GestureDet: Real-time Student Gesture Analysis with Multi-dimensional Attention-based Detector

Rui Zheng, **Fei Jiang*** and **Ruimin Shen**
Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
zhengr, jiangf, rmshen@sjtu.edu.cn

Abstract

Students' gestures, hand-raising, stand-up, and sleeping, indicates the engagement of students in classrooms and partially reflects teaching quality. Therefore, fast and automatically recognizing these gestures are of great importance. Due to limited computational resources in primary and secondary schools, we propose a real-time student behavior detector based on light-weight MobileNetV2-SSD to reduce the dependency of GPUs. Firstly, we build a large-scale corpus from real schools to capture various behavior gestures. Based on such a corpus, we transfer the gesture recognition task into object detections. Secondly, we design a multi-dimensional attention-based detector, named GestureDet, for real-time and accurate gesture analysis. The multi-dimensional attention mechanisms simultaneously consider all the dimensions of the training set, aiming to pay more attention to discriminative features and samples that are important for the final performance. Specifically, the spatial attention is constructed with stacked dilated convolution layers to generate a soft and learnable mask for re-weighting foreground and background features; the channel attention introduces the context modeling and squeeze-and-excitation module to focus on discriminative features; the batch attention discriminates important samples with a new designed reweight strategy. Experimental results demonstrate the effectiveness and versatility of GestureDet, which achieves 75.2% mAP on real student behavior dataset, and 74.5% on public PASCAL VOC dataset at 20fps on embedding device Nvidia Jetson TX2.

1 Introduction

Student behaviors in real classrooms are an important part of teaching quality assessment. Previous student behaviors depend on the observations of teachers, which can hardly cover all the students in classrooms. In this paper, we focus on developing a real-time student behavior analysis system, aiming to automatically recognize student behaviors and assist in teaching quality evaluation.

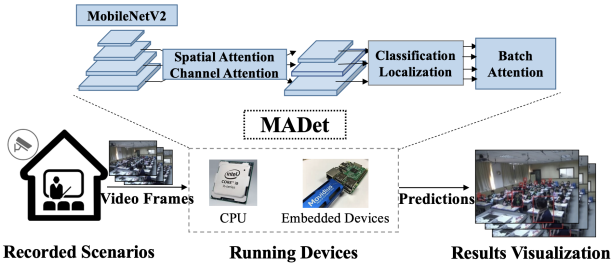


Figure 1: Overview of the real-time student behavior analysis system based on GestureDet. The whole system uses video frames from recorded scenarios as the input. Then our proposed GestureDet running on CPU or embedded devices outputs the detection results for further analysis and visualization. Particularly, our GestureDet is based on MobileNetV2 and three kinds of attention mechanisms (batch attention, channel attention, and spatial attention) are fused.

To capture various gestures of student behaviors, we build a large-scale corpus from 200 classrooms, 30+ schools, and change the behavior recognition tasks into object detections, where each behavior (hand-raising, stand-up and sleeping) corresponds to one object.

For object detection, CNN-based algorithms have achieved impressive results, which can be roughly divided into two categories: two-stage detectors [Girshick *et al.*, 2014; Ren *et al.*, 2015] and one-stage detectors [Liu *et al.*, 2016; Redmon *et al.*, 2016]. Two-stage detectors utilize Region Proposal Network (RPN) [Ren *et al.*, 2015] as the first stage to generate Region of Interests (RoIs) and these RoIs are further refined through the detection head for better accuracy. Although two-stage detectors have achieved state-of-the-art results on the public object detection benchmark [Everingham *et al.*, 2010; Lin *et al.*, 2014], they are too heavy for real-life computational-constrained scenarios. On the other hand, one-stage detectors directly predict bounding boxes and class probabilities, which usually involve less computation. For this reason, one-stage detectors are widely regarded as the key to real-time detection [Redmon *et al.*, 2016]. However, there are still accuracy gaps compared to two-stage detectors because of the large foreground-background imbalances and aligned feature representations in one-stage detectors [Lin *et*

* Corresponding author.

al., 2017; Chen *et al.*, 2019].

