Less Can Be More: Sound Source Localization With a Classification Model

Arda Senocak* Hyeonggon Ryu* Junsik Kim†* In So Kweon
KAIST †Harvard University

Abstract

In this paper, we tackle sound localization as a natural outcome of the audio-visual video classification problem. Differently from the existing sound localization approaches, we do not use any explicit sub-modules or training mechanisms but use simple cross-modal attention on top of the representations learned by a classification loss. Our key contribution is to show that a simple audio-visual classification model has the ability to localize sound sources accurately and to give on par performance with state-of-the-art methods by proving that indeed "less is more". Furthermore, we propose potential applications that can be built based on our model. First, we introduce informative moment selection to enhance the localization task learning in the existing approaches compare to mid-frame usage. Then, we introduce a pseudo bounding box generation procedure that can significantly boost the performance of the existing methods in semi-supervised settings or be used for large-scale automatic annotation with minimal effort from any video dataset.

1. Introduction

We live in an environment full of audio and visual signals. Human perception has been developed to recognize semantic information from raw signals and understand cross-modal relations. In order to achieve human-like perception, modeling of audio-visual modalities is essential. For this reason, the audio-visual learning field is growing fast based on successful advancements in audio [19] and visual [25, 41, 18] recognition.

There are a vast amount of video data piled on the web. However, most audio-visual datasets are not annotated, and the annotation process is labor-intensive. Thereby, most of the research on audio-visual learning is based on self-supervised or weakly supervised learning with category labels to avoid a fine-grained annotation. One key idea in audio-visual learning is to utilize audio-visual correspondence with the assumption that audio and visual information from the same source, *i.e.* video, are correlated. Since audio-visual correspondence is annotation-free information, learning by audio-visual correspondence has been adopted in a broad range of audio-visual studies.

One typical and challenging problem in audio-visual learning is to identify a sound source in a visual scene, which is also known as sound source localization. A simple and straightforward way to learn sound source localization is to use supervision in form of sounding source segmentation mask or bounding box. However, such fine-grained annotation is hard to be obtained in a large scale. Recent studies introduce self-supervised or weakly-supervised learning based on audio-visual relations such as cross-modal atten-

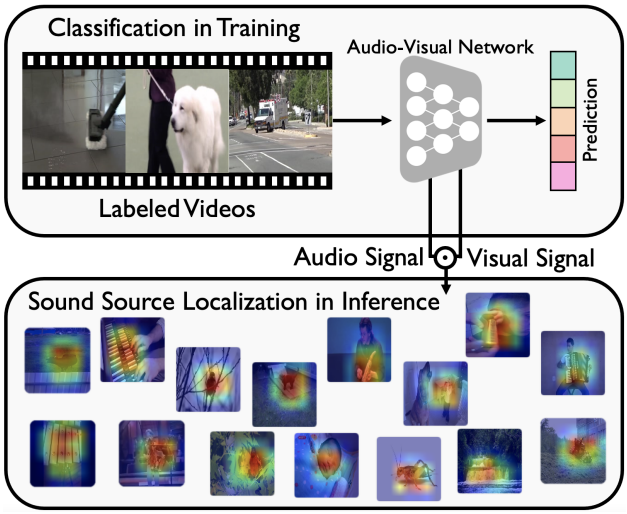

Figure 1: **Localizing Sound Sources by Classifying Videos.** Sound localization appears as an outcome of video classification model without any explicit training for this task. The audio and visual representations learned during training are used to find sounding objects in the scenes.

*Equal contributions alphabetically. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) 2021-0-02068.† The part of this work was done when J. Kim was in KAIST and supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2021R1I1A1A01059778).

tion [39, 40, 7, 27], temporal matching [43, 34, 37, 36, 35], or semantic correspondences [38] along with audio-visual correspondence.

In this paper, we show that a very simple audio-visual classification model works surprisingly well on sound source localization without any additional architectural modification, *e.g*., sub-modules for localization, or complex learning mechanisms. The classification model is trained with cross-entropy loss. Then we apply cross-modal attention to audio-visual features extracted from the classification model to predict sound source location (Figure 1). We quantitatively validate that the sound localization performance of a classification model performs on par with the state-of-the-art methods.

The strong localization capability of a classification model implies that the learned representation is suitable for sound source localization. We further utilize the representation of a classification model to guide sound source localization learning in the existing methods. To this end, we propose informative moment selection and pseudo label generation (in the form of the bounding box) methods from the representation of a classification model. The purpose of the moment selection is to sample moments in a video that captures informative audio-visual events, filtering out uninformative or semantically mismatching moments. The localization result of a classification model is converted to a bounding box form to be used as a pseudo label for semi-supervised sound source localization training in the existing approaches. We validate a sound source localization model guided by a classification model (via generated bounding boxes) gains a significant performance improvement and outperforms the competing methods with a sizable margin.

