Learning Sound Localization Better From Semantically Similar Samples

Arda Senocak[1]* Hyeongon Ryu[1]* Junsik Kim[2]* In So Kweon[1]
[1]KAIST [2]Harvard University

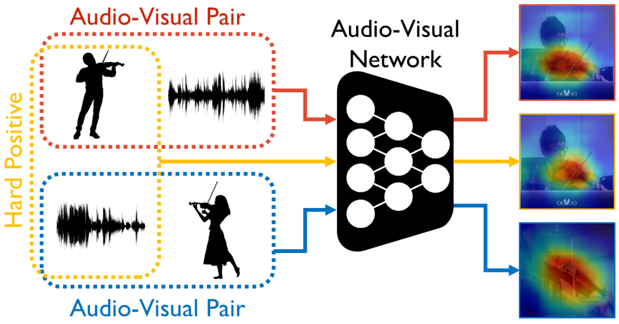

Figure 1. **Localizing Sound Sources by Introducing Hard Positive Samples.** Here, two audio-visual pairs from the red box and blue box are semantically related. When we pair the image of the red box and the audio of the blue box, which forms the yellow box, the response maps from the yellow box and the red box localize similar regions.

1. Introduction

During daily events of our lives, we are continuously exposed to various sensory signals and their interactions with each other. Because of this continuous stream of information, human perception has been developed to organize incoming signals, recognize the semantic information and understand the relationship between these cross-modal signals to combine or separate them. Among these sensory signals, inarguably the most dominant ones are vision and audition. In order to have human-level perceptional understanding, modeling proper audio-visual learning that can associate or separate the audio-visual signals is essential. Thus, the audio-visual learning is an emerging research topic with variety of tasks, such as audio-visual source separation [1, 4–6, 17, 20], audio spatialization [12], audio-visual video understanding [7, 18, 19] and sound localization [2, 3, 14, 15].

One of the main challenging problems in audio-visual learning is to discover a sound source in a visual scene by taking advantage of the correlation between the visual and audio signals. A simple way to do this is using sounding source segmentation masks or bounding boxes as supervision. However, obtaining sufficient annotated data is difficult and costly for audio-visual tasks, where annotation often requires listening and watching samples. As a result, a key challenge is proposing unsupervised approaches, without any manual annotation, that can solve this task successfully. A widely used self-supervised approach in sound localization is using the correspondence between the audio and visual signals by using them as supervision to each other [2, 13–15]. Additionally, [14, 15] use an attention mechanism to refine visual features by feature weighting with sound source probability map. Other models [9] deploy the clustering of audio-visual samples to reveal the cross-modal relationships and improve the accuracy. While prior self-supervised works ignore the category information, [10] incorporates class-aware object dictionaries and distribution alignment in a self-supervised way. More recently, in the light of the success of the noise contrastive learning, [3] introduces a state-of-the-art method that uses a contrastive learning scheme with hard negative mining from background regions within an image.

Inarguably, audio-visual correspondence in a video, *i.e.* audio and image pairs, is the key assumption in sound localization approaches. Vast amount of audio-visual research leverage contrastive learning by assigning corresponding audio-visual pair from the same source as positives while mismatched pairs, *i.e.* audio and visual pairs from different sources, as negatives. One problem of contrastive learning, when applied with this audio-visual correspondence assumption, is that the negative pairs may contain semantically matched audio-visual information. For example, if there are multiple videos with a person playing the violin in a dataset, every pairwise relation of these videos are semantically correct pairs (Figure 1). However, when contrastive learning is applied without consideration of semantic information contained in videos, a model is falsely guided by these false negatives, *i.e.* audio and video from different sources but containing similar semantic information [8, 11].

We mine and incorporate these semantically similar samples into the training. As a consequence, we can reduce down the chance of falsely paired audio-visual samples as negatives during the training. Moreover, since these pairs are semantically correlated, they can be used as positive

*Equal contribution.

pairs, *i.e.* hard positives, in sound localization tasks. For instance, if the audio samples of two different instances are semantically related, then the attention maps of those audios paired with the same base image should be similar to each other as in Figure 1. Note that this observation is valid when the order of the audio and vision samples in the previous scenario are swapped as well. We show this simple approach boosts sound localization performance on standard benchmarks.

