scientific reports

Check for updates

OPEN

Multimodal learning audio-visual detection for obtaining object-level sound sources in Japanese-language teaching room

Lu Li[1]✉, Xiuxiu Bai[2]✉, Junxiu Xu[2], Dingkang Wang[3]✉ & Tao Jiang[1]

The combination of artificial intelligence and education is one of the current trends in research. While observing the daily teaching and learning process at school, we have considered the possibility of using multimodal learning, in particular audio-visual detection (AVD), to improve the teaching and learning process in Japanese-language teaching rooms. AVD can be effectively used to locate sounding objects (e.g. clapping, sneaking, organizing things, etc.) from unknown sources in online or physical classrooms. This study proposes a novel deep learning-based approach for audio-visual detection (AVD) in Japanese-language teaching rooms, combining audio and visual information to detect sound sources at the object level. To evaluate the proposed method, we construct an AVD benchmark that provides object-level annotations according to the sound sources in the videos. The feasibility of applying our proposed method in the classroom is demonstrated by designing evaluation metrics for AVD and comparing it with similar works.

Keywords Japanese-language teaching, Artificial intelligence, Multimodal learning, Audio-visual detection

Deep learning technologies have facilitated the rapid development of a wide range of industries. With the increasing development of computer vision technology, more and more researchers are focusing their efforts on the application of specific technologies in campus settings. In typical classroom scenarios, the most prominent feature is crowding. This is especially evident in large classrooms, where hundreds of students may be attending class simultaneously. For faculty, managing a large number of students is more difficult to handle. For example, it can be difficult to find students who are talking in secret or making noise in the classroom, or to notice when a student is answering a question in his or her seat without the teacher noticing.

Audio and visual information are related. On the temporal level, they often occur simultaneously, and on the spatial level, they both come from the same location in the video frame. Therefore, they are similar in feature space. The goal of audio-visual detection (AVD) is to identify the location and category of specific events in classroom videos by using audio and visual information. In terms of practical application requirements and task logistics, what AVD needs to achieve is to find the sound-emitting students or other objects in the classroom, which represents the event. Therefore, from another point of view, AVD can be seen as a multimodal object detection task that utilizes both audio and visual information.

Many state-of-the-art detectors can be applied to various scenarios that require accurate object detection. There have been many applied studies based on object detection in intelligent education systems, such as student counting[1] and student behavior detection[2]. One of the difficulties with object detection in the classroom is that the image features are not distinct. For example, when a student is answering or asking a question from his or her seat, it is difficult to capture robust visual information from the image alone, and in this case, it is beneficial to include sound features.

To efficiently combine object detection with sound, we propose an audio-visual detector (AVDor) based on multimodal learning. In this study, the main contributions are summarized as follows: (1) For intelligent education applications, we introduce the audio-visual detection (AVD) task which combines audio and visual information for detecting events of interest in classroom scenes. (2) To achieve AVD, we propose a novel multimodal-based AVDor that receives audio and visual information as input and outputs the object location

[1]School of Japanese Culture and Economics, Xi'an International Studies University, Xi'an 710128, China. [2]School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China. [3]School of Public Health, Global Health Institute, Xi'an Jiaotong University, Xi'an 710061, China. ✉email: li.lu@xisu.edu.cn; xiubai@xjtu.edu.cn; dkwang@xjtu.edu.cn

and class. (3) For evaluation, we construct a benchmark for AVD, which provides object-level annotations based on the sound sources in the videos.

Related work

In the field of audio-visual multimodal research, there are two broad categories based on the coarseness of the localization results. Research such as audio-visual correspondence[3] (AVC), audio-visual event localization[4] (AVEL), and audio-visual video parsing[5] (AVVP), which segment videos into events based on audio and visual information, can be considered coarse-grained. Considering the applications in the classroom, these methods match audio and video clips but do not meet the needs of teachers. Other fine-grained methods, such as sound source localization[6] (SSL) and audio-visual segmentation[7] (AVS), can locate the region of the video frame where the sound is made or the pixels of a specific instance in the video. SSL methods locate only the location of the sound source, not specifically an instance itself, which is still not enough to solve the problem.