Our findings imply that semantic supervision, *i.e*. class label, alone is strong information to achieve a competitive sound source localization performance. This observation raises a research question on weakly supervised sound source localization models whether the localization performance is dominantly resulting from a classification loss or from architectural modifications and task-oriented loss designs.

The main contributions of this work are summarized as follows:

- We show that a simple audio-visual classification model is highly capable of localizing sound sources and it performs on par with state-of-the-art methods.
- An informative moment selection using audio-visual features of the classification model is proposed to enhance sound source localization learning.
- Pseudo labels generated by localization results of a classification model in the form of bounding boxes significantly improves the performance of the existing sound source localization models.

2. Related Work

2.1. Audio-visual representation learning

The goal of audio-visual representation learning is to learn a general representation that achieves high performance when adapted to downstream tasks such as image, audio, video classification or sound source localization. For the purpose of large scale learning without annotations, representation learning has been progressed based on a self-supervised learning. A series of studies have shown that audio and visual contents in a video are correlated, thereby a visual representation learned by sound prediction [33] or audio representation distilled from visual representation [4, 12] show strong performance. Later, a variety of joint audio-visual representation learning methods are proposed with an assumption that there is a semantic [2, 20, 31, 30] or temporal [32, 23] correspondence between them. The representation learned jointly by a simple audio-visual correspondence [2] are shown to be more effective than the representations without joint learning [4, 33]. However, simply learning by audio-visual correspondence by instance discrimination ignores similarity of audio-visual contents between samples, *i.e*., videos, leading to a sub-optimal representation. In order to mitigate this issue, clustering [20], sampling [31], weighting [30], and hard mining [23] are proposed.

2.2. Sound source localization

Similar to audio-visual representation learning, the localization of the sound source has been progressed by exploiting audio-visual correspondences. In [3], an audio-visual representation is learned by binary classification whether the features from each modality corresponds or not. Another widely used approach for sound source localization is the cross-modal attention [39, 40, 7, 27]. The cross-modal attention [39, 40] is used to refine visual features by feature weighting with sound source probability map. Later, the attention based methods are improved by intra-frame hard sample mining [7] and iterative contrastive learning with pseudo labels [27]. The aforementioned methods localize sound source regions but do not inference category information. In order to understand the category of sound sources, [21] propose to use an object dictionary and train a model with distribution matching. In addition to the localization, there has been an attempt to localize sounding objects and recover the separated sounds simultaneously, also known as the cocktail party problem [17, 29]. The separation of sound mixture is achieved by predicting masks of spectrogram guided by visual features [9, 51, 50, 13, 49, 10, 1, 52, 14, 45, 42]. Furthermore, a number of recent papers are presented on audio-visual navigation for a given sound source [6, 11].

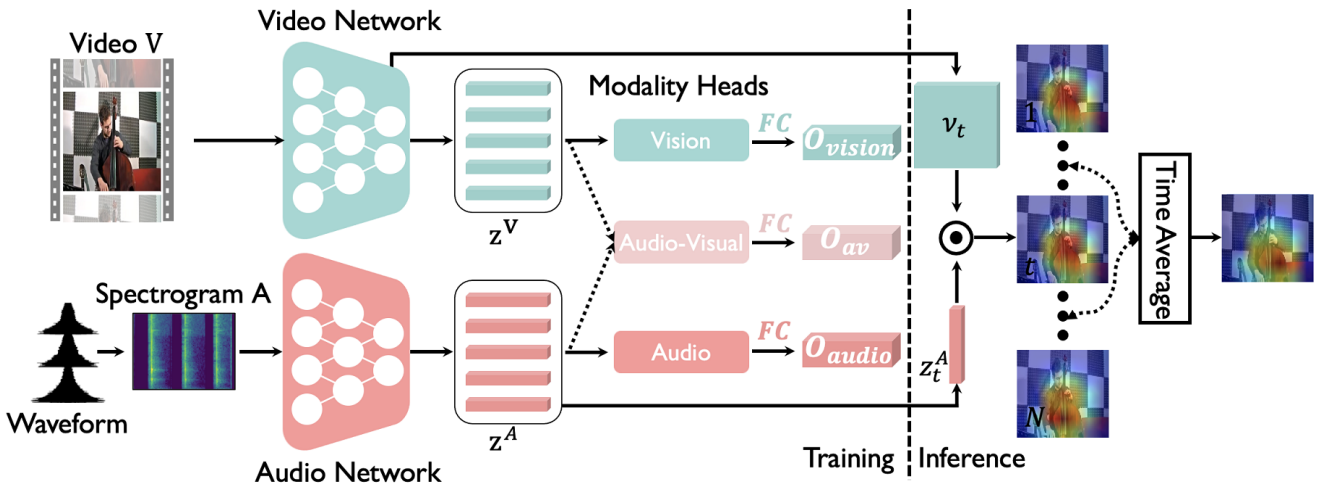

Figure 2: **Algorithm overview.** The model consists of video and audio backbone networks that extract video-level features, z^V and z^A, for each modality. Modality heads take these features as input based on the modality type. After training, sound localization responses are obtained by the dot product of the visual activation from the last convolution layer of the video network, $\mathcal{V}_t$, and the audio embedding on every temporal time step t.