Although several works are proposed to improve the performance of one-stage detector, few works focus on exploring multi-dimensional attention mechanisms. There are several reasons for introducing attention mechanisms to one-stage detector. First, the importance of samples in each mini-batch are not the same. One-stage detectors directly predict the classification and localization results, which highly rely on the training samples. It is crucial that such detectors pay more attention to the important samples rather than overwhelmed by those wrong or missing outliers in the dataset. Second, SENet [Hu *et al.*, 2018] has proved that extracting channel-wise information is important for both classification and localization. Channel attention can improve the representation ability of detectors and generate more effective features for detection. Third, one-stage detectors face large background-foreground imbalances due to the lack of region proposal part. The detectors need to discriminate objects of interests from numerous background regions. Thus, focusing more on the foreground regions is extremely important. A mechanism similar to region proposal part in two-stage detectors is needed.

Based on the above discussions, we propose a multi-dimensional attention-based one-stage detector, named GestureDet, and build a real-time student behavior analysis system, shown in Fig. 1. The multi-dimensional mechanisms, including batch attention, channel attention, and spatial attention, are introduced to the classical one-stage detector, MobileNetV2-SSD. First, considering that samples in each mini-batch are not equally important, we propose a novel re-weight strategy to put more focus on these important samples (high IoU with ground-truth but low class confidence) and suppress these outliers. Moreover, we adopt the context modeling and squeeze-and-excitation modules for channel attention to generate more representative features. Finally, to solve the background-foreground class imbalances, we use stacked dilated convolutions to generate a soft weight mask for detection features, thus increase the foreground feature responses.

Our main contributions can be concluded as follows:

(1) We build a large-scale student gesture detection dataset, including 70k hand-raising samples, 20k stand-up samples, and 3k sleeping samples.

(2) We design a real-time student gesture analysis system based on classical real-time detector MobileNetV2-SSD with multi-dimensional attention mechanisms to improve detection performance with little overhead. A novel feature fusion strategy is also proposed to further improve the detection performance.

(3) We demonstrate our proposed methods both on our student behavior dataset and public PASCAL VOC dataset to show the effectiveness and versatility.

2 Related Works

With the fast development of deep learning, CNN has been widely used in traditional vision tasks and achieved impressive results. In this section, we briefly present the CNN-based object detection and the attention mechanisms in CNNs.

2.1 CNN-based Object Detection

In recent decades, CNN-based object detectors greatly improve the detection performances.

CNN-based Detectors. CNN-based detectors can be roughly divided into two kinds: two-stage object detectors (also called region-proposal based detectors) and one-stage detectors. In two-stage detectors, R-CNN [Girshick *et al.*, 2014] is among the earliest CNN-based detectors. Since then, numerous improvements have been proposed for better accuracy and faster speed. On the other hand, one-stage detectors [Redmon *et al.*, 2016; Liu *et al.*, 2016] achieve faster inference with competitive accuracy. In this paper, we present GestureDet based on classical one-stage detectors.

Real-time Object Detection. Real-time object detection is a huge challenge for CNN-based detectors due to the heavy computation of deep convolution layers. One-stage detectors are considered as the key to real-time detection due to the simpler network design and faster inference speed. For example, YOLO [Redmon *et al.*, 2016] and SSD [Lin *et al.*, 2017] can run in real-time on GPU. Integrated with light-weight backbone networks like MobileNet [Howard *et al.*, 2017], these light-weight one-stage detectors such as SSD-lite [Howard *et al.*, 2017], Tiny-YOLO [Redmon *et al.*, 2016] and Pelee [Wang *et al.*, 2018], can achieve real-time inference on mobile devices. In this paper, we use the most popular real-time detector MobileNetV2-SSD [Sandler *et al.*, 2018] as the baseline and incorporate multi-dimensional attention mechanisms to it.