For the majority of cases, finding a sounding object in a classroom scene, as in the case of an object detection task, is sufficient. Therefore, AVS is unnecessary at such a fine-grained level, as it segments objects down to their shapes. Also, the instances segmented by AVS do not contain their categories, thus their output does not reflect the type of the event.

As a reference task for the realization of AVD, a superior object detector also plays an important role. As a fundamental task in computer vision, object detection has been well developed. There are two main types of object detectors: (1) One-stage detectors, such as the YOLO (You Only Look Once) series detectors[8] and SSD (Single-Shot Detector)[9]; (2) Two-stage detectors, such as Faster R-CNN[10] and Cascade R-CNN[11], etc. These mainstream models have demonstrated satisfactory performance on natural image data.

Object detection is also widely used in smart education, especially in classroom scenarios. Liu, et al.[1] proposed a student counting system based on object detection, which focuses on solving the problem of small object detection. Rao and Chen[12] developed a YOLOv5-based method for real-time student counting in classrooms, improving detection accuracy for small objects like students' heads through optimized anchor boxes. This method achieved 91.59% accuracy on the CUHK Occlusion dataset. Lv et al.[2] proposed a student behavior detection system based on SSD, which uses an improved SSD[9] to recognize student behavior. Gan et al.[13] explored IoT and multimodal analysis integration in smart education, combining audio, video, and sensor data. Their method demonstrated potential in enhancing learning analytics through comprehensive data fusion. Tian et al.[14] proposed a Transformer-based framework for audio-visual event localization in videos. Though specific success rates were not provided, their method demonstrated a 61.56% mean attack success rate. These studies highlight the potential of multimodal approaches in enhancing the effectiveness of smart education systems.

Audio-visual detector

In this section, we provide a comprehensive overview and intricate architecture of the AVDor, which is designed to address the challenging task of combining audio and visual information for robust detection of objects or events that occur in Japanese-language teaching scenarios.

Figure 1 shows the overall architecture of the proposed AVDor. From the perspective of object detection combined with audio information, the audio can be considered as an enhancement to vision. Thus, in our study, the AVDor is designed based on a typical one-stage object detector paradigm, which is composed of (1) a feature extraction part, (2) a feature fusion module, and (3) a detection head, followed by post-processing.

Feature extraction

We use a common visual backbone network (i.e. ResNet[15]) to extract visual features from images, and an audio backbone (i.e. VGGish[16]) to extract audio information from sounds. The ResNet is stage-wise designed, and the output of each stage is a feature map (for T images) with different resolutions (H × W) and channels (Ci). As with other common detectors, the features from the backbone are then fused by a feature pyramid network[17] (FPN) for better detecting objects of different scales.

For audio feature extraction, we use VGGish[16], which employs a deep convolutional neural network that processes input audio spectrograms to extract high-level representations for various audio analysis tasks.

Feature Fusion

For common 2D object detection, RGB images are fed to the detector, which processes 3-channel data. To combine audio information, which is single-channel data, the AVDor needs to include a feature fusion module. As shown in the Feature Fusion part in Fig. 1, the image feature and the audio feature are sent to the TPAVI + module together, which is an upgraded version of the TPAVI[7] module that encodes audio-visual relations through a non-local[18] block. The detailed structure of the TPAVI + module is illustrated in Fig. 2. Compared with TPAVI, the TPAVI + module is inserted with 2 enhancers. Figure 3 illustrates the structure of the enhancer.