2.3. Video classification

Various deep learning based approaches have been proposed for video classification (or action recognition) by incorporating visual and audio modalities. The effort to boost an accuracy of video classifiaction has been proposed using attention mechanism [28], sub-sequence sampling by saliency [24], mid-level fusion [22], hierarchical synchronization [48], knowledge distillation [15] and gradient blending [46]. However, these advancements in video classification are based on architectural modification [28, 22, 48] and complex learning and inference strategy [15]. Although there are advanced video classification approaches, in this paper, we focus on a basic audio-visual classification model trained by classification loss to investigate the sound source localization capability of a classification model. Our classification model follows the simplified version of [46] where the architectural modification or complex loss design is not required.

2.4. Weakly supervised audio-visual learning

Unlike a video classification task, spatial localization (sound source localization) or temporal localization (audio-visual event localization) require a fine grained annotation in the form of masking, bounding box, or time stamps. However, this is not an affordable annotation cost when the data size is large. Weakly supervised audio-visual learning aims to alleviate the annotation cost by leveraging category information, *i.e*. video tags. Audio-visual event localization methods are proposed using a global and local feature fusion [26], and bi-directional attention matching between global and local features [47]. Some other works tackle both spatial and temporal localizations [43, 34, 37, 36, 35] by exploiting off-the-shelf object and sound proposals [34], cross-modal attention with a recurrent model [43, 37, 36], or use Grad-CAM [38] to localize visually semantic regions [35]. Unlike the recently proposed weakly supervised models, we investigate that a simple classification model without any task-specific architecture design performs surprisingly well on sound source localization, competing with state-of-the-art methods.

3. Approach

3.1. Problem Formulation

The goal of our model is to localize sound sources that are visible in a video as a natural outcome of video classification. Most of the existing works [1, 2, 3, 7, 27, 35, 39, 40] use either audio-visual correspondence or contrastive learning based methods to reveal the sounding object location and the correlation between audio and visual signals. During these training schemes, they incorporate positive and negative samples.

Different from the aforementioned methods, we investigate the possibility of sound localization as a result of multi-modal (audio-visual) video classification without using any explicit sub-modules or training mechanisms and we have only access to a video-level class information.

3.2. Architecture

Similar to [46], we build our multimodal video classification model with individual modality heads, *e.g*., *audio, vision, audio-visual*, and multi-task learning as in Figure 2. Differently, we design our backbone two-stream networks

to represent the entire video with temporally more fine-grained features (per-frame) for each modality.

Backbone Networks. Given video clip $\mathbf{V}$ with its corresponding audio $\mathbf{A}$, our backbone networks extract features for each modality. We use a two-stream architecture similar to other existing audio-visual learning works. Our backbone networks take an entire sequence of video and audio frames and extract features per frame for each modality. The video network is a spatio-temporal network, similar to MCx [44], that is the mixed convolution networks starting with 3D convolutions and followed by 2D convolutions. It takes a video $\mathbf{V}$ of T frames as input and generates a video embedding $\mathbf{z}^{\mathbf{V}}$ with dimensions $T \times D$. The audio network, similar to [1], takes the log-mel spectrogram $\mathbf{A}$ of $10T$ frames and passes it through 2D convolution layers to extract an audio embedding $\mathbf{z}^{\mathbf{A}}$ with dimensions $T \times D$ similar to video features. Thus, there is a corresponding audio feature for every video feature and we do not need any replication or tile operations to match audio and video feature dimensions.

Modality Heads. As shown in Figure 2, our model contains individual modality heads. Defining i the index of each head, the head takes $\mathbf{z}^{\mathbf{V}}$ and $\mathbf{z}^{\mathbf{A}}$ from backbone networks and outputs video-level prediction O_i. We explain each modality head detail below.

- **Audio-Visual Head.** This head is designed to integrate audio and visual signals by performing temporal aggregation to each frame features from both modalities. The integrated audio-visual feature $\mathbf{z}_{av}$ can be computed as follows:

$$\mathbf{z}_{av} = \frac{1}{T}\sum_{t=1}^{T} \texttt{concat}(\mathbf{z}_t^V, \mathbf{z}_t^A), \tag{1}$$

  where `concat` denotes the concatenation of two vectors and t denotes time step. The multi-modal head feature $\mathbf{z}_{av}$ is obtained by temporal aggregation over all time steps T by average pooling.

- **Vision Head.** Vision head assigns zero-valued audio features for each visual frame feature and applies Eq. (1) to compute $\mathbf{z}_{vision}$.

- **Audio Head.** Conversely to the vision head, audio modality head assigns zero-valued visual features to each audio frame feature and computes $\mathbf{z}_{audio}$ with Eq. (1).