To be clear, we do not propose a new architecture or a new loss function in this paper, but instead, we provide a new training mechanism by discovering semantically matched audio-visual pairs and incorporating them as hard positives. We make the following contributions: 1) We demonstrate hard positive audio-visual pairs produce similar localization results to that of corresponding pairs. 2) We mine and incorporate hard positives into the positive set of contrastive loss. 3) We show that incorporating hard positives improves the sound localization performance and outperforms prior works on the standard benchmarks.

2. Approach

Audio-visual attention is commonly used in sound localization studies [3,10,14]. We build the baseline audio-visual model based on the most recent work LVS [3] and validate our method on top of the baseline. We first introduce the audio-visual attention mechanism and the baseline model named vanilla-LVS, and then introduce our approach.

2.1. Preliminaries

We use an image frame $v \in \mathbb{R}^{3\times H_v \times W_v}$ and the corresponding spectrogram $a \in \mathbb{R}^{1\times H_a \times W_a}$ of the audio from a clip $X = \{v, a\}$. With the two-stream models, $f_v(\cdot;\theta_v)$ for vision embedding and $f_a(\cdot;\theta_a)$ for audio embedding, the signals are encoded into the features:

$$\begin{aligned} V &= f_v(v;\theta_v), \quad V \in \mathbb{R}^{c\times h\times w} \\ A &= f_a(a;\theta_a), \quad A \in \mathbb{R}^{c} \end{aligned} \tag{1}$$

The vision and audio features, V_j and A_i, are fed into the localization module and audio-visual response map $\alpha_{ij} \in \mathbb{R}^{h\times w}$ is computed with cosine similarity as:

$$[\alpha_{ij}]_{uv} = \frac{\langle A_i, [V_j]_{:uv}\rangle}{\|A_i\|\ \|[V_j]_{:uv}\|}, \quad uv \in [h]\times[w], \tag{2}$$

where i and j denote the audio and image sample indices respectively. Following [3], pseudo-ground-truth mask m_{ij}, is obtained by thresholding the response map as follows:

$$m_{ij} = \sigma((\alpha_{ij} - \epsilon)/\tau), \tag{3}$$

where σ refers to the sigmoid function, ϵ to the thresholding parameter and τ is the temperature. The inner product between the mask m_{ij} and the response map α_{ij} is computed to emphasize positively correlated regions of the response map.

2.2. Semantically Similar Sample Mining

Our method is based on the observation that hard positives make similar localization results to the original pairs. These additional audio-visual response maps from the hard positives can be easily incorporated into the contrastive learning formulation. Semantically related samples in each modality are mined to form hard positives based on the similarity scores in the feature space. To get the reliable representations within each modality, we train the baseline model without tri-map, *i.e.* hard negative mining introduced in [3].

Given sample i and an arbitrary sample j, we compute the cosine similarity between the features within each modality, $A_i^T A_j$ and $\mathcal{V}_i^T \mathcal{V}_j$, where $\mathcal{V}_i$ is the spatial-wise average pooled vector of visual the feature map V_i. Sets of semantically similar items in both modality for the given sample, $\mathcal{P}_{i_A}$ and $\mathcal{P}_{i_V}$, are constructed based on the computed scores, $\mathcal{S}_{i_A}$ and $\mathcal{S}_{i_V}$. K number of semantically related samples are retrieved from the sorted similarity scores. All the samples that are not contained in the aforementioned sets are considered as negative samples and form a negative set, $\mathcal{N}_i$:

$$\begin{aligned} \mathcal{S}_{i_A} &= \{A_i^T A_j | 1 \leq j \leq n\}, \\ \mathcal{P}_{i_A} &= \{X_t\}_{t\in S[1:K]}, \quad S = argsort(\mathcal{S}_{i_A}), \\ \mathcal{S}_{i_V} &= \{\mathcal{V}_i^T \mathcal{V}_j | 1 \leq j \leq n\}, \\ \mathcal{P}_{i_V} &= \{X_t\}_{t\in S[1:K]}, \quad S = argsort(\mathcal{S}_{i_V}), \\ \mathcal{N}_i &= \overline{\mathcal{P}_{i_A} \cup \mathcal{P}_{i_V}}. \end{aligned} \tag{4}$$

