SEMIPL: A SEMI-SUPERVISED METHOD FOR EVENT SOUND SOURCE LOCALIZATION

Yue Li[1], *Baiqiao Yin*[1], *Jinfu Liu*[1], *Jiajun Wen*[1], *Jiaying Lin*[1], *Mengyuan Liu*[†2]

[1]School of Intelligent Systems Engineering, Sun Yat-sen University
[2]National Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School
Corresponding author is Mengyuan Liu (e-mail: liumengyuan@pku.edu.cn).

ABSTRACT

In recent years, Event Sound Source Localization has been widely applied in various fields. Recent works typically relying on the contrastive learning framework show impressive performance. However, all work is based on large relatively simple datasets. It's also crucial to understand and analyze human behaviors (actions and interactions of people), voices and sounds in chaotic events in many applications, e.g., crowd management, and emergency response services. In this paper, we apply the existing model to a more complex dataset, explore the influence of parameters on the model, and propose a semi-supervised improvement method SemiPL. With the increase in data quantity and the influence of label quality, self-supervised learning will be an unstoppable trend. The experiment shows that the parameter adjustment will positively affect the existing model. In particular, SSPL achieved an improvement of 12.2% cIoU and 0.56% AUC in Chaotic World compared to the results provided. The code is available at: https://github.com/ly245422/SSPL

1. INTRODUCTION

Vision can influence our perception of sound. For example, when we view a picture, visual images can change our interpretation and emotional experience of the sound. Visual images, colors, and motion can guide our perception of sound and the construction of scenarios. Hearing can also influence our perception and experience of visuals [1]. For example, when we hear a sound, the pitch, volume, and rhythm of the sound can influence our perception and emotion of visual objects. Sound can enhance or diminish our attention, emotions, and feelings about our visual environment.

When we hear the roar of a car engine, we can infer the speed and motion of the car based on the pitch and volume of the engine sound. Early experiments [2] have shown that sudden sounds enhance perceptual processing of subsequent visual stimuli. This sound perception of speed directly affects our visual perception of the road and vehicles around us. This interplay and interaction stems from the brain's ability to synthesize and process different sensations. By integrating existing sensory inputs, the brain develops a more comprehensive and accurate perceptual experience. This interaction also reflects the connections and coordination between perceptions, helping us to better understand and adapt to our surroundings.

Thus, acoustic event detection is a rather broad topic [3] in the field of environmental sound detection and classification, with a wide range of applicability in surveillance and monitoring, assistive technologies, and multimedia indexing. Recent works showing impressive localization performance typically rely on the contrastive learning framework. With the increase in data quantity and the influence of label quality, self-supervised learning will be an unstoppable trend in the future [4]. For datasets with partial labels, undoubtedly, semi-supervised learning is the best choice and also the inevitable trend for the future development of sound source localization. So, we propose a semi-supervised method SemiPL.

The main contributions of this paper are:

- We apply SSPL to the Chaotic World dataset achieving an improvement of 12.2% cIoU and 0.56% AUC.
- We explore the application of the semi-supervised method SemiPL and the effect of different parameters on SSPL on the Chaotic World dataset.

2. RELATED WORK

2.1. Sound Localization in Visual Scenes

Sound source localization within visual scenes seeks to pinpoint the location of objects producing sound within a specified image. Initial efforts to address this challenging task predominantly focused on leveraging statistical modeling of crossmodal relationships, utilizing techniques such as mutual information and canonical correlation analysis to effectively capture the underlying associations. Nonetheless, these shallow models primarily demonstrate their strengths in relatively straightforward audio-visual scenarios. Currently, employing deep learning techniques to tackle this task has become the mainstream approach. For example, Senocak et al. [5] introduce an innovative unsupervised algorithm for localizing sound sources by utilizing an attention mechanism. This mechanism is guided by auditory information in conjunction with paired sound and video frames, offering a sophisticated

approach to address the problem. In response to the daunting challenge of visually localizing multiple sound sources in unconstrained videos without the availability of pairwise sound-object annotations, Qian et al. [6] engineered a two-stage audiovisual learning framework. This innovative solution separates audio and visual representations of different categories within complex scenes and subsequently executes a cross-modal feature alignment in a progressively refined manner.