3.3. Training

With the proposed backbone networks and modality heads, we obtain different representations from each modality head by given identical inputs. To make each head produce the final C-class prediction output O_i, letting i be the index of each head, separate fully-connected layers are used as in Figure 2. Thus, we propose multi-task joint learning with multiple objectives. This training methodology uses individual heads with their loss functions and supervisory label, where the same single label is given for each task. Considering this as a classification problem, cross entropy loss for each modality head is computed as:

$$\mathcal{L}_i(O_i, y) \quad \text{where } O_i = \texttt{FC}_i(\mathbf{z}_i), \tag{2}$$

where $i \in \{vision, audio, av\}$, $\mathcal{L}$ is the cross entropy loss, `FC` is the linear classifier, O and y are the prediction output and ground-truth label respectively. The final learning objective of the network is minimizing the sum of individual losses:

$$\mathcal{L}_{multi} = \mathcal{L}_{visual} + \mathcal{L}_{audio} + \mathcal{L}_{av}. \tag{3}$$

where each loss has equal weight. This type of losses (auxiliary losses) are commonly used in multi-task learning schemes and we use it for multimodal learning as in [46].

3.4. Sound Localization

After training our model with the video classification objective, we leverage backbone features, $\mathbf{z}^{\mathbf{V}}$ and $\mathbf{z}^{\mathbf{A}}$, to have sound localization results as a natural output without any explicit sound localization operation or module during training. Considering backbone features have $T \times D$ dimensions, sound localization responses can be computed as $\boldsymbol{\alpha}_t = \mathcal{V}_t \cdot \mathbf{z}_t^A$, $\mathcal{V}_t \in \mathbb{R}^{H' \times W' \times D}$ is the visual activation from the last convolution layer of video backbone network and $\mathbf{z}_t^A$ is the audio embedding at moment t within a video [40, 1]. As recent work [7] uses 3 sec. audio segments around mid-frame, we compute the 3 sec. time average of localization responses around the reference time t. The final sound localization result is computed as follows:

$$\boldsymbol{\alpha}_{final} = \frac{1}{|\mathcal{K}|}\sum_{i \in \mathcal{K}} \boldsymbol{\alpha}_{\mathtt{i}}, \tag{4}$$

where $\mathcal{K} = [t - \frac{d}{2}, t + \frac{d}{2}]$ and d is the duration, *e.g.*, 30 for 3 seconds when the video frame rate is 10fps. It is noteworthy that our sound localization computation is different than Grad-cam [38, 5] based approaches as it is cross-modal attention by using learned audio and visual features.

4. Experiments

4.1. Datasets

We train our method on VGGSound [8] dataset and test on VGG-SS [7] and SoundNet-Flickr [39] test sets for quantitative analysis. Additionally, AVE [43] is used for training and visualization for qualitative analysis only. VGGSound is a recently released audio-visual dataset containing around 200K videos. The AVE is an audio-visual

Method	cIoU	AUC
Attention10k [39]	0.185	0.302
AVEL [43]	0.291	0.348
AVobject [1]	0.297	0.357
LVS [7]†	0.303	0.364
Ours	**0.322**	**0.366**

Table 1: **Quantitative results on the VGG-SS test set**. All models are trained on VGG-Sound 144k and tested on VGG-SS. † is the result of the model released on the official project page and the authors report 3% drop in cIoU performance comparing to their paper.

dataset formed for audio-visual event localization. Datasets that are used for quantitative analysis, VGG-SS and Flickr-SoundNet-test, have spatial localization annotations and contain around 5K and 250 samples, respectively.

4.2. Implementation Details

The input audio is 10 seconds. We sample audio data with 16kHz sampling rate. As recent studies[48, 1] do, we compute log-mel spectrogram with size of 1000×80. We use MC3-18[44] as the video network and it takes 100 frames, 10 sec. video at 10fps, of size 112×112 as input. Thus, T = 100 time steps in Section 3.1. We train the network using SGD optimizer with starting learning rate 1×10^2 and reduce it by a factor of 10 if validation accuracy does not increase for 3 epochs. See Supp. for network architectures.

4.3. Quantitative Results

We first compare our sound localization results with existing approaches on the VGG-SS dataset. All of the methods that are used for comparison are trained with the same amount of training data. As shown in Table 1, even though our method is not trained with the sound localization objective, it still outperforms (0.303% *vs.* 0.322%) or gives comparable results to explicitly trained sound localization methods. We extend the comparison to another existing method [43], which is similar to our proposed method, that uses video labels but also incorporates audio-visual attention module into training. Our proposed architecture gives a higher performance in this comparison as well. These results justify that the proposed method has the ability to localize sound sources as a natural outcome of the model without explicit learning mechanisms or specialized modules in any video based dataset, *i.e*. VGGSound and AVE.