2.2 Attention for Object Detection

Attention mechanism [Vaswani *et al.*, 2017] was first proposed in Nature Language Processing (NLP) tasks and has been widely used in the following works. In general, attention means focusing on the more important parts such as particular locations or feature representations. Due to the effectiveness of attention in NLP fields, there are several existing works introducing attention mechanisms in computer vision tasks and achieving impressive results.

Batch Attention. Recently, some hard example mining strategies such as OHEM [Shrivastava *et al.*, 2016] and Focal loss [Lin *et al.*, 2017] are proposed to focus on these hard samples for boosting detection performance. These hard samples are selected or re-weighted according to their loss values, higher loss value leading to larger weights. Although these simple methods achieved impressive boosts on public datasets, they cannot get optimal results on our student behavior dataset due to some wrong and missing annotations. These outliers usually affect the stability of the models since they tend to have higher loss values and larger gradients. A recent study [Cao *et al.*, 2019] proposed the definition of prime samples, which means the most important samples for training an object detector. And the main strategy is to assign higher weights to those positive samples with higher IoUs with the ground-truth objects. However, the prime samples are only defined on positive samples and most samples in one-stage detectors are negative samples.

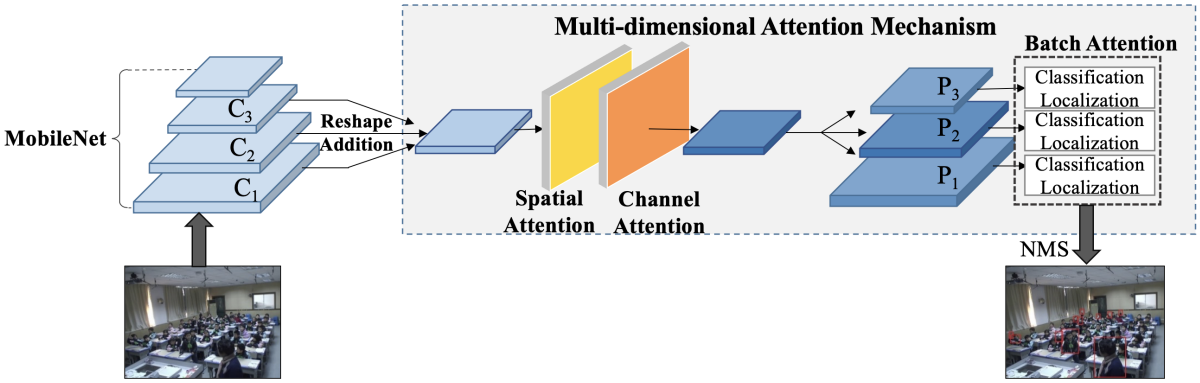


Figure 2: The overall architecture of our proposed GestureDet, which contains the feature fusion module and multi-dimensional attention modules including batch attention, channel attention, and spatial attention.

Channel Attention. SENet [Hu *et al.*, 2018] first explored the relationship between channels in CNNs and won first place on the public image classification challenge. The Squeeze-and-Excitation (SE) block is proposed to improve the quality of representations generated by the CNN backbones. Typically, SE blocks model the dependencies between channels of the convolutional features and perform feature recalibration. This learnable mechanism can utilize global context information to selectively emphasize informative features while suppressing the less important ones.

Spatial Attention. Residual Attention Network [Wang *et al.*, 2017] first proposed combining residual modules and attention modules to generate attention feature representations. ThunderNet [Qin *et al.*, 2019] used the RPN output as the soft-weight mask to generate re-weighted feature maps, which can be seen as spatial attention in object detection. However, spatial attention module in one-stage detectors has been rarely studied.

3 Our Method

In this section, we first give a detailed introduction to the overall architecture of our proposed GestureDet. Then we elaborate on the strategies, including the feature fusion module, and multi-dimensional attention mechanisms, which we design for alleviating challenges in the real scenarios.