2.2. Self-Supervised Visual Representation Learning

Self-supervised learning (SSL) has made significant strides and achieved remarkable breakthroughs in large-scale computer vision benchmarks. The majority of contemporary SSL methods rely on the implementation of contrastive learning strategies. Fundamentally, these approaches transform a single image into multiple views, simultaneously repelling distinct images (negatives) while attracting various perspectives of the same image (positives). Numerous efforts [7, 8, 9] have been dedicated to alleviating the reliance on negatives and streamlining the SSL framework beyond the scope of traditional contrastive learning approaches. For instance, to minimize the computational burden associated with positive and negative sample pairs, Ermolov et al. [8] chart a different course by proposing a novel loss function for SSL, which is founded on the whitening of latent-space features. Grill et al. [9] presents a fresh approach to self-supervised image representation learning, known as Bootstrap Your Own Latent (BYOL). This innovative method hinges on the synergy between two neural networks, dubbed online and target networks, which collaborate and continuously learn from one another.

2.3. Audio-Visual Representation Learning.

Often, vision and sound function as complementary senses that naturally aid in supervising audio-visual learning [10, 11, 12]. For instance, in [10], visual features drawn from a pre-trained teacher network assist a student network in acquiring more distinctive audio representations, and the reverse is also true [12]. Korbar et al. [11] and Owens and Efros [41] leveraged the synchronization between audio and visual streams to create negative samples and formulate contrast loss, respectively, aiming to derive generalized multisensory characteristics. Certain investigations have also delved into audio-visual correspondences via feature clustering. Primarily, these methodologies concentrate on acquiring task-independent representations for downstream classification-related endeavors, including action/scene recognition, audio event categorization, and video retrieval. Nevertheless, as they are not tailored for sound source localization, their performance on this particular task remains constrained.

3. CHAOTIC WORLD DATASET

The SSPL [13] model has achieved excellent results on the traditional Flickr-SoundNet [14] dataset and VGG-Sound [15] with the leading edge. However, It's also crucial to understand and analyze human behaviors (actions [16, 17, 18] and interactions of people [19]), voices, and sounds in chaotic events in many applications, e.g., crowd management, and emergency response services. Unlike human behavior in everyday life, human behavior during chaotic events is often more complex and their actions and impacts on others are often unusual. Thus, this paper aims to utilize the Chaotic World dataset [20], which is a large and challenging multimodal video dataset.

Chaotic World dataset [20] consists of a total of 378,093 annotated instances for triangulating the source of sound for Event Sound Source Localization. The dataset aims to analyze multiple dimensions of a scene to study human behavior in chaotic situations in a comprehensive and detailed manner, and to understand the nature of human behavior (including human actions and interactions) during chaotic events. Sound is critical in the understanding and response to chaotic events in many applications, such as crowd management and emergency response services. Therefore, it is important to provide a comprehensive analysis of human behavior during chaotic events.

In this paper, we use the Chaotic World dataset [20], which contains data on the localization of event sound sources as well as acoustic aspects, to fully understand human behavior in chaotic situations.

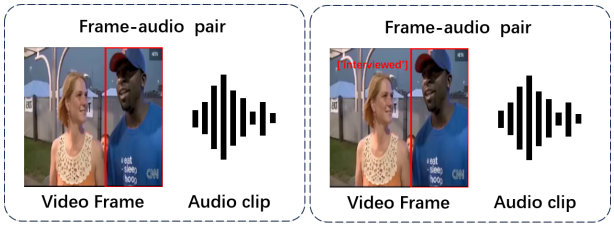

Fig. 1: Left: SSPL (w/ PCM) input data format. Right: SSPL (w/o PCM) input data format.

This paper uses 456 videos from this dataset, 384 training videos, and 72 test videos. We first extract each video frame by frame, and then extract the middle frame of each video and 1.5s of audio before and after the middle frame, for a total of 3s of audio, to form image-audio pairs. After that, we train the model using two random subsets of 1595 image-audio pairs and provide 304 annotated image-audio pairs for evaluation. The annotations in the Chaotic World dataset provide the fps corresponding to the start and end times of the videos and the bounding box coordinates in non-normalized xyxy format from start time to end time, we extract the intermediate video frames and the corresponding audio and annotation boxes based on the video frame index, as shown in