In order to see the effect of architectural design choices and training multi-modalities with single optimization, we report ablative evaluation in Table 2. As [46] explains that naive training of multi-modalities jointly by late fusion is not optimal for video classification task, our ablation study also shows similar trend on sound localization task that *Single Multi-Modal* setting performs lower than multi-task settings. Single Multi-Modal setup here is designed as using the only audio-visual head with single classification loss optimization by disabling audio, and vision heads in Figure 2. Additionally, we try another setting that uses multi-task learning with multi-modality heads, as in Eq. (2), but using *one Shared FC* layer rather than *Individual FCs*. Since this architecture is still not a single optimization model, it performs better than *Single Multi-Modal* but gives lower performance than our design choice as *one Shared FC* layer suffers from handling different modality outputs at once.

Method	cIoU	AUC
Single Multi-Modal	0.307	0.355
Shared FC	0.313	0.359
Individual FCs	**0.322**	**0.366**

Table 2: **Ablation Study**. We investigate the effects of architectural choices in the proposed method.

4.4. Qualitative Results

In qualitative comparisons, we mainly visualize the localization response of our method and compare it with other existing methods based on the used test set.

VGG-SS. We visualize the attention maps of VGG-SS test samples in Figure 3 and compare them with the state-of-the-art [7] method on this dataset. Our results are more accurate in comparison to the competing approach. As seen in the orchestra example that includes both orchestra people behind and the musician who is playing cornet, our method focuses on the location that cornet sound comes from as it is the sound source. However, LVS [7] attends to a wider area that contains the entire stage.

AVE. Our algorithm can work on any video dataset with label information. Thus, we train our network on the AVE [43] dataset as well. However, since this dataset does not contain spatial localization annotation, we can only show the qualitative results. Figure 4 shows qualitative results compare to the method that this dataset is introduced. As it can be seen, not only our results are more accurate, they also cover a wider area of the sounding objects.

Other Results. At first glance, it may look like our method attends to the moving areas or spots in the visual scene rather than the context of the sound. It should be noted that our network outputs accurate results even on video clips that have a fixed frame during the entire video as in Figure 5. This means that our method localizes a visual region guided by sound regardless of a motion cue.