3.1 Overall Architecture

The overall network architecture of GestureDet is illustrated in Fig. 2. GestureDet is based on MobileNetV2-SSD [Sandler *et al.*, 2018] with several improvements. Specifically, we first propose a novel feature fusion module to enhance the scale-invariant detection. Then the channel attention module and spatial attention module are sequenced to strengthen the original detection features, which assign learned soft weights to different channels and spatial locations. We use the enhanced features to re-construct the feature pyramid for detecting objects of various scales. Moreover, during the training process, we re-weight the samples in each mini-batch according to the localization loss and classification loss and put more efforts into training the selected important samples, which we called batch attention.

3.2 Feature Fusion Module

Original SSD [Liu *et al.*, 2016] used the outputs from different layers for predictions. However, SSD simply attaches localization and classification branches to each level without any feature fusion, which leads to the poor performance on detection objects of various scales.

To solve the above-mentioned issues, we propose a novel feature fusion module to aggregate multi-level spatial details and context information for more discriminative features. The multi-level features C_1, C_2, C_3 from MobileNetV2 backbone are resized into the intermediate size same to C_2, with nearest upsampling and stride-2-convolutions downsampling. Then we apply 1x1 convolutions to each feature map for squeezing the channels and reducing computation cost. The aggregated feature map is then fed into the channel attention module and spatial attention module to get more robust and discriminative feature responses. Finally, the after-attention feature map is used to re-construct the feature pyramid with opposite reshape operations, as shown in Fig. 2.

By leveraging both high-resolution spatial details at low levels and semantic information at high levels, our proposed feature fusion module effectively enhances the representation ability of this shallow and thin backbone network. Moreover, compared to the prior feature pyramid structure, our feature fusion modules only involve non-parameter resize operation and 1×1 convolutions, which are more computation-friendly.

3.3 Multi-dimensional Attention Mechanisms

CNNs extract hierarchical features from original images using convolution operators. Typically, the base data format of the images and features is $[N, C, H, W]$, which denotes batch, channel, and spatial (height, width) respectively. Original CNNs extract batch, channel, and spatial information in a unified way.

In this paper, we propose a multi-dimensional attention mechanism including batch attention, channel attention, and spatial attention to the original MobileNetV2-SSD for better

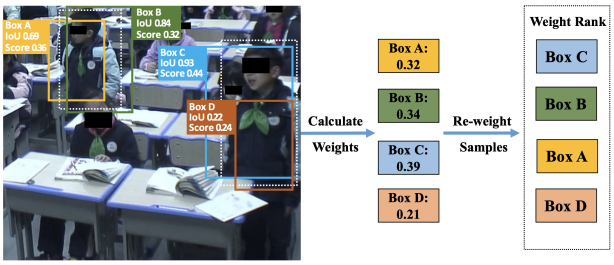


Figure 3: Batch Attention. The white boxes denote the ground-truth boxes, while boxes with other colors represent the predictions. These predictions are re-weighted according to the classification losses and localization loss.

detection performances. This comprehensive attention mechanism covers each dimension of the tensors ($[N, C, H, W]$) and enables more discriminative features for detection. In the following sections, we will give a detailed analysis of these three kinds of attention modules.

Batch Attention

CNN-based detectors, including two-stage detectors and one-stage detectors, are usually trained to classify and localize the sampled regions. Therefore, the selection of these sampled regions is crucial to the final detection results. Most detectors simply treat these sampled regions as equally important and are trained to optimize the average loss among these samples. However, in our scenarios, most of the sampled regions are located in the background, which leads to an inaccurate average loss for poor optimization.

Inspired by the prior works, we propose a batch attention mechanism that combines hard negative mining and positive samples re-weighting. Specifically, instead of using all the negative examples, we sort them using the highest confidence loss for each default box and pick the top ones so that the ratio between the negatives and positives is at most 3:1, following the hard mining practice in original SSD [Liu *et al.*, 2016].