2.3. Training

The training objective of our method makes positive pair responses higher while negative pair responses are reduced. Since we incorporate responses from hard positives, we extend the contrastive learning formulation of [3] by adding each hard positive response. Defining the response map of the base pairs as P_b, hard positive responses are computed as follows: P_a is the response of base visual signal and semantically similar audio, P_v is the response of base audio signal and semantically similar image, finally P_c is the cross-modal hard positive pair's response. All the responses from the base audio and negatively correlated image pairs are considered as the negative responses N_i. Definition of positive and negative responses are given as:

$$
\begin{aligned}
P_b &= \frac{1}{|m_{ii}|}\langle m_{ii},\ \alpha_{ii}\rangle \\
P_a &= \frac{1}{|m_{ji}|}\langle m_{ji},\ \alpha_{ji}\rangle, j \in \mathcal{P}_{i_A} \\
P_v &= \frac{1}{|m_{ik}|}\langle m_{ik},\ \alpha_{ik}\rangle, k \in \mathcal{P}_{i_V} \\
P_c &= \frac{1}{|m_{jk}|}\langle m_{jk},\ \alpha_{jk}\rangle, j \in \mathcal{P}_{i_A}, k \in \mathcal{P}_{i_V} \\
P_i &= \exp(P_b) + \exp(P_a) + \exp(P_v) + \exp(P_c) \\
N_i &= \sum_{l \in \mathcal{N}_i} \exp(\frac{1}{hw}\langle \mathbf{1},\ \alpha_{il}\rangle).
\end{aligned}
\tag{5}
$$

After constructing positive and negative responses, our model can be optimized by the loss function $\mathcal{L}$ as below:

$$
\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\left[\log \frac{P_i}{P_i + N_i}\right] \tag{6}
$$

Method	VGG-SS cIoU	VGG-SS AUC	Flickr cIoU	Flickr AUC
Attention [14]$_{\text{CVPR18}}$	0.185	0.302	0.660	0.558
AVEL [16]$_{\text{ECCV18}}$	0.291	0.348	-	-
AVObject [1]$_{\text{ECCV20}}$	0.297	0.357	-	-
Vanilla-LVS$_{\text{CVPR21}}$	0.278	0.350	0.692	0.563
LVS [3]†$_{\text{CVPR21}}$	0.303	0.364	0.724	0.578
Random HP	0.207	0.314	0.572	0.528
Ours	**0.346**	**0.380**	**0.768**	**0.592**

Table 1. **Quantitative results on the VGG-SS and SoundNet-Flickr test sets**. All models are trained with 144K samples from VGG-Sound and tested on VGG-SS and SoundNet-Flickr. † 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.

3. Experiments

3.1. Quantitative Results

In this section, we compare our results with existing sound localization approaches on VGG-SS and SoundNet-Flickr-Test datasets. We recall that our model is based on vanilla-LVS trained with semantically similar sample mining. In Table 1, we show the performance of the proposed model together with several prior works on VGG-SS and SoundNet-Flickr-Test datasets by following [3]. The comparison methods here are trained with the same amount of training data, 144K, as in [3]. AVEL [16] and AVObject [1] models are based on video input. Thus, the SoundNet-Flickr dataset, which contains static image and audio pairs, can not be evaluated. The proposed model achieves significantly higher performance on both the VGG-SS and SoundNet-Flickr-Test datasets than the other existing works including LVS. This shows that inter-sample relation across-modality is more important than the intra-sample relation.

Method	cIoU	AUC
Attention [14]$_{\text{CVPR18}}$	0.660	0.558
Vanilla-LVS$_{\text{CVPR21}}$	0.704	0.581
LVS [3]†$_{\text{CVPR21}}$	0.672	0.562
Ours	**0.752**	**0.597**

Table 2. **Quantitative results on the SoundNet-Flickr test set.** All models are trained and tested on the SoundNet-Flickr dataset. † is the result of the model from the official project page.

Next, we compare the performance on the SoundNet-Flickr-Test set by training our method separately with 144K samples from SoundNet-Flickr. As shown in Table 2, the proposed model gives the highest performance in this comparison. As [3] reports, our method also achieves higher accuracy on this test set when it is trained with VGGSound.