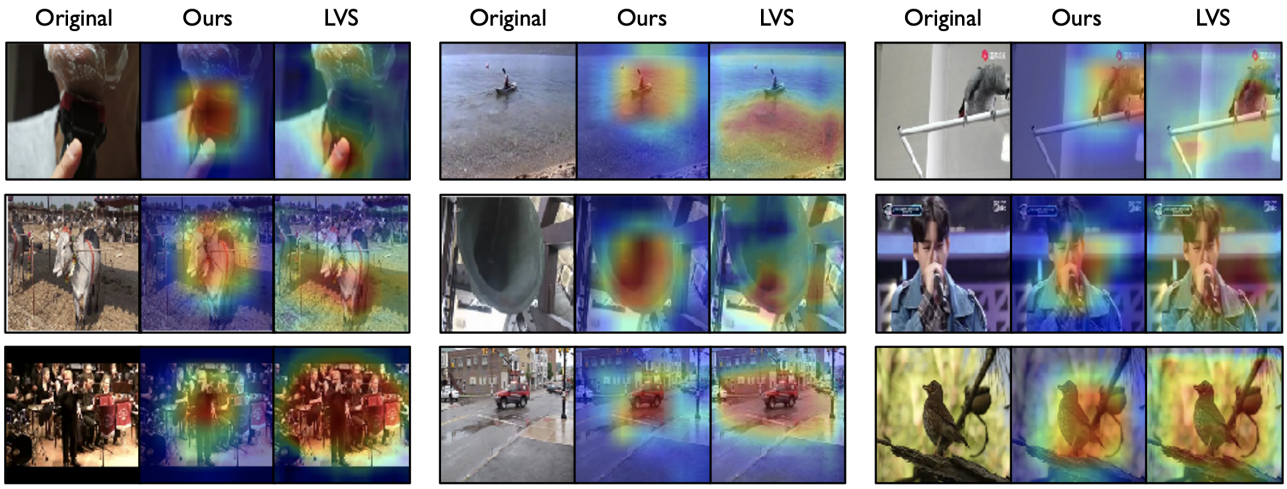

Figure 3: **Sound Localization Results on VGG-SS and comparison with LVS [7].**

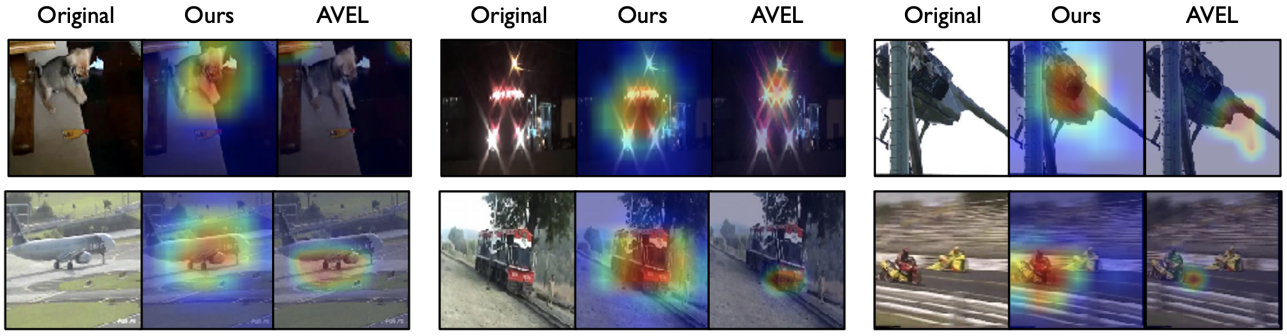

Figure 4: **Sound Localization Results on AVE and comparison with AVEL [43].**

Figure 5: **Sound Localization Responses on Fixed-frame Videos.** These video samples contain one fixed frame during the entire 10 sec. video. Results show that our network does not only focus on moving spots but object contexts as well.

5. Applications

5.1. Informative Moment Selection

Recent work [7] uses the center frames of the videos for sound localization training on the VGGSound dataset. However, using a mid-frame can bring noisy, non-informative frames/segments for training as in Figure 6. Examples in the first row show that the audio segment corresponding to the mid-frame may not contain an informative audio signal. Also, the second row depicts that mid-frame may contain inappropriate visual signals. Thus, knowing which frame/segment in the video is useful and informative for sound localization training is an important issue to tackle. Thus, we pose a problem to select an informative time step within a video to use as an alternative to a mid-frame in training.

Our backbone networks enable us to have audio and visual features at fine temporal time steps within a video. We can select a time step that has the highest correlation score between audio and visual features and use this time step as an alternative to the mid-frame. The assigned task here is performed by finding the time steps (moments) that have the highest correlation scores between audio and visual features, $\mathbf{z}^{\mathbf{V}}$ and $\mathbf{z}^{\mathbf{A}}$ respectively. Correlation scores are computed by pairwise dot products between audio and visual embeddings [16, 1] at the same time step and computed as $\mathbf{S}_{av}[t] = \mathbf{z}_t^V \cdot \mathbf{z}_t^A$. Later, time steps in S_{av} are sorted by $\texttt{top-k}(\mathbf{S}_{av})$ and alternative frames to mid-frame is selected. In this paper, we used top-1 moment for training. Figure 6 shows some of the selected moments.

To show that these alternative selected moments to mid-frames are beneficial, we report sound localization results as a comparison on Flickr-SoundNet and VGG-SS datasets. We use the publicly available method [39] as a baseline by modifying its audio and visual networks, replacing them

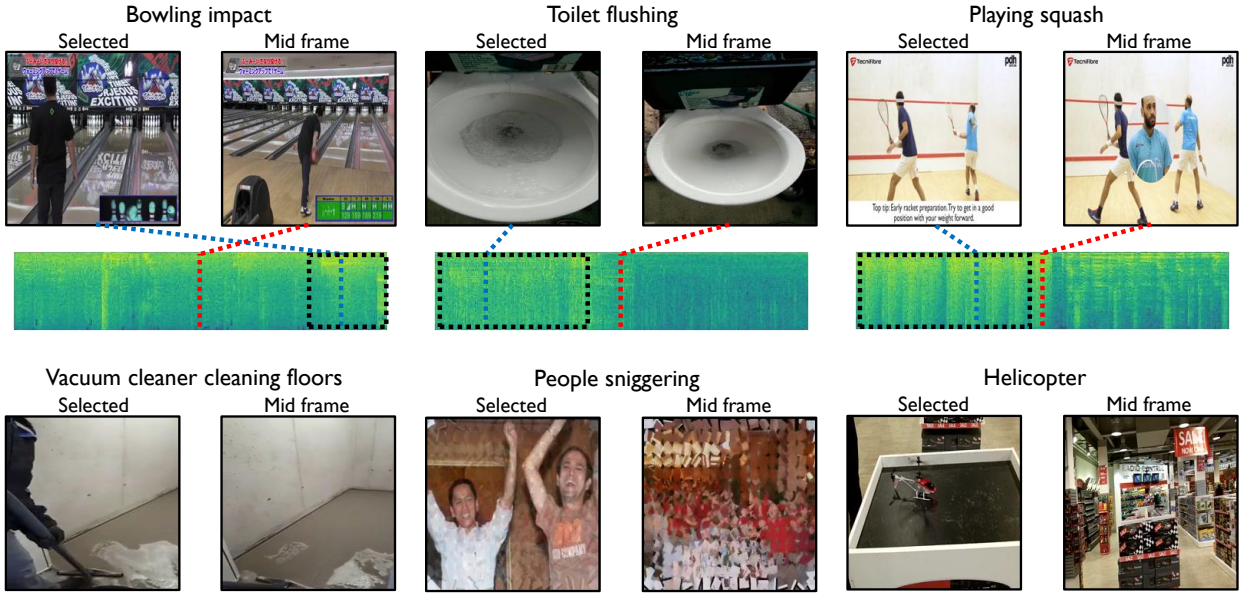

Figure 6: **Qualitative results of mid-frame vs. our selected moment.** LVS[7] uses middle frames for training. The red dashed lines are the mid frame moment, and the blue depicts our selected moment. The black dashed boxes contain the proper information in the audio modality. The first row shows the samples that mid frames have insufficient audio information, *e.g.*, silence or noise. The second row visualizes the mid frames with in-appropriate visual signals whereas our selected moments are proper.

