Sound Source Localization is All about Cross-Modal Alignment

Arda Senocak[1*] Hyeonggon Ryu[1*] Junsik Kim[2*] Tae-Hyun Oh[3,4]
Hanspeter Pfister[2] Joon Son Chung[1]

[1] Korea Advanced Institute of Science and Technology [2] Harvard University
[3]Dept. of Electrical Engineering and Grad. School of Artificial Intelligence, POSTECH
[4]Institute for Convergence Research and Education in Advanced Technology, Yonsei University

Abstract

Humans can easily perceive the direction of sound sources in a visual scene, termed sound source localization. Recent studies on learning-based sound source localization have mainly explored the problem from a localization perspective. However, prior arts and existing benchmarks do not account for a more important aspect of the problem, cross-modal semantic understanding, which is essential for genuine sound source localization. Cross-modal semantic understanding is important in understanding semantically mismatched audio-visual events, e.g., silent objects, or off-screen sounds. To account for this, we propose a cross-modal alignment task as a joint task with sound source localization to better learn the interaction between audio and visual modalities. Thereby, we achieve high localization performance with strong cross-modal semantic understanding. Our method outperforms the state-of-the-art approaches in both sound source localization and cross-modal retrieval. Our work suggests that jointly tackling both tasks is necessary to conquer genuine sound source localization.

1. Introduction

Humans can easily perceive where the sound comes from in a scene. We naturally attend to the sounding direction and associate incoming audio-visual signals to understand the event. To achieve human-level audio-visual perception, sound source localization in visual scenes has been extensively studied [50, 51, 4, 47, 8, 35, 31, 33, 53, 54, 52, 36, 39, 38, 20]. Motivated by that humans learn from natural audio-visual correspondences without explicit supervision, most of the studies have been developed on a fundamental assumption that audio and visual signals are temporally correlated. With the assumption, losses of the sound source localization task are modeled by audio-visual correspondence as a self-supervision signal and are implemented by contrasting audio-visual pairs, *i.e.*, contrastive learning.

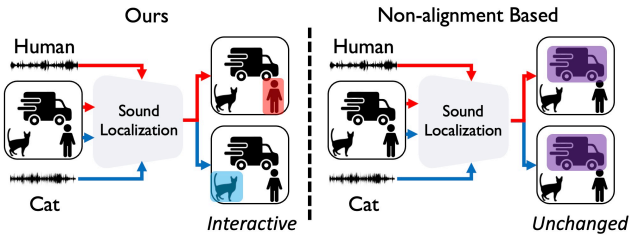

Figure 1. **A conceptual difference between prior approaches and our alignment-based sound source localization.**

While these approaches appear to be unsupervised methods, they strongly rely on partial supervision information; *e.g.*, using supervisedly pretrained vision networks [50, 51, 47, 53, 54, 20] and visual objectness estimators for post-processing [39, 38]. Without leveraging such strong initial representations, the performance is degraded. Thus, the previous methods are not purely self-supervised approaches. Even further, there are recent studies [45, 39, 38] that point out visual objectness bias in existing sound source localization benchmarks and exploit the objectness prior to improve the localization accuracy. They show that, even without interaction between visual and audio signals, a model may achieve strong accuracy in localization by only referring visual signals alone, which is not the true intention of the sound source localization task. In short, the current evaluation and setting of the sound source localization do not capture the true sound source localization performance.

In this work, we first sort out evaluating sound source localization methods by introducing a cross-modal retrieval task as an auxiliary evaluation task. By this task, we can measure whether the learned representation have the capability to accurately interact between audio and visual modalities; *i.e.*, more fine-grained audio-visual correspondence which is essential for genuine sound source localization. This aspect has been missed in existing sound source localization benchmarks. Indeed, our experiments show that higher sound localization performance does not guarantee higher cross-modal retrieval performance.

*These authors contributed equally to this work.

Second, given this additional criterion, we revisit the importance of semantic understanding shared across audio and visual modalities in both sound source localization and cross-modal retrieval. In the previous methods [50, 51, 54, 47], the cross-modal semantic alignment is induced by instance-level cross-modal contrastive learning, *i.e*., cross-modal instance discrimination between visual and audio features. However, they are aided by labels or supervisedly pretrained encoder [2] for easing challenging cross-modal feature alignment. Instead, our method learns from scratch supporting the lack of guidance by incorporating multiple positive samples into cross-modal contrastive learning. Specifically, we construct a positive set for each modality using both multi-view [10] and conceptually similar samples [17]. Thereby, we enhance feature alignment and achieve high localization performance and strong cross-modal semantic understanding.