3.2. Qualitative Results

We provide our sound localization results on VGG-SS and SoundNet-Flickr test samples in Figure 2 and compare them with LVS [3]. Our results present more accurate response maps in comparison to the competing approach. Additionally, we demonstrate attention maps of the hard positives. The left part of the Figure 3 visualizes the scenario where hard positive is obtained with the visual signal. Here, the original audio is paired with the semantically related image. Similarly, the right part of the Figure 3 depicts that the hard positive is obtained with the audio signal. Results show that in both scenarios, response maps of the hard positives are very similar to the response maps of the original pairs, $att_{original}$. As expected, response maps of the negative samples, att_{neg}, are inaccurate since those instances are not correspondent.

4. Conclusion

In this paper, we address the problem of self-supervised sound source localization in contrastive learning formulation. We identify a source of noise due to the random negative sampling, where semantically correlated pairs are contained among negative samples. We suggest a simple positive mining approach and employ these hard positives into training. We validate our method on standard benchmarks showing state-of-the-art performance. The proposed method is applicable to any type of model using contrastive loss, therefore audio-visual learning studies can benefit from our work.

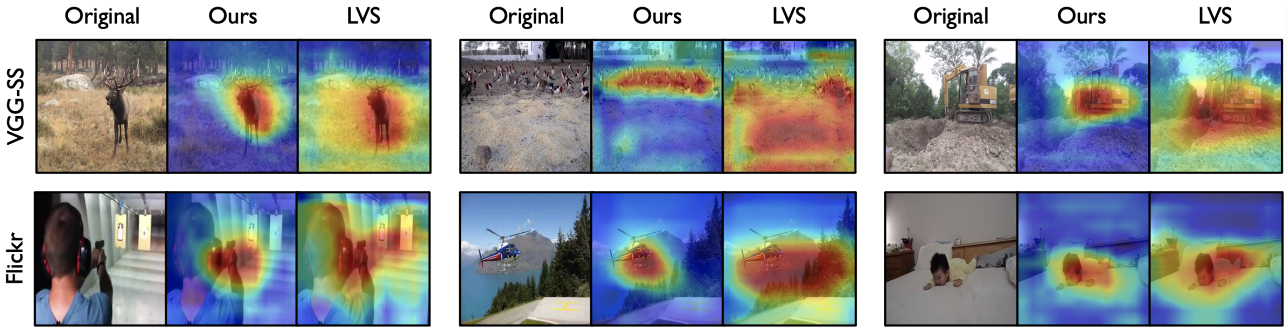

Figure 2. **Sound localization results on VGG-SS and SoundNet-Flickr and comparison with LVS [3].**

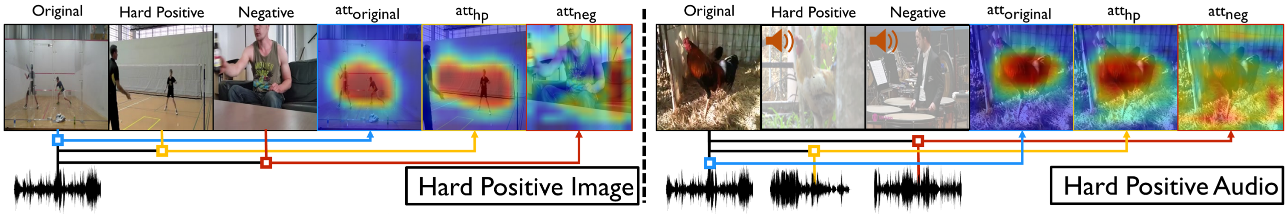

Figure 3. **Response maps of hard positives.** Left refers to the scenario that the hard positive is obtained with the visual signal. Right shows the hard positive pair which is obtained with the audio signal. In both cases, the response maps of hard positives resemble the original pair response maps.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *ECCV*, 2020. 1, 3

[2] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 1

[3] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *CVPR*, 2021. 1, 2, 3, 4

[4] 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 TOG*, 2018. 1

[5] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019. 1

[6] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021. 1

[7] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020. 1

[8] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *NeurIPS*, 2020. 1

[9] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, 2019. 1

[10] 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. *NeurIPS*, 2020. 1, 2

[11] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *CVPR*, 2021. 1

[12] Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360° video. In *NeurIPS*, 2018. 1

[13] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 1

[14] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 1, 2, 3

[15] 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. *TPAMI*, 43(5):1605–1619, 2021. 1

[16] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 3

[17] 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 *ICLR*, 2021. 1

[18] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020. 1

[19] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 1

[20] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019. 1