https://doi.org/cvpr. 2018.00510

[11] Zhu, Z., Wu, W., Zou, W. et al. (2018). End-to-End Flow Correlation Tracking with Spatial-Temporal Attention. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2018.00064

[12] Woo, S., Park, J., Lee, J. et al. (2018). CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision (ECCV)*, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[13] Cao, Y., Xu, J., Lin, S. et al. (2019). GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. https://doi.org/10.1109/iccvw.2019.00246

[14] Fu, J., Liu, J., Tian, H. et al. (2019). Dual Attention Network for Scene Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3146-3154. https://doi.org/10.1109/CVPR.2019.00326

[15] Wang, F., Jiang, M., Qian, C. et al. (2017). Residual Attention Network for Image Classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3156-3164. https://doi.org/10.1109/CVPR.2017.683

[16] Huang, Z., Ke, W., & Huang, D. (2020). Improving Object Detection with Inverted Attention.*2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. https://doi.org/10.1109/WACV45572.2020.9093507

[17] Goodfellow, I., Pouget-Abadie, J., Mirza, M. et al. (2014). Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 3, 2672-2680. https://doi.org/10.1145/3422622

[18] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *Computer Science*.

[19] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. *Computer Science*.

[20] Yang, W. & Mori, G. (2008). Multiple Tree Models for Occlusion and Spatial Constraints in Human Pose Estimation. *European Conference on Computer Vision (ECCV)*, *2*, 710-724. https://doi.org/10.1007/978-3-540-88690-7_53

[21] Dantone, M., Gall, J., Leistner, C. et al. (2013). Human Pose Estimation Using Body Parts Dependent Joint Regressors. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3041-3048. https://doi.org/10.1109/CVPR.2013.391

[22] Kiefel, M. & Gehler, P. (2014). Human Pose Estimation with Fields of Parts. *European Conference on Computer Vision (ECCV)*, 331-346. https://doi.org/10.1007/978-3-319-10602-1_22

[23] Toshev, A. & Szegedy, C. (2013). DeepPose: Human Pose Estimation via Deep Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2014.214

[24] Ouyang, W., Xiao, C., & Wang, X. (2014). Multi-source Deep Learning for Human Pose Estimation. *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2014.299

[25] Newell, A., Yang, K., & Jia, D. (2016). Stacked Hourglass Networks for Human Pose Estimation. *European Conference on Computer Vision (ECCV)*. https://doi.org/10.1007/978-3-319-46484-8_29

[26] Fang, H., Xie, S., Tai, Y. et al. (2017). RMPE: Regional Multi-person Pose Estimation. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2353-2362. https://doi.org/10.1109/ICCV.2017.256

Contact information:

Tong ZHAO
(Corresponding author)
Foreign Languages Department,
Shenyang Pharmaceutical University,
Shenyang 110016, China
E-mail: joanna_zt2021@163.com

Tianyi SONG
School of Information Science and Engineering,
Shenyang Ligong University,
Shenyang 110159, China