We evaluate our method on the VGG-SS and SoundNet-Flickr benchmarks for sound source localization and cross-modal retrieval. As aforementioned, the sound source localization task is closely related to the cross-modal retrieval task, but our experiments show that existing works have a weak performance correlation between them. This implies that we need to evaluate both tasks for evaluating the genuine sound source localization. The proposed method performs favorably against the recent state-of-the-art approaches in both tasks.

We summarize the contributions of our work as follows:

- We analyze that sound source localization benchmarks are not capable of evaluating cross-modal semantic understanding, thereby sound source localization methods may perform poorly in cross-modal retrieval tasks.
- We propose semantic alignment to improve cross-modal semantic understanding of sound source localization models.
- We expand semantic alignment with multi-views and conceptually similar samples which leads to state-of-the-art performance on both sound source localization and cross-modal retrieval.

2. Related work

Sound source localization. Sound source localization in visual scenes has been investigated by exploiting correspondences between audio and visual modalities. The most widely used approach for sound source localization is cross-modal attention [50, 51, 57] with contrastive loss [13, 29, 42]. Later, the attention-based method is improved by intra-frame hard sample mining [8], iterative contrastive learning with pseudo labels [35], feature regularization [36], positive mining [52], negative free learning [54] with stop-gradient operation [12], or momentum encoders [38].

[2]Typically, an image encoder is pretrained on ImageNet [16] and an audio encoder is pretrained on AudioSet [25] in supervised ways.

Some sound localization approaches exploit additional semantic labels [47, 33, 53] or object prior [39, 63]. Semantic labels are used to pretrain audio and vision encoders with classification loss [33, 53] or refine audio-visual feature alignment [47]. A more explicit way to refine localization output is to use object prior. EZVSL [39] proposes post-processing to combine attention based localization output with a pretrained visual feature activation map. Similarly, Xuan *et al*. [63] propose to combine off-the-shelf object proposals with attention based sound localization results. However, postprocessing by object prior may generate a false positive output as it is solely based on vision without audio-visual interaction.

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 [27, 37]. The separation of sound mixture is achieved by predicting masks of spectrogram guided by visual features [19, 1, 64, 23, 62, 21, 2, 65, 24, 58, 56]. Furthermore, a number of recent papers are presented on audio-visual navigation for a given sound source [7, 22].

Self-supervised representation learning. In a broader categorization, sound source localization belongs to self-supervised multimodal learning. Our work is also relevant to self-supervised audio-visual representation learning, and other multimodal learning studies.

Contrastive learning aims to learn robust representations from large-scale raw data without annotations. Recent representation learning approaches [60, 10, 28, 11] use instance discrimination by contrastive learning [13, 29, 42] as a pretext task with notable advancements in visual recognition tasks. Recently, positive mining by nearest-neighbor search are used to learn representations of images [17, 18, 61], videos [26], neural recordings [6], and text-image [34]. In this work, we expand the previous works by incorporating both multi-views and conceptually similar samples into audio-visual modalities for cross-modal feature alignment.

A series of audio-visual representation learning studies have shown that audio and visual contents in a video are correlated, therefore a visual representation can be learned by sound prediction [44] or audio representation can be distilled from visual representation [5, 55]. Later, a variety of joint audio-visual representation learning methods are proposed with an assumption that there is a semantic [3, 30, 41, 40] or temporal [14, 43, 32, 15] correspondence between them. However, simply learning sound source localization by audio-visual correspondence with instance discrimination ignores the semantic similarity of audio-visual contents among samples, introducing false negatives or positives. In order to mitigate this issue, clustering [30], sampling [41], weighting [40], and hard mining [32] are proposed. Similarly, in this work, we go beyond instance discrimination by using multiple positive samples