The features that go through the enhancer can be expressed by Eq. 1.

$$F' = F_{eh} \times F \tag{1}$$

in which the F is the input feature of size $Th_iw_i \times Th_iw_i$ or $T \times h_i \times w_i \times C$. F_{eh} is the enhanced feature, and F′ is the output feature of the enhancer. F_{eh} is calculated by Eq. 2.

$$F_{ch} = Sigmoid(\mathrm{ReLU}(Conv(MP(F)))) \tag{2}$$

where MP denotes the max pooling layer. Conv denotes a 1 × 1 convolution. ReLU and Sigmoid are activation functions. The enhancer controls the relative importance of the audio-visual features, affecting their contribution

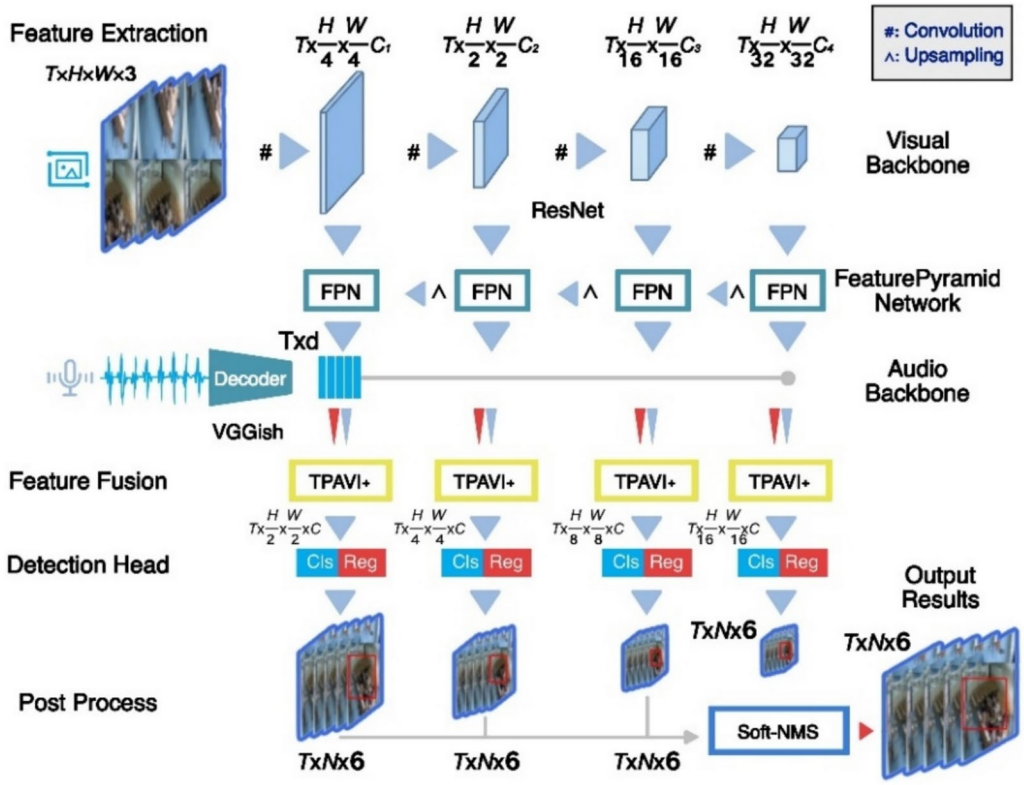

Fig. 1. Overall architecture of audio-visual detector (AVDor). (**a**) the feature extraction part, including a vision backbone (ResNet[15]) combined with FPN[17], and an audio backbone (VGGish[16]); (**b**) the feature fusion module, TPAVI+; (**c**) the detection head, which outputs the object location and class; (**d**) the post processing part, which uses Soft-NMS to filter the detection results.