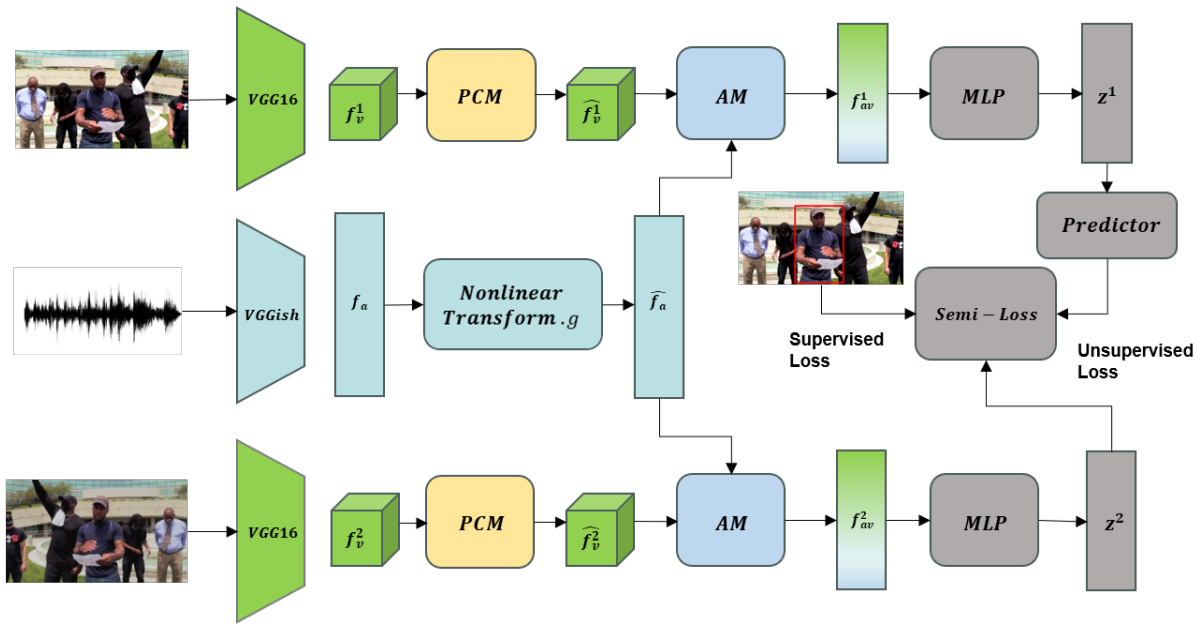

Fig. 2: **Framework of our Semi-Supervised Model SemiPL.** The AM and PCM modules remain consistent with SSPL, with the addition of an unsupervised loss.

Figure 1. For SSPL [13] without PCM(a predictive coding module), we also extract the keywords ssl(type of sound) in the annotations as the class labels, as shown in Figure 1. If there exists only one annotation frame for a certain video, the first annotation frame is selected as the true value.

4. METHOD

In this section, we delved deeper into the influence of various parameters on the effectiveness of SSPL (Self-Supervised Predictive Learning: A Negative-Free Method for Sound Source Localization in Visual Scenes) [13]. Through rigorous experimentation, we observed how changes in these parameters affected the algorithm's accuracy, efficiency, and stability. Additionally, we introduced our novel semi-supervised learning approach, SemiPL, which incorporates advancements in both supervised and unsupervised learning techniques. By leveraging unlabeled data more effectively, SemiPL aims to enhance the overall performance and generalizability of machine learning models, especially in scenarios where labeled data is scarce.

4.1. Parameter adjustments

During the research process, we found that learning rates have a significant impact on the performance of SSPL [13]. Therefore, we attempted to adjust the learning rates and batch size per GPU to observe their effects on the performance of SSPL [13]. We reduced the ssl head learning rate from 5e-5 to 3e-5. Also, we reduce batch size per GPU from 128 to 64. The results and analysis will be provided in the section experiment.

4.2. Semi-supervised Learning

To achieve the same performance, semi-supervised learning requires only a small amount of labeled data and a certain amount of unlabeled data, while self-supervised learning usually requires a large amount of unlabeled data [21]. Therefore, for the Chaotic World dataset, self-supervised learning may not obtain enough information to learn effective feature representations. We devised a semi-supervised method SemiPL under the self-supervised learning setup. We incorporated supervised loss into the SSPL [13] network architecture. To achieve this, we formulated a semi-supervised loss,

$$L_{SSPL} = L_S\left(\hat{y}_i, y_i\right) + L_U\left(p_1, p_2, z_1, z_2\right) \tag{1}$$