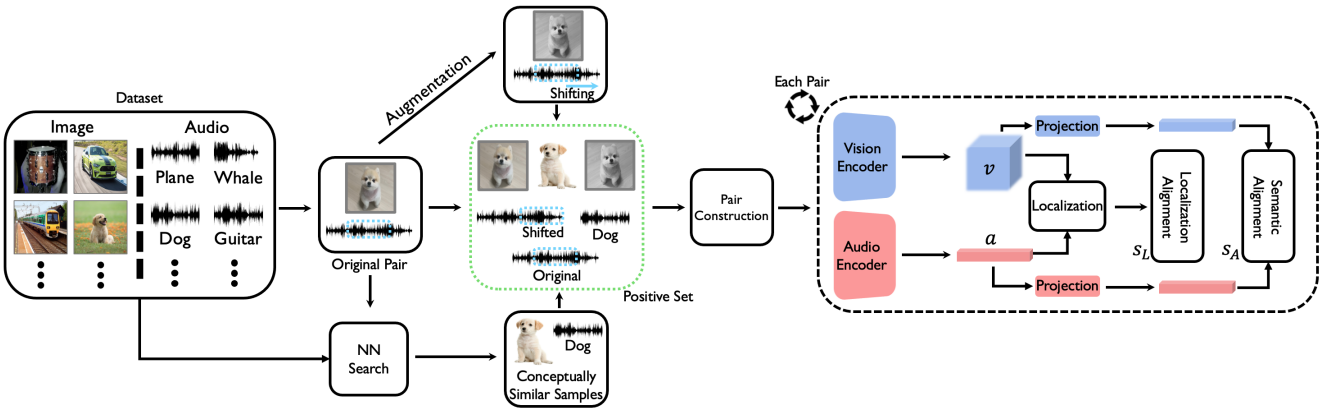

Figure 2. **Our sound source localization framework.** Our model construct multiple positive pairs with augmentation and Nearest Neighbor Search (Conceptually Similar Samples). By using these newly constructed 9 pairs, our model employs spatial localization, s_L, and semantic feature alignment, s_A, for each pair to learn a better sound source localization ability.

to enforce semantic understanding across modalities.

3. Method

3.1. Preliminaries

Contrastive learning learns representation by containing positive and negative pairs. Given an encoded query sample q and its encoded positive pair k^+ and negative pairs k, the loss can be defined as:

$$\mathcal{L} = -\log \frac{\exp(q \cdot k^+/\tau)}{\sum_i \exp(q \cdot k_i/\tau)} \tag{1}$$

where τ is the temperature parameter.

Cross-modal contrastive learning extends contrastive learning across multiple modalities. In sound source localization, audio-visual correspondence is used to define positive and negative cross-modal pairs. With an audio-visual dataset $\mathcal{D} = \{(v_i, a_i) : i = 1, ..., N\}$ and its encoded features $\mathbf{v}_i = f_v(v_i)$ and $\mathbf{a}_i = f_a(a_i)$, cross-modal contrastive learning loss is defined as:

$$\mathcal{L}_i = -\log \frac{\exp(s(\mathbf{v}_i, \mathbf{a}_i)/\tau)}{\sum_j \exp(s(\mathbf{v}_i, \mathbf{a}_j)/\tau)} \tag{2}$$

where s is a cross-modal similarity function. The cross-modal contrastive loss Eq. (2) can be extended to symmetric form [48] as used in a few previous works [39, 38].

3.2. Cross-Modal Feature Alignment

We consider both spatial localization and semantic feature alignment for sound source localization. To this end, we use two different similarity functions s_L and s_A for contrastive learning (Eq. (2)), s_L for localization and s_A for cross-modal feature alignment.

Recent studies rely on audio-visual spatial correspondence maps to learn sound source localization by contrasting them. Given a spatial visual feature $\mathbf{v} \in \mathbb{R}^{c \times h \times w}$ and audio feature $\mathbf{a} \in \mathbb{R}^{\mathbf{c}}$, audio-visual similarity with a correspondence map can be calculated as follows:

$$s_L(\mathbf{v}, \mathbf{a}) = \sum_{xy \in M} \frac{1}{|M|} \frac{\mathbf{v}^{xy} \cdot \mathbf{a}}{\|\mathbf{v}^{xy}\| \|\mathbf{a}\|} \tag{3}$$

where $\mathbf{v}^{xy}$ is a feature vector at location (x, y), and M is an optional binary mask when an annotation or pseudo-mask [8, 36] is available. Since we assume no supervision for sound source localization, we do not use any mask, therefore, $M = \mathbf{1}$.