And the positive samples are re-weighted according to the localization loss and classification loss, as shown in Eqn. (1). In general, we want the detectors focusing more on samples that have higher IoUs with ground-truth objects but have low confidence scores. These truly hard examples are more important and have more possibilities to have objects we interested in. Moreover, the classification losses represent the confidence scores of objects, while the localization losses are a good estimation of IoUs with ground-truth. We give higher weights to samples with low localization loss and high classification loss to avoid these important samples being filtered by the Non-Maximum-Suppression post-processing. The re-weight strategy is shown in Eqn. 1, where l'_{cls} and l'_{loc} denote the original classification loss and localization loss of each sample. We choose α and β as 0.5 for simplicity and more hyper-parameter combinations will be explored in the future work.

$$w_i = \alpha \frac{\frac{1}{l'_{loc}}}{\sum_{k=1}^{n} \frac{1}{l'_{loc}}} + \beta \frac{l'_{cls}}{\sum_{k=1}^{n} l'_{cls}} \tag{1}$$

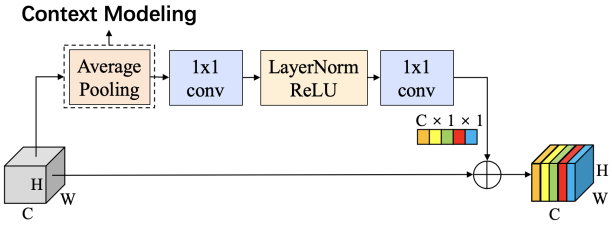


Figure 4: Channel Attention. The feature maps are shown as feature dimensions, where C×H×W denotes a feature map with channel number C, height H and width W. $\bigoplus$ denotes broadcast element-wise addition.

With the proposed re-weight strategy, the classification loss (L_{cls}) can be rewritten as Eqn. 2, where n and m are the numbers of positive and negative samples respectively, c and $\hat{c}$ denote the predictions and targets. Following the practice in PISA [Cao *et al.*, 2019], we also normalize the weights to remain total classification losses unchanged. CE represents Cross Entropy loss here.

$$L_{cls} = \sum_{i=1}^{n} w'_i CE\left(c_i, \hat{c}_i\right) + \sum_{j=1}^{m} CE\left(c_j, \hat{c}_j\right) \tag{2}$$

$$w'_i = w_i \frac{\sum_{k=1}^{n} CE\left(c_k, \hat{c}_k\right)}{\sum_{k=1}^{n} w_k CE\left(c_k, \hat{c}_k\right)} \tag{3}$$

What's more, most object detectors use a multi-task loss to solve the classification and regression (localization) tasks simultaneously. This leads to the issues of possible range inconsistencies among classification and regression losses. A technique to solve it is to assume classification and regression tasks are correlated and combine these two loss terms. In this paper, with our designed batch attention mechanisms shown in the above equations, the classification and regression losses are explicitly linked together. Thus, these two branches can get additional supervision from others and enable extra gradient flow, which benefits the training of detectors.

Channel Attention

To indicate the channel-wise feature dependencies, a channel attention mechanism is designed followed the practice in SENet, as shown in Fig. 4. First, a global average pooling is adopted to model the global context. Then, we use 1×1 convolutions with LayerNorm to estimate the relative importance between channels. The main difference between our channel attention module and SE block is that we use broadcast element-wise addition to generate the final enhanced features rather than sigmoid activation and multiplication in the SE block.