where L_U and L_S denote unsupervised and supervised losses respectively. The unsupervised loss L_U function is defined as,

$$\mathrm{L}_U = \frac{1}{2}\left(\mathrm{NCS}(p_1, z_2) + \mathrm{NCS}(p_2, z_1)\right) \tag{2}$$

where p_1 and p_2 are the outputs of the two projection heads of the network, and z_1 and z_2 are the outputs of the corresponding augmented samples, and negative cosine similarity (NCS) is calculated by:

$$\mathrm{NCS}(p, z) = \frac{p \cdot z}{\|p\|\|z\|} \tag{3}$$

L_S utilizes the cross-entropy loss function, defined by,

$$L_S = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right) \quad (4)$$

where N is the size of the flattening pixel of a 224×224 image. $\hat{y}_i$ and y_i are one-dimensional vectors obtained by flattening the ground-truth heatmap and the interpolated heatmap of the predicted values resized to 224×224 size, respectively.

The rest retains the original model structure.

5. EXPERIMENT

In this section, we will present the experimental results of parameter tuning and SemiPL, analyze their strengths and weaknesses, and discuss potential improvements.

5.1. Evaluation Metric.

We follow [22, 13]and use consensus-IoU (cIoU@0.5), whereby the score of each pixel is computed based on the consensus of multiple annotations [22], as well as Area Under the Curve (AUC) [22] for Event Sound Source Localization (ESSL).

Considering that the annotations provided by the dataset are based on bounding boxes of images sized 328×120, we will first convert the bounding box annotations to size 224×224. Then, we will convert the bounding box annotations into binary maps $\{gt_i\}_{i=1}^{N}$, where N is the number of subjects. Infer maps $\{infer_i\}_{i=1}^{N}$ are binary maps of prediction maps, we use 0.5 for the cIoU threshold in the experiments, cIoU's formula is:

$$\text{cIoU} = \frac{\sum_{i=1}^{N} \text{infer}_i \cap \text{gt}_i}{\sum_{i=1}^{N} [\text{gt}_i + \text{infer}_i \cup (\text{gt}_i == 0)]} \quad (5)$$

$$\text{infer}_i \cap \text{gt}_i = (\text{infer}_i \times \text{gt}_i) \quad (6)$$

$$\text{infer}_i \cup (\text{gt}_i == 0) = (\text{infer}_i \times (\text{gt}_i == 0)) \quad (7)$$

Using the common practice in object detection, we use 0.5 for the cIoU threshold in the experiments.

5.2. Qualitative Results and Analysis.

In qualitative comparisons, we mainly visualize the localization results in figure 3 and figure 4.

We observe that when the scene is complex, the model is prone to overlook target objects (e.g., the first and second rows in Figure 3), while when the scene is simple (e.g., the first, second, and third rows in Figure 4), the model can locate target objects well but cover unrelated background details (e.g., ground and trees). It is hypothesized that because most of the occurring objects in the dataset are people, the model did not learn the vocal characteristics of the rest of the vocal objects (e.g., drums) very well. The semi-supervised model may have been disturbed by the data where some of the vocalized objects were not people, and the training effect was instead reduced.

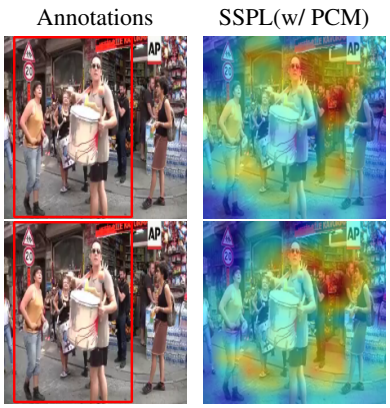

Fig. 3: The first row is the self-supervised model, and the second row is the semi-supervised model SemiPL. It can be seen that self-supervised model has a somewhat larger recognition area for vocalized objects.

Similarly, the model tends to underestimate or overestimate the extent of the sound-emitting object, as evident in the first and second rows of Figure 2. This may be because positive and negative regions in different images cannot be easily distinguished using the same threshold parameter.

In contrast, the SSPL method can cover the regions of interest, and the use of PCM (possibly referring to a specific technique or method) further helps reduce the influence of background noise, resulting in more accurate localization.