The contrastive loss with localization similarity s_L enforces location dependent alignment giving sparse but strong audio-visual correspondence which enables to perform localization. However, our empirical studies on cross-modal retrieval indicate that strong localization performance does not guarantee semantic understanding. To overcome the low semantic understanding in recent studies, we propose to add instance-level contrastive loss. Instance-level contrasting encapsulates the whole context in a scene, enforcing better audio-visual semantic alignment. However, instance-level contrasting may smooth out spatial discriminativeness learned by Eq. (3). Inspired by SimCLR [10], we adopt a projection layer to align audio-visual semantics in a projection space. The projection layer separates the latent space of localization and semantic alignment, thereby preventing the alignment loss smoothing out the spatial discriminativeness. The similarity function for cross-modal feature alignment is defined as follows:

$$s_A(\mathbf{v}, \mathbf{a}) = \frac{p_v(\mathsf{avg}(\mathbf{v})) \cdot p_a(\mathbf{a})}{\|p_v(\mathsf{avg}(\mathbf{v}))\| \|p_a \mathbf{a}\|} \tag{4}$$

where $\mathsf{avg}(\cdot)$ is spatial average pooling, p_v is a projection

layer for visual features, and p_a is a projection layer for audio features.

3.3. Expanding with Multiple Positive Samples

Typically, contrastive learning contrasts between one positive pair and multiple negative pairs as shown in Eq. (1). In audio-visual learning, by an audio-visual correspondence assumption, an audio-image pair from the same clip is used as a positive pair while negative pairs are sampled from different clips. However, single-instance discrimination may not be sufficient to achieve strong cross-modal alignment. In this section, we expand contrastive learning beyond single instance discrimination by positive set construction and pairing them. To construct a positive set, we incorporate both hand-crafted positive and conceptual positive samples for each modality. Later, we adjust the contrastive learning to incorporate multiple positive pairs to enforce cross-modal alignment.

Obtaining hand-crafted positive samples. Using randomly augmented samples as positive multi-view pairs are widely adopted in self-supervised representation learning, *i.e.*, instance discrimination. Similarly, we extend a single anchor audio-image pair to multiple positive pairs by applying simple augmentations on image and audio samples separately. While we utilize common image transformations on images, we apply temporal shifting to audios. It is worth noting that sound source localization task learns from the underlying semantic consistency rather than subtle time differences as in videos. Thus, a slight shift in the audio may not alter contextual information significantly. As a result of hand-crafted multi-view positive pair generation, we obtain additional $\mathbf{v}^{aug}$ and $\mathbf{a}^{aug}$ samples.

Obtaining conceptual positive samples. Apart from manually created augmented views, we additionally expand our positive set with conceptually similar samples. The sampling strategy with nearest neighbor search can be performed in a various way, such as on-the-fly sampling [17, 49, 61, 34], sampling by pretrained encoders [52], or guided sampling [26, 18] using another modality. For selecting our conceptually similar samples, we utilize pretrained encoders. Note that pretrained encoders trained either with supervised or self-supervised learning are effective in positive sample mining as shown in the experiment section. By employing readily available image and audio encoders, we use the k-nearest neighborhood search to sample semantically similar samples in both modalities. In particular, given a pair of image and audio, we compute cosine similarity with all other samples and choose the top-k most similar samples among the training set for each modality. From a set of k samples, we randomly select one sample to obtain conceptually similar samples for each modality, $\mathbf{v}^{conc.}$ and $\mathbf{a}^{conc.}$. By utilizing the conceptually similar samples as positive samples, our model expands semantic understanding.

Pair Construction. Once we obtain the conceptual and hand-crafted positive samples for each modality, we proceed to create 9 distinct audio-visual pairs by pairing $\mathbf{V} = \{\mathbf{v}, \mathbf{v}^{aug}, \mathbf{v}^{conc}\}$ and $\mathbf{A} = \{\mathbf{a}, \mathbf{a}^{aug}, \mathbf{a}^{conc}\}$. This is done to ensure semantic alignment and consistency between them through contrastive learning. The negative pairs are randomly paired from the remaining samples in a training set. It is worth noting that some of these pairs are a combination of hand-crafted and conceptually similar samples, which further enhances the feature alignment of our model during training.

3.4. Training

Our loss formulation incorporates both localization and instance-level similarity functions with multiple positive pairs constructed by augmentation and conceptually similar sample search. The final loss term is defined as follows:

$$\mathcal{L}_i = -\sum_{\mathbf{v}_i \in \mathbf{V}} \sum_{\mathbf{a}_i \in \mathbf{A}} \left[\log \frac{\exp(s_L(\mathbf{v}_i, \mathbf{a}_i)/\tau)}{\sum_j \exp(s_L(\mathbf{v}_i, \mathbf{a}_j)/\tau)} + \log \frac{\exp(s_A(\mathbf{v}_i, \mathbf{a}_i)/\tau)}{\sum_j \exp(s_A(\mathbf{v}_i, \mathbf{a}_j)/\tau)} \right] \tag{5}$$

where $\mathbf{V}$ and $\mathbf{A}$ indicate positive sample sets.