Spatial Attention

Spatial attention in object detection aims to filter unimportant background and focus more on the foreground objects. As we mentioned above, the region proposal part in two-stage detectors can be viewed as a special kind of spatial attention mechanisms and is the key to the impressive detection performances. Due to the lack of the region proposal part, one-stage detectors face extremely imbalanced classification and

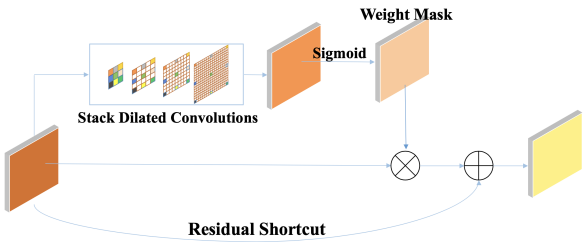


Figure 5: Spatial Attention. The convolution layers with gradually increasing dilation rates are sequentially stacked to generate the attention weight mask. Moreover, a residual shortcut is used to construct identity mapping.

perform poorly compared to two-stage detectors. To make matters worse, as GestureDet utilizes a light-weight backbone network and the small input image, it is harder for such thin model to learn a proper feature distribution. Thus, we softly re-weight the feature maps with learned weight masks to address the extreme background-foreground imbalances.

Inspired by the recent state-of-the-art semantic segmentation model DeepLabV3 [Chen *et al.*, 2017], we propose to use stacked dilated convolutions with gradually increasing dilation rate for generating the learnable weight masks. Suppose the original feature map is F and $\theta(\cdot)$ denotes the stacked dilated convolutions, the output feature map F^{SP} is defined as:

$$F^{SP} = F \cdot (1 + sigmoid(\theta(F))) \tag{4}$$

As shown in Fig. 5, convolution layers with increasing dilation rates (1, 2, 4, 8) are sequentially stacked to capture both local and global context. The last convolution layer will output a same-size mask with attention weights. Then this mask is fed into a sigmoid activation to constrain the value within [0, 1] and estimate the relative importance between each spatial region. The spatial attention module will output the same-size soft attention weights, which are then multiplied with the original feature value for better feature distributions. Therefore, during the training process, these attention masks can be used to distinguish foreground features from numerous background features. What's more, to stabilize the earlier training process with initial noisy attention weights, we add a residual shortcut to combine the original features with the after-attention features.

Our proposed spatial attention module can not only alleviate the imbalanced classification in one-stage detectors but also achieve more focused and effective training due to the re-weighted backward gradients.

4 Experiments

To demonstrate the effectiveness and versatility of our proposed GestureDet, we conduct extensive experiments both on our student behavior dataset and the public PASCAL VOC dataset. For both datasets, we show the results with metrics of PASCAL VOC [Everingham *et al.*, 2010]: mean Average Precision (mAP).

	baseline	ablations				ours
+feature fusion?	–	✓	✓	✓	✓	✓
+batch attention?	–		✓			✓
+channel attention?	–			✓		✓
+spatial attention?	–				✓	✓
mAP (%)	70.7	71.6	74.5	72.7	74.7	**75.2**

Table 1: Experiment results of baseline and our methods on our student behavior dataset.

4.1 Implementation Details

We implement on PyTorch library. The detectors are trained end-to-end on one GPU using SGD with a weight decay of 0.0001 and a momentum of 0.9, following the settings in prior works. The input image resolution is 300×300 pixels for efficiency and the batch size is set to 64 images. The data augmentation strategies are the same as the original SSD [Liu *et al.*, 2016]. The learning rate starts from 0.002 with warm-up epochs and decays exponentially every step. Note that, we do not use other tricks like multi-scale training and better post-processing like Soft-NMS. What's more, we use network inference time (without pre-processing and post-processing) tested on GPU (1080Ti), CPU (Intel i7-8700K) and embedding devices (Jetson TX2) to evaluate the inference speed of models and FLOPs (floating point operations) to evaluate the computation cost of models.