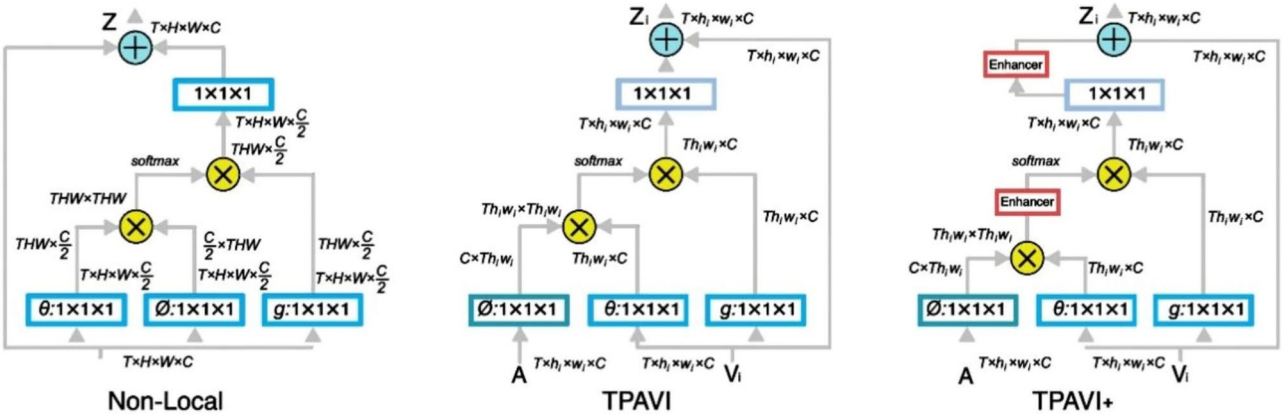

Fig. 2. Illustration of the Non-Local, TPAVI, and TPAVI + structure. Note that we use the same color to represent the same type of feature map. The reshape operation is omitted for simplicity.

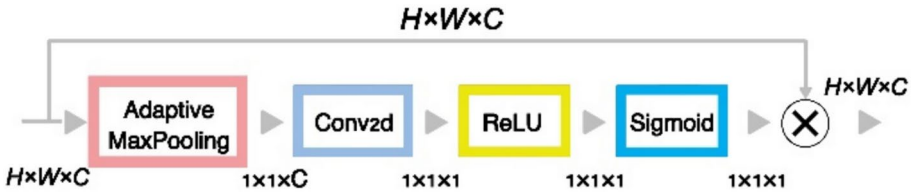

Fig. 3. Structure of the enhancer.

to the final prediction. The ReLU-Sigmoid function ensures that the feature importance is scaled between 0.5 and 1, preventing drastic feature disappearance during training.

The shape of the image feature from each stage varies, while the dimension of the audio feature is fixed (T × 128). Therefore, the audio feature A is transformed into the same channel (dimension) as the image feature V_i by a linear layer (T × 128 to T × C). Then A is reshaped to the same shape as V_i by duplicating (T × C to T × h_i × w_i × C). In TPAVI+, the audio feature and the image feature of ith stage are fused by a non-local paradigm. The fusion can be expressed by Eq. 3.

$$Z_i = V_i + E_1\mu(E_2\alpha_i g(V_i)) \tag{3}$$

where g and μ are 1 × 1 × 1 convolutions. $Zi \in R^{T \times hi \times wi \times C}$. E_1 and E_2 are 2 enhancers. α_i denotes the audio-visual similarity, which can be calculated by Eq. 4.

$$\alpha_i = \frac{\theta(V_i)\phi(\hat{A})^\top}{N} \tag{4}$$

in which θ, ϕ are 1 × 1 × 1 convolutions. N is a normalization factor, which equals T × hi × wi. Through the TPAVI + module, the audio information is fused with the image feature, and the audio-visual relation is encoded.