4. Experiments

Our proposed method for sound source localization is validated through experiments conducted on VGGSound [9] and SoundNet-Flickr [5]. First, we conduct a quantitative analysis to evaluate the accuracy of the localization, cross-modal retrieval, and the impact of various components of our model. Then, we visualize our sound source localization results across different categories of sounds.

4.1. Experiment Setup

Datasets. Our method is trained using the VGGSound [9] and SoundNet-Flickr-144K [50, 51]. VGGSound is an audio-visual dataset containing around ∼200K videos. SoundNet-Flickr-144K set is the subset of SoundNet-Flickr [5]. After training, we test the sound localization performance with VGG-SS [8] and SoundNet-Flickr-Test [50] datasets for the main experiments. These evaluation sets have bounding box annotations of sound sources for ∼5K and 250 samples, respectively. Moreover, we employ the AVSBench [66] and Extended VGGSound/SoundNet-Flickr [38] datasets for additional evaluations. AVSBench dataset provides binary segmentation maps that show the

Method	Pre. Vision	VGG-SS cIoU ↑	VGG-SS AUC ↑	Flickr-SoundNet cIoU ↑	Flickr-SoundNet AUC ↑
Attention [50]$_{\text{CVPR18}}$	✓	18.50	30.20	66.00	55.80
CoarseToFine [47]$_{\text{ECCV20}}$	✓	29.10	34.80	-	-
LCBM [53]$_{\text{WACV22}}$	✓	32.20	36.60	-	-
LVS [8]†$_{\text{CVPR21}}$	✗	30.30	36.40	72.40	57.80
LVS [8]$_{\text{CVPR21}}$	✗	34.40	38.20	71.90	58.20
HardPos [52]$_{\text{ICASSP22}}$	✗	34.60	38.00	76.80	59.20
SSPL (w/o PCM) [54]$_{\text{CVPR22}}$	✓	27.00	34.80	73.90	60.20
SSPL (w/ PCM) [54]$_{\text{CVPR22}}$	✓	33.90	38.00	76.70	60.50
EZ-VSL (w/o OGL) [39]$_{\text{ECCV22}}$	✓	35.96	38.20	78.31	61.74
SSL-TIE [36]$_{\text{ACM MM22}}$	✗	38.63	39.65	79.50	61.20
SLAVC (w/o OGL) [38]$_{\text{NeurIPS22}}$	✓	37.79	39.40	**83.60**	-
Ours					
↳ NN Search w/ Supervised Pre. Encoders	✗	**39.94**	**40.02**	79.60	**63.44**
↳ NN Search w/ Self-Supervised Pre. Encoders	✗	39.20	39.70	79.20	63.00
with OGL:					
EZ-VSL (w/ OGL) [39]$_{\text{ECCV22}}$	✓	38.85	39.54	83.94	63.60
SLAVC (w/ OGL) [38]$_{\text{NeurIPS22}}$	✓	39.80	-	**86.00**	-
Ours (w/ OGL)					
↳ NN Search w/ Supervised Pre. Encoders	✗	**42.64**	**41.48**	82.40	**64.60**
↳ NN Search w/ Self-Supervised Pre. Encoders	✗	42.47	41.42	82.80	64.48
with Optical Flow:					
HearTheFlow [20]$_{\text{WACV23}}$	✓	39.40	40.00	84.80	64.00

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. SLAVC [38] does not provide AUC scores.

audio-visually correspondent pixels for roughly 5k five-second videos belonging to 23 categories. Lastly, the Extended VGGSound /SoundNet-Flickr dataset, proposed by [38], is used to understand non-visible sound sources.