4.2 Experiments on Our Student Behavior Dataset

We perform our proposed GestureDet on the student behavior dataset, including 70k hand-raising samples, 20k stand-up samples, and 3k sleeping samples. The behaviors we captured are collected from 1080P cameras in 30+ different primary and middle schools in Minhang distinct, Shanghai, China.. Our dataset are really challenging due to the various scales, large class imbalances, and less high-quality annotations. There are large scale variations among different behaviors of almost 25 times, such as hand-raising (about 40×40 pixels) and standing (about 200×200 pixels). To make matters worse, nearly 70% of the objects in our dataset only occupy less than 0.5% part of the whole image, which introduces the challenge of detecting very small objects.

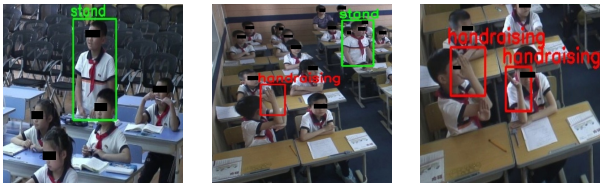
We use 29k images (out of 40k images in total) for training, then validate performances on the rest 11k subset. Original MobileNetV2-SSD is used as the baseline for comparison to demonstrate the effectiveness of our proposed methods.

As shown in Table 1, our proposed GestureDet achieves better performance than the baseline. The feature fusion strategy already improves the baseline to 71.6% mAP. Based on this higher result, the multi-dimensional attention mechanisms still outperform by 3.6% mAP. From the ablations columns, we can see continuous improvements of our multi-dimensional attention mechanisms. The batch attention and channel attention significantly improves the mAP by 3.8% and 2%, respectively. Moreover, the batch attention module is only performed on the training process and the channel attention module introduces little computation overhead. With the spatial attention module, the mAP is increased by 4% with a slightly higher computation cost.

Model	GPU (ms)	CPU (ms)	Jestson TX2 (ms)
baseline	4.5	240.7	54.0
+feature fusion	5.9	253.2	54.4
ours	7.3	259.9	54.9

Table 2: Inference time tests of baseline (MobileNetV2-SSD) and our methods on our student behavior dataset.

(a) Detection results of baseline.



(b) Detection results of ours.

Figure 6: Examples of detection results obtained on some images from the test-set. The top and bottom rows show the detection results of the baseline and our methods, respectively. Compared with the results of baseline, our methods can detect more behaviors.

Table 2 shows that our proposed GestureDet can still achieve a fast speed on CPU and embedding devices (Nvidia Jetson TX2), which shows a more suitable accuracy/speed trade-off for real classrooms. Note that, the inference time only calculates the network forward time on GPU or CPU without pre-processing and post-processing time for a fair comparison. The image-preprocessing and post-processing are usually executed asynchronously with other tasks on CPU. Thus, the actual inference speed should be very close to our test results.

4.3 Experiments on PASCAL VOC

PASCAL VOC [Everingham *et al.*, 2010] dataset consists of natural images drawn from 20 classes. The detectors are trained on the union set of VOC 2007 trainval and VOC 2012 trainval, and tested on VOC 2007 test. The results are shown in Table 3. Note that, for the comparison algorithms, experimental settings are the same as in their publications.

GestureDet outperforms original MobileNetV2-SSD with 2.2% mAP. Moreover, our proposed GestureDet significantly surpass prior state-of-the-art light-weight one-stage detectors such as Tiny-YOLO [Redmon *et al.*, 2016], Pelee [Wang *et al.*, 2018] and Tiny-DSOD [Li *et al.*, 2018] while maintain similar computation cost. Furthermore, GestureDet achieves comparable results with state-of-the-art one-stage object detectors such as SSD300 [Liu *et al.*, 2016] and YOLOv2 [Redmon and Farhadi, 2017], but significantly reduce the computation cost and accelerate the inference.