Loss function

The loss function is designed to reflect the relationship between sound and vision. In addition to the loss of the predicted box outputted by the detection head, it should also contain a distance between the event represented by the predicted box and the sound. We compute the loss function in two parts, as illustrated by Eq. 5.

$$L = L_{det} + \lambda L_{avd} \tag{5}$$

where L_{det} is the common loss of the object detection task, which contains regression loss and classification loss. λ is a balance weight. L_{avd} is the loss of audio-visual relation. We refer to the loss function of AVS[7], which uses Kullback–Leibler (KL) divergence to match the similarity between audio and visual features. The difference is that our losses are calculated and summarized head-wise. The L_{avd} is calculated by Eq. 6.

$$L_{avsd} = \sum_{i=1}^{n} KL(AvgPool(F_i^{tpavi+}), A_i) \tag{6}$$

where n is the number of detection heads. AvgPool is the average pooling layer. F_i^{tpavi+} is the output of the ith TPAVI + module. A_i is the audio feature of the ith stage. KL denotes the Kullback–Leibler divergence.

Benchmark

To the best of our knowledge, there is no previous work or benchmark for AVD in Japanese-language teaching rooms. Therefore, we construct a benchmark for AVD by collecting a dataset and designing evaluation metrics.

Dataset

The dataset is collected from our Japanese-language classroom in Xi'an International Studies University, Xi'an 710128, China. Some sample images are shown in Figure S1. The dataset contains 4500 annotated images, which are divided into 3500 training images and 1000 test images. We define a total of 6 behaviors associated with sounds, which are sneaky talk, clapping, organizing stuff, laughing, answering, and teacher speaking. The detailed statistics of the dataset are shown in Table 1.

The format of the dataset should consider the needs of AVDor. Each event that produces a sound corresponds to a video clip as well as a sound clip. We isometrically sample T frames from the video clip as visual information and use the whole sound clip as audio information. Thus, the audio information will be repeated T times to match the visual information. We manually annotate the bounding box of the object that emits the sound, and the event category of the object.

Behavior	Training	Test
Sneaky talk	835	258
Clapping	606	210
Organizing stuff	320	131
Laughing	857	270
Answering	544	188
Teacher speaking	1,983	906

Table 1. Detailed statistics of the dataset. Note that the numbers indicate the amount of objects in images.

Evaluation metric

For evaluation, we design a metric for the AVD task, which is based on the common metric for object detection, mean average precision (mAP), and audio-visual similarity.

In object detection, the mAP is calculated by Eq. 7.

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{7}$$

where n is the number of classes. AP is the average precision of each class, which is calculated by the enclosed area of PR curve (Precision-Recall curve). Specifically, the precision and recall can be computed by Eq. 8.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}'}, \quad Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

where TP, FP, and FN are the numbers of true positive, false positive, and false negative boxes, respectively.

Specifically, for AVD, the accuracy of audio-vision matching is important. For each sound clip S_i, we calculate the accuracy of detected events corresponding to this clip. Therefore, we use Eq. 9 to calculate the audio-visual match rate (AVMR).

$$AVMR = \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\frac{N_{TP}}{N_{Pred}} \cdot \frac{N_{Pred}}{N_{gt}}\right) \tag{9}$$

where N_s denotes the number of sound clips. N_{TP}, N_{Pred}, N_{gt} are the numbers of true positive, predicted, and ground truth boxes, respectively. The predicted box with IoU and confidence score over the threshold is considered a true positive.

Combining mAP and AVMR, we use mean average matching precision (mAMP) as the evaluation metric, which is calculated as in Eq. 10.

$$mAMP = mAP \times AVMR \tag{10}$$

from which we can conclude that the closer the mAMP is to 1, the better the performance of the AVDor.

Experimental results and discussion

In this section, we introduce the implementation details and evaluate the proposed AVDor based on the constructed benchmark.

Implementation details

The experiments were carried out on a computer equipped with 2 NVIDIA RTX 3090 GPUs (24 GB memory). We use PyTorch-1.8.0[19] to implement the AVDor. We train the AVDor for 24 epochs using the AdamW optimizer with a batch size of 16 and an initial learning rate of 0.0001. The image size is set to 800 × 800.