Implementation details. We use two ResNet18 models for both audio and vision encoding. Unlike prior approaches, we do not fine-tune (or use a pretrained) a visual encoder from ImageNet pretrained weights. Instead, we train both the audio and vision encoders from scratch. We preprocess images and audios following the previous works [8, 52]. To create multiple pairs, we utilize both NN search and generic augmentation approaches. For NN search, we experiment on two different setups to retrieve k conceptually similar samples: (1) For supervisedly pretrained encoder experiments, We employ ResNet and VGGSound models pretrained on ImageNet and VGGSound respectively, (2) For self-supervisedly pretrained encoder experiments, we utilize the CLIP [48] Vision Encoder and Wav2CLIP [59] Audio Encoder. We use k=1000 for the experiments. To perform image augmentations, we follow the augmentations used in SimCLR [10]. For audios, we randomly select time-window shifts in a time axis. The model is trained for 50 epochs with Adam Optimizer and a learning rate of 0.0001. τ is set to 0.07 in contrastive learning.

4.2. Quantitative Results

Comparison with strong baselines. In this section, we conduct a comparative analysis of our sound source localization method against existing approaches. We carry out our evaluations in two settings, following previous approaches. Firstly, we train our model on VGGSound-144K and evaluate it on VGG-SS and SoundNet-Flickr test sets. Secondly, we train our model on SoundNet-Flickr-144K and evaluate it on the SoundNet-Flickr test set. It is important to note that all the compared models are trained using the same amount of data. AVEL [57], AVObject [2], and LCBM [53] models rely on video input, and as such, they cannot be evaluated on the SoundNet-Flickr dataset, which contains static image and audio pairs. We present our results in Table 1 and Table 2.

Method	Pre. Vision	cIoU ↑	AUC ↑
Attention[50]$_{\text{CVPR18}}$	✓	66.00	55.80
DMC[30]$_{\text{CVPR19}}$	✓	67.10	56.80
LVS [8]†$_{\text{CVPR21}}$	✗	67.20	56.20
LVS [8]$_{\text{CVPR21}}$	✗	69.90	57.30
HardPos [52]$_{\text{ICASSP22}}$	✗	75.20	59.70
SSPL (w/o PCM) [54]$_{\text{CVPR22}}$	✓	69.90	58.00
SSPL (w/ PCM) [54]$_{\text{CVPR22}}$	✓	75.90	61.00
EZ-VSL (w/o OGL) [39]$_{\text{ECCV22}}$	✓	71.89	58.81
SSL-TIE [36]$_{\text{ACM MM22}}$	✗	81.50	61.10
SLAVC (w/o OGL) [38]$_{\text{NeurIPS22}}$	✓	-	-
Ours			
↳ NN Search w/ Supervised Pre. Encoders	✗	**85.20**	62.20
↳ NN Search w/ Self-Supervised Pre. Encoders	✗	84.80	**62.66**
with OGL:			
EZ-VSL (w/ OGL) [39]$_{\text{ECCV22}}$	✓	83.13	63.06
SLAVC (w/ OGL) [38]$_{\text{NeurIPS22}}$	✓	-	-
Ours (w/ OGL)			
↳ NN Search w/ Supervised Pre. Encoders	✗	84.00	64.16
↳ NN Search w/ Self-Supervised Pre. Encoders	✗	**84.40**	**64.38**
with Optical Flow:			
HearTheFlow [20]$_{\text{WACV23}}$	✓	86.50	63.90

Table 2. **Quantitative results on the SoundNet-Flickr test set.** All models are trained and tested on the SoundNet-Flickr 144K dataset. † is the result of the model from the official project page. SLAVC [38] does not provide results with SoundNet-Flickr 144K.

Our proposed model achieves higher performance compared to prior approaches on both test sets. Specifically, it yields a +2.15% cIoU and +0.6% AUC improvement on VGGSS, as well as a +3.7% cIoU improvement on SoundNet-Flickr compared to the state-of-the-art methods that uses pretrained vision encoder. It is worth highlighting that unlike the majority of previous works, our proposed model does not utilize a vision encoder pretrained on ImageNet in a sound source localization backbone. This is because, as discussed in Mo *et al*. [38], using supervisedly pretrained vision encoders makes the sound source localization problem a weakly supervised problem. However, it is worth noting that even without using a pretrained vision encoder, our method achieves state-of-the-art performance on both experiments that are presented in Table 1 and Table 2. We demonstrate the performance of our model with the pretrained models learned through supervised learning (NN Search w/ Supervised Pre. Encoders) and with models that are pretrained through self-supervised learning (NN Search w/ Self-Supervised Pre. Encoders) in NN Search module. As the results indicate, using self-supervised pre-