Model	Input	MFLOPs	mAP
SSD300 [Liu *et al.*, 2016]	300 × 300	31750	77.5
YOLOv2 [Redmon and Farhadi, 2017]	416 × 416	17400	76.8
Tiny-YOLOv2 [Redmon and Farhadi, 2017]	416 × 416	3490	57.1
MobileNet-SSD [Howard *et al.*, 2017]	300 × 300	1150	68.0
Pelee [Wang *et al.*, 2018]	304 × 304	1210	70.9
Tiny-DSOD [Li *et al.*, 2018]	300 × 300	1060	72.1
MobileNetV2-SSDLite **(baseline)**	300 × 300	1480	72.3
GestureDet **(ours)**	300 × 300	**1830**	**74.5**

Table 3: Experiment results on VOC 2007 test. GestureDet achieves superior performances with similar computation cost.

5 Conclusion

Privacy Issues. One of the potential issues of analyzing classrooms is privacy protection. In our system, we try our best to balance between positive use cases and abuses. Firstly, these behavior samples in the dataset are captured from open excellent courses in 30+ different schools in our distinct, where the teachers and students are informed and agreed that the classes are recorded and analyzed by experts of schools. And we gain the permissions from schools and parents to analyze behaviors of students and teachers to assist teaching quality estimation. Secondly, the dataset cannot access without permission to avoid possible leakage. Thirdly, our system only provides masking data rather than personalized data for school administrators. We will always put privacy protection as our primary concerns and are open to public scrutiny.

System Performances. In this paper, we propose a real-time detector named GestureDet with multi-dimensional attention mechanisms, including batch attention, channel attention, and spatial attention. These three kinds of attention modules are proposed and fused to the classical real-time detector MobileNetV2-SSD. The batch attention is constructed with re-weighting the samples in each mini-batch and put more effort into these important samples. The channel attention consists of context modeling and squeeze-and-excitation blocks for more discriminative feature representations. The spatial attention uses stacked dilated convolutions to generate a learnable weight mask for addressing the extremely imbalanced classification in one-stage detectors. Experiments demonstrate the effectiveness and versatility of our proposed methods both on real classroom scenarios and the public PASCAL VOC benchmark. Our proposed GestureDet can run at a fast speed on CPU or embedded devices.

Future Work. In the future, we would like to further improve the detection performances with other attention mechanisms and integrate incorporate more student behaviors into our system to help better understand the teaching status.

Acknowledgments

The authors would like to thank the editors and anonymous reviewers for their constructive suggestions to improve the manuscript. The work was supported by National Nature Science Foundation of China (No. 61671290), China Postdoctoral Science Foundation (No. 2018M642019), and Shanghai Municipal Commission of Economy and Information (No. 2018-RGZN-02052).

References

[Cao *et al.*, 2019] Yuhang Cao, Kai Chen, Chen Change Loy, and Dahua Lin. Prime sample attention in object detection. *arXiv preprint arXiv:1904.04821*, 2019.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[Chen *et al.*, 2019] Yuntao Chen, Chenxia Han, Naiyan Wang, and Zhaoxiang Zhang. Revisiting feature alignment for one-stage object detection. *arXiv preprint arXiv:1908.01570*, 2019.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[Howard *et al.*, 2017] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[Li *et al.*, 2018] Yuxi Li, Jiuwei Li, Weiyao Lin, and Jianguo Li. Tiny-dsod: Lightweight object detection for resource-restricted usages. *arXiv preprint arXiv:1807.11013*, 2018.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[Liu *et al.*, 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[Qin *et al.*, 2019] Zheng Qin, Zeming Li, Zhaoning Zhang, Yiping Bao, Gang Yu, Yuxing Peng, and Jian Sun. Thundernet: Towards real-time generic object detection. *arXiv preprint arXiv:1903.11752*, 2019.

[Redmon and Farhadi, 2017] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[Redmon *et al.*, 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[Sandler *et al.*, 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[Shrivastava *et al.*, 2016] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Wang *et al.*, 2017] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[Wang *et al.*, 2018] Robert J Wang, Xiang Li, and Charles X Ling. Pelee: A real-time object detection system on mobile devices. In *Advances in Neural Information Processing Systems*, pages 1963–1972, 2018.