As for the composition of the training data, for each event corresponding to a video with sound, we isometrically extract video frames with T = 5. The entire sound clip is used as audio information. The 5 frames of the video and the entire sound clip are paired as a training sample.

Comparison experiment

First, we compare the performance of the AVDor with other state-of-the-art object detectors that do not use audio information. The results are shown in Table 2. We report mAP@0.5:0.95 and $AVMR_{thresh=0.5}$ to show the performance of the methods. We can observe that the AVDor outperforms other detectors by a large margin, even with a simple ResNet-50 backbone. The results show that the audio information is helpful for object detection in the classroom.

Detector	mAP (%)	AVMR (%)
YOLOv5s	43.74	38.70
YOLOv6s	46.60	40.23
YOLOv8s	48.02	40.64
YOLOv8x	54.27	47.76
Faster Rcnn-r101	50.82	44.21
Cascade Rcnn-×101	53.20	45.22
Ours-r50	55.45	49.18
Ours-r101	**56.19**	**52.54**

Table 2. Comparison experiment of the AVDor with other state-of-the-art object detectors. r50, r101, and×101 represent ResNet-50, ResNet-101 [15], and ResNeXt-101 [20], respectively. Significant values are in bold.

Method	mAP (%)	AVMR (%)
r50&A + V	47.75	40.08
r50&TPAVI	53.78	48.88
r50&TPAVI +	**55.45**	**49.18**
r101&A + V	51.33	44.89
r101&TPAVI	55.02	51.95
r101&TPAVI +	**56.19**	**52.54**

Table 3. Comparison experiment of the TPAVI + and TPAVI. A + V represent addition operation of audio feature and visual feature. Significant values are in bold.

Then, we compare the proposed TPAVI + module with the original TPAVI module and simple feature addition. The results are shown in Table 3. It is evident that the TPAVI + module outperforms the original version. With the enhancer, the audio-visual relationship is better encoded, which is beneficial to the AVDor. Through the experiments, we can conclude that the AVDor performs better than other object detectors in the classroom, which proves the feasibility of the AVD.

Discussion

Through the experiments, we constructed a benchmark and designed AVDor to demonstrate the feasibility of AVD. Combining audio information improves the object detection performance, reaching 56.19% mAP and 52.54% AVMR. For more intuitive illustration, in Figure S2, we visualize some detection results selected from the test set. As observed, AVDor is able to accurately detect various events that produce sound. This capability proves that AVD is particularly valuable for audio-visual detection in classroom scenarios, as instructors do not have complete control over all information. The application of multimodal AI algorithms contributes to the advancement of smart education systems.

We also acknowledge the limitations of the current study. The dataset remains relatively constrained in scale and may not fully represent the wide spectrum of classroom settings across different schools or teaching styles. Moreover, the model's ability to distinguish overlapping or concurrent sound sources can be further improved. To address these issues, training on larger and more diverse datasets could significantly enhance both detection accuracy and resilience to noise.

Conclusion

In this study, we propose a novel multimodal task for intelligent education, namely audio-visual detection. AVD can be used to locate sound-emitting objects with unclear sources in online or physical classrooms. In order to accomplish AVD, we propose a brand-new multimodal-based AVDor that outputs the object location and class after receiving audio and visual input. We also construct a benchmark for AVD, which provides object-level annotations and an evaluation metric according to the sound sources in the videos. Through experiments, we demonstrate that the proposed AVD can better detect sound-producing persons or events in classroom settings compared to common object detectors, thereby effectively assisting lectures as one of the components of an intelligent education systems.

Data availability

Data is provided within the manuscript or supplementary information files.