5.3. Quantitative Results and Analysis.

Table 1 shows results using different parameters, the best outcome is observed when batch size = 128. Larger batch size improves memory utilization as well as parallelization efficiency for large matrix multiplications, requires fewer iterations to run through an epoch (the full dataset), and is processed faster than a smaller batch size for the same amount of data. Within a certain range, generally speaking, the larger the batch size is, the more accurate the direction of descent it determines, and the less training oscillations it causes. As depicted in Figure 5, when examining the im-

Table 1: Results of Parameter Adjustments

Method	bz	lr	cIoU	AUC
SSPL(Unsupervised)	128	3e-5	41.02	42.23
SSPL(Unsupervised)	64	5e-5	42.03	43.38
SSPL(Unsupervised)	128	5e-5	47.70	44.14

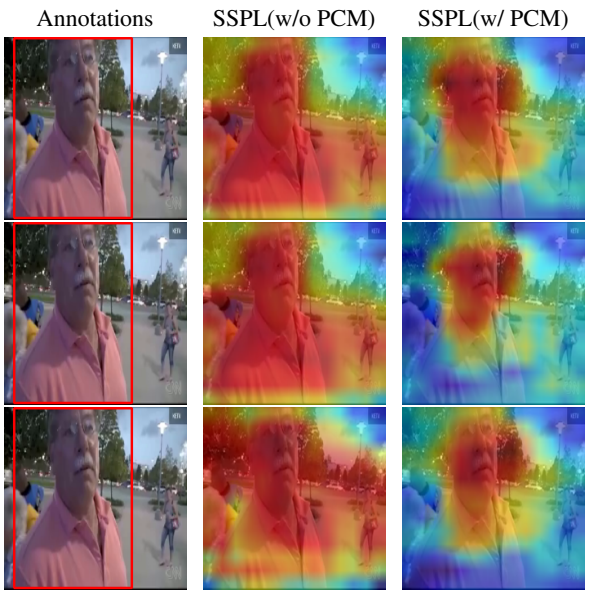

Fig. 4: The first row is batch size 64, learning rate 5e-5, The second row is batch size 128, learning rate 3e-5, The third row is batch size 128, learning rate 5e-5,

pact of varying learning rates on the training process, a noticeable trend emerges. Decreasing the learning rate appears to mitigate the fluctuations in accuracy throughout the training epochs. However, this seemingly beneficial effect comes at a cost: a significant reduction in the convergence speed of the model. In practical terms, this slowdown in convergence renders the training process prohibitively slow, thereby undermining the overall efficacy of the learning algorithm. Consequently, striking a balance between stability and speed becomes paramount in optimizing the learning rate for the given task or model architecture.

Table 2: Results of different kinds of supervision

Methods	cIoU	AUC
LSS(Unsupervised)	23.85	36.56
SemiPL(Semi-supervised)	36.84	41.86
SSPL(Unsupervised)	47.70	44.14

The performance of our SemiPL, while promising, is not quite on par with that of self-supervised models. One potential reason for this could be the relatively large bounding box range used during training. This extensive range may have interfered with the model's ability to learn precise localization, as it introduced too much noise and irrelevant information into the learning process. This hypothesis suggests that refining the bounding box or exploring new data annotation strategies may improve the performance of semi-supervised learning models. Future research in this area is expected to provide valuable insights and advances in the overall development of semi-supervised learning.

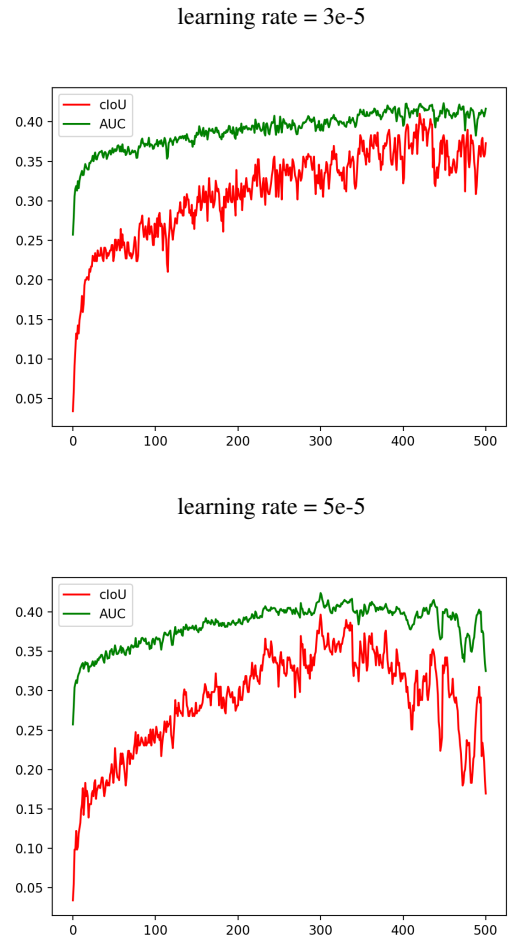

Fig. 5: Different learning rate parameter results. The top learning rate is 3e-5, the bottom learning rate is 5e-5.