Training frame	Test set	cIoU	AUC
Top-1(Ours)	VGG-SS	**0.264**	**0.318**
Mid-Frame	VGG-SS	0.256	0.315
Top-1(Ours)	Flickr	**0.764**	**0.597**
Mid-Frame	Flickr	0.749	0.590

Table 3: **Quantitative results of Top-1 moment *vs.* mid-frame.** The modified [39] model is trained on VGGSound with mid-frame or Top-1 moment selection and tested on VGG-SS and Flickr-SoundNet test sets.

with ResNet-18. See Supp. for details. We fix the training set to be VGGSound with the full data. Table 3 shows the results. Top-1 selected moments give 1.5% higher accuracy compare to the mid-frame inputs on Flickr-SoundNet. Similarly, alternative moment inputs perform better than mid-frame on the VGG-SS dataset though the performance gap is smaller (0.8%). Our analysis confirms that our proposed approach has ability to pick informative moments within a video to use in training phase of existing sound localization methods.

5.2. Automatic Bounding Box Generation for Sounding Objects

Having high quality bounding box annotations for sounding objects is very important not only for evaluating audio-visual localization algorithms but also for enabling semi-supervised learning approaches [39, 40]. However, as in other fields, having annotations on a large scale is time-consuming and costly. Automating this process, even without human level precision, will be useful for further research in this community such as it can provide a good starting point for faster manual annotations or provide large number of samples for semi-supervised methods. We use our audio and visual features' correlations for automatic bounding box generation on VGGSound [8].

Instead of using visual object detectors and manual image annotations as in [7, 39, 40], we use the attention map results of informative moments, *i.e.*, top-k time steps that are computed based on the correlation of audio and visual features in the entire video. To get more accurate bounding boxes, we select the top-1 time step that has the highest correlation score and compute the attention map. Then, bounding boxes for sounding objects are generated as follows: 1) binarize the attention map with thresholding and it results in connected segments of pixels, and 2) draw a bounding box around the single largest segment. We emphasize that some of these automatically computed annotation boxes are not as precise and accurate as they can be in human annotation (second row of Figure 7). However, they are sufficient to use in sub-tasks, such as helping human annotators for faster annotation or semi-supervised sound localization. We show qualitative results of our automatic bounding box gen-

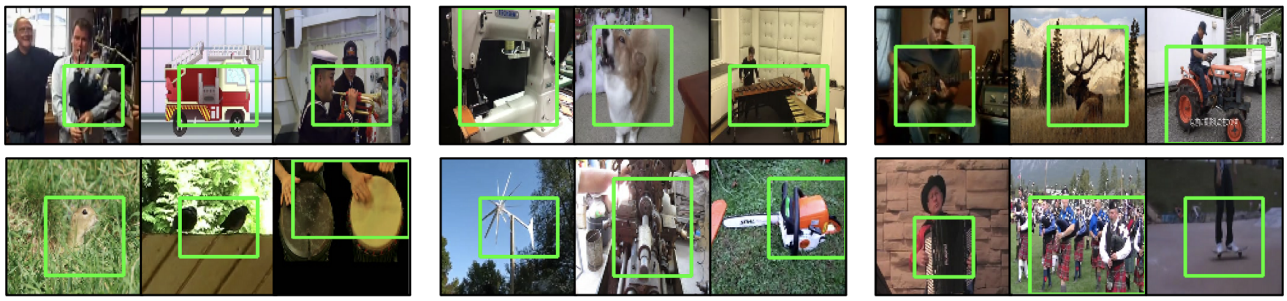

Figure 7: **Qualitative Results of Automatic Bounding Box Generation.** Our method accurately generates bboxes for sounding objects to use in sub-tasks, such as faster human annotation and semi-supervised sound localization.

Training set	Test set	cIoU	AUC
w/o bbox	VGG-SS	0.264	0.318
w/ bbox	VGG-SS	**0.382**	**0.393**
w/o bbox	Flickr	0.764	0.597
w/ bbox	Flickr	**0.805**	**0.617**

Table 4: **Performance evaluation of VGG-SS and Flickr-SoundNet test sets with auto generated bboxes.** The modified [39] model is trained on full VGGSound together with 2.5K pseudo-bboxes generated from our proposed method. Results show that our auto generated bboxes significantly boost the performance.

eration in Figure 7. As it can be seen, giving these bounding boxes to the annotators can be helpful in reducing down the time of annotation drastically.