Received: 22 January 2025; Accepted: 29 April 2025
Published online: 13 May 2025

References

1. Liu, M., Zhang, X., Han, Y. Intelligent counting system for Japanese-language teaching room numbers based on video surveillance. In *Proceedings of the 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2020, 242–245. https://doi.org/10.1109/ICPICS50287.2020.9201999.
2. Lv, W., Huang, M., Zhang, Y., Liu, S. Research on intelligent recognition algorithm of college students' classroom behavior based on improved SSD. In *Proceedings of the 2022 IEEE 2nd International Conference on Computer Communication and Artificial Intelligence (CCAI)*, 2022, 160–164. https://doi.org/10.1109/CCAI55564.2022.9807756.
3. Aytar, Y., Vondrick, C., Torralba, A. SoundNet: Learning sound representations from unlabeled video. in proceedings of the advances in neural information processing systems; Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; Garnett, R., Eds. Curran Associates, Inc., 2016, arXiv:1610.09001. https://doi.org/10.48550/arXiv.1610.09001.
4. Lin, Y. B., Wang, Y. C. F. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2021, 12627, 274–290. https://doi.org/10.1007/978-3-030-69544-6_17.
5. Tian, Y., Li, D., Xu, C. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of the Computer Vision—ECCV 2020*; 2020, arXiv:2007.10558. https://doi.org/10.48550/arXiv.2007.10558.
6. Senocak, A., Oh, T. H., Kim, J., Yang, M. H., Kweon, I. S. Learning to localize sound source in visual scenes. In *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, arXiv:1803.03849. https://doi.org/10.48550/arXiv.1803.03849.
7. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y. Audio-visual segmentation. In *Proceedings of the Computer Vision - ECCV 2022*; 2022, arXiv:2207.05042. https://doi.org/10.48550/arXiv.2207.05042.

8. Wang, C. Y., Bochkovskiy, A., Liao, H. Y. M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721.
9. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., Berg, A. C. SSD: Single shot multibox detector. In *Proceedings of the Computer Vision—ECCV*, 2016, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
10. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031.
11. Cai, Z., Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, arXiv:1712.00726. https://doi.org/10.48550/arXiv.1712.00726.
12. Rao, J. & Chen, M. YOLOv5-based student counting software design. *J. Intell. Knowl. Eng.* **2**(1), 64–69. https://doi.org/10.62517/jike.202404109 (2024).
13. Gan, W., Dao, M. S., Zettsu, K., Sun, Y. IoT-based multimodal analysis for smart education: Current status, challenges and opportunities. In *Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval* 2022; 32–40. https://doi.org/10.1145/3512731.3534208
14. Tian, Y.; Shi, J.; Li, B.; Duan, Z.; Xu, C. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018: 247–263. https://doi.org/10.48550/arXiv.1803.08842
15. He, K., Zhang, X.; Ren, S., Sun, J. Deep residual learning for image recognition, In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, 770–778. https://doi.org/10.1109/CVPR.2016.90.
16. Hershey, S., Chaudhuri, S. , Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., Wilson, K. CNN architectures for large-scale audio classification. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, 131–135. https://doi.org/10.1109/ICASSP.2017.7952132.
17. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, 936–944, https://doi.org/10.1109/CVPR.2017.106.
18. Wang, X., Girshick, R., Gupta, A., He, K. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, arXiv:1711.07971. https://doi.org/10.48550/arXiv.1711.07971.
19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in Neural Information Processing Systems*. 2019, arXiv:1912.01703. https://doi.org/10.48550/arXiv.1912.01703.
20. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, 5987–5995. https://doi.org/10.1109/CVPR.2017.634.

Acknowledgements

L. Li and X. Bai contributed equally to this study. Thanks for the support from the National Social Science Foundation Chinese Academic Translation Project (23WZSB004), Scientific Research Program Funded by Shaanxi Provincial Education Department (24JK0190), Xi'an International Studies University Graduate Education Comprehensive Reform Research and Practice Project Teaching Reform Project (22XWYJGA17).

Author contributions

L. L., J. T. and X. B. wrote the main manuscript. L. L. prepared Figs. 1–3. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-00588-0.

Correspondence and requests for materials should be addressed to L.L., X.B. or D.W.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com