6. CONCLUSION AND FUTURE WORK

In this study, we implemented the Self-Supervised Predictive Learning method (SSPL) and our novel Semi-supervised Predictive Learning approach, SemiPL, on the Chaotic World dataset. Our primary objective was to elevate the accuracy of visual-audio localization by explicitly mining positive correspondences. To achieve this, we utilized a tri-stream network coupled with a strategic training regimen to establish the relationship between audio signals and their corresponding video frames within the same video clip. Although SSPL proved effective in single-source audio localization tasks, it fell short in multi-source scenarios primarily due to the limited size of the dataset. Despite this limitation, the expanding availability of data and the escalating importance of label quality indicate that self-supervised learning is destined to become a prevailing trend in machine learning. A potential solution is to develop a fully supervised approach to multi-source localization

tasks, which we leave for future work.

7. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (No. 62203476), Natural Science Foundation of Shenzhen (No. JCYJ20230807120801002).

8. REFERENCES

[1] William W Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[2] Francesca Frassinetti, Nadia Bolognini, and Elisabetta Làdavas, "Enhancement of visual perception by crossmodal visuo-auditory interaction," *Experimental brain research*, vol. 147, pp. 332–343, 2002.

[3] John Hershey and Javier Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," *Advances in neural information processing systems*, vol. 12, 1999.

[4] Jesper E Van Engelen and Holger H Hoos, "A survey on semi-supervised learning," *Machine learning*, vol. 109, no. 2, pp. 373–440, 2020.

[5] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon, "Learning to localize sound source in visual scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4358–4366.

[6] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin, "Multiple sound sources localization from coarse to fine," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 292–308.

[7] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proceedings of the International conference on machine learning (ICML)*, 2021, pp. 12310–12320.

[8] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe, "Whitening for self-supervised representation learning," in *Proceedings of the International conference on machine learning (ICML)*. PMLR, 2021, pp. 3015–3024.

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al., "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

[10] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Oct 2016.

[11] Bruno Korbar, Du Tran, and Lorenzo Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Jun 2018.

[12] Andrew Owens, Jiajun Wu, JoshH. McDermott, WilliamT. Freeman, and Antonio Torralba, "Ambient sound provides supervision for visual learning," *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Aug 2016.

[13] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang, "Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes," 2022.

[14] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," 2016.

[15] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "Vggsound: A large-scale audio-visual dataset," 2020.

[16] Mengyuan Liu, Fanyang Meng, Chen Chen, and Songtao Wu, "Novel motion patterns matter for practical skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023, pp. 1701–1709.

[17] Jinfu Liu, Xinshun Wang, Can Wang, Yuan Gao, and Mengyuan Liu, "Temporal decoupling graph convolutional network for skeleton-based gesture recognition," *IEEE Transactions on Multimedia*, pp. 1–13, 2023.

[18] Jinfu Liu, Runwei Ding, Yuhang Wen, Nan Dai, Fanyang Meng, Shen Zhao, and Mengyuan Liu, "Explore human parsing modality for action recognition," *CAAI Transactions on Intelligence Technology*, 2024.

[19] Yuhang Wen, Zixuan Tang, Yunsheng Pang, Beichen Ding, and Mengyuan Liu, "Interactive spatiotemporal token attention network for skeleton-based general interactive action recognition," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 7886–7892.

[20] Kian Eng Ong, Xun Long Ng, Yanchao Li, Wenjie Ai, Kuangyi Zhao, Si Yong Yeo, and Jun Liu, "Chaotic world: A large and challenging benchmark for human behavior understanding in chaotic events," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 20156–20166.

[21] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer, "S4l: Self-supervised semi-supervised learning," 2019.

[22] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon, "Learning to localize sound source in visual scenes," 2018.