To further demonstrate the usefulness of our auto generated bounding boxes, we use bounding boxes in the existing semi-supervised sound localization approach [39, 40]. Since this kind of learning setup requires a set of supervised samples that have ground truth sounding area, it is not easily applicable to different and new datasets or scaling up the number of supervised samples requires tremendous effort. Existing datasets have either only test set with annotation [7] or have supervised training set in a single dataset with a small amount of data [39]. By using our method, we can address these shortcomings. Firstly, we automatically build a new set of samples with pseudo bounding boxes on recent VGG-Sound dataset. It contains 2500 samples. However, it can be easily extended to any number. Later, we use the modified version of [39], introduced in Section 5.1, by incorporating our auto generated annotations. Table 4 shows the quantitative results of semi-supervised learning setup. Even though these boxes are automatically generated, not as precise as human annotations, they are still useful to give informative cues to the model in semi-supervised learning setting as it shows 11.8% and 4.1% improvements on VGG-SS and SoundNet-Flickr respectively. This shows that our auto generated bounding boxes are indeed informative and they can significantly boost the sound localization performance of the relatively older methods on standard benchmarks, and even surpass the state-of-the-art methods. To the best of our knowledge, there is no other recent work that provides annotated training samples for the VGGSound dataset and reports semi-supervised performance on VGG-SS. It is also noteworthy that our method enables to generate any number of samples with their bounding box in any video based dataset.

6. Conclusion

We present a multimodal video classification model that learns sound source localization as a natural outcome of the model. Unlike previous sound localization models, we do not use any explicit training mechanism or sub-module for this task. However, it achieves on par performance with state-of-the-art methods. Moreover, we propose interesting potential applications that can be built based on our model. Firstly, we introduce informative moment selection, an alternative to mid-frame usage, for enhancing the sound source localization learning in the existing approaches. Secondly, we introduce an automatic bounding box generation ability of our model for sounding objects. This can be potentially very useful to the community as it provides a good starting point for faster human annotation for dataset construction or use these bounding boxes to boost the performance of the existing sound localization methods in semi-supervised setting. With the moment selection or bounding box generation, we show that an accurate sound source localization model can be trained and it leads to significant performance improvements.

The empirical analyses in our paper demonstrate that category supervision with the simple learning process achieves competitive performance with competing sound source localization models. This motivates us to take a closer look at weakly supervised sound source localization methods to examine whether a dominant learning signal attributes to category supervision or task-oriented techniques. In future work, we will analyze the significance of each sound source localization oriented technique with and without category supervision.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, 2020.

[2] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision*, 2017.

[3] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision*, 2018.

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems*, 2016.

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2018.

[6] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. 2020.

[7] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[8] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2020.

[9] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 2018.

[10] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[11] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B. Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. 2020.

[12] Chuang Gan, Hang Zhao, Peihao Chen, David D. Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. 2019.

[13] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *IEEE International Conference on Computer Vision*, 2019.

[14] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[15] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[16] Tavi Halperin, Ariel Ephrat, and Shmuel Peleg. Dynamic temporal alignment of speech to lips. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[17] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[19] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017.

[20] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[21] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 2020.

[22] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *IEEE International Conference on Computer Vision*, 2019.

[23] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, 2018.

[24] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *IEEE International Conference on Computer Vision*, 2019.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 2012.

[26] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[27] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Unsupervised sound localization via iterative contrastive learning. *arXiv preprint arXiv:2104.00315*, 2021.

[28] Xiang Long, Chuang Gan, Gerard De Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. Attention clusters: Purely attention based local feature integration for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[29] Josh H McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009.

[30] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[31] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[32] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision*, 2018.

[33] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision*, 2016.

[34] Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc QK Duong, Patrick Pérez, and Gael Richard. Weakly supervised representation learning for audio-visual scene analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:416–428, 2019.

[35] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *European Conference on Computer Vision*, 2020.

[36] Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2020.

[37] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2020.

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 2017.

[39] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[40] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound sources in visual scenes: Analysis and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1605–1619, 2021.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[42] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[43] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European Conference on Computer Vision*, 2018.

[44] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[45] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel P. W. Ellis, and John R. Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *International Conference on Learning Representations*, 2021.

[46] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[47] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *IEEE International Conference on Computer Vision*, 2019.

[48] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[49] Xudong Xu, Bo Dai, and Lin Dahua. Recursive visual sound separation using minus-plus net. In *IEEE International Conference on Computer Vision*, 2019.

[50] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *IEEE International Conference on Computer Vision*, 2019.

[51] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *European Conference on Computer Vision*, 2018.

[52] Hang Zhou, Xudong Xu, Lin Dahua, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *European Conference on Computer Vision*, 2020.