Test Class	Method	Pre. Vision	cIoU ↑	AUC ↑
Heard 110	LVS [8]$_{\text{CVPR21}}$	✗	28.90	36.20
	EZ-VSL(w/o OGL) [39]$_{\text{ECCV22}}$	✓	31.86	36.19
	SLAVC(w/o OGL) [38]$_{\text{NeurIPS22}}$	✓	35.84	-
	Ours	✗	**38.31**	**39.05**
	with OGL:			
	EZ-VSL(w/ OGL) [39]$_{\text{ECCV22}}$	✓	37.25	38.97
	SLAVC(w/o OGL) [38]$_{\text{NeurIPS22}}$	✓	38.22	-
	Ours(w/ OGL)	✗	**41.85**	**40.93**
	with Optical Flow:			
	HearTheFlow [20]$_{\text{WACV23}}$	✓	37.30	38.60
Unheard 110	LVS [8]$_{\text{CVPR21}}$	✗	26.30	34.70
	EZ-VSL(w/o OGL) [39]$_{\text{ECCV22}}$	✓	32.66	36.72
	SLAVC(w/o OGL) [38]$_{\text{NeurIPS22}}$	✓	36.50	-
	Ours	✗	**39.11**	**39.80**
	with OGL:			
	EZ-VSL(w/ OGL) [39]$_{\text{ECCV22}}$	✓	39.57	39.60
	SLAVC(w/o OGL) [38]$_{\text{NeurIPS22}}$	✓	38.87	-
	Ours(w/ OGL)	✓	**42.94**	**41.54**
	with Optical Flow:			
	HearTheFlow [20]$_{\text{WACV23}}$	✓	39.30	40.00

Table 3. **Comparison results on open-set audio-visual localization experiments trained and tested on the splits of [8, 39, 20].**

Test Class	Method	Pre. Vision	cIoU ↑	AUC ↑
Heard 110	SSSL-TIE [36]$_{\text{ACM MM22}}$	✗	39.00	40.30
	Ours	✗	**41.20**	**41.00**
Unheard 110	SSSL-TIE [36]$_{\text{ACM MM22}}$	✗	36.50	38.60
	Ours	✗	**36.90**	38.59

Table 4. **Comparison results on open set audio-visual localization experiments trained and tested on the splits of [36].**

trained encoders in NN Search performs on par with the supervised pretrained encoders in NN Search. This shows that our model does not depend on supervised pretrained encoders for the NN search module and can utilize any type of pretrained encoder feature for nearest neighbor search. Note that these pretrained encoders are not used in the backbone networks of the sound source localization module but only in the NN Search Module, as illustrated in Figure 2.

We also discuss the methods employed by previous studies, such as SSPL [54] which utilizes a sub-module called PCM to reduce the impact of background noise, HTF [20] which utilizes Optical Flow, and EZ-VSL [39] which refines its initial audio-visual localization outcomes through object guidance obtained from an ImageNet pretrained visual encoder. Our model, on the other hand, and any of its variations do not require any task-specific modules or operations to achieve the state-of-the-art (SOTA) results. This suggests that using additional semantic and multi-view correspondence, as well as feature alignment, provides more varied and robust supervision for better aligned audio and visual features, as opposed to using task-specific approaches.

The quantitative results presented in Table 1 and Table 2 also showcase the performance of previous methods that utilize object guidance to evaluate their final sound source localizations. Our model outperforms all previous methods that employ object guidance on the VGG-SS test set and achieves comparable results on the SoundNet-Flickr test set, even though our model *does not use object guided refinement (OGL)*. Additionally, we acknowledge that the addition of OGL to our audio-visual localization results in improvement on the VGGSS test set, while degrading performance on the SoundNet-Flickr test set. In contrast, prior methods see modest improvements when utilizing OGL. This can be explained by the fact that our model is already accurately localizing the sounding objects, and object guidance can interfere with localization results by introducing visual regions that are not sounding (refer to Section 4.4 for visual results). Unlike prior methods, we do not use OGL in our architecture for the remainder of this paper, unless it is being directly compared with OGL-based methods.

Finally, in comparison to HearTheFlow, which utilizes an additional Optical Flow modality, our method outperforms it on the VGGSS test set, and achieves slightly lower performance on the SoundNet-Flickr test set without utilizing any additional modalities, but instead relying on better audio-visual correspondence and alignment.

Test Set	Method	Pre. Vision	mIoU ↑	F-Score ↑
S4	LVS (w/o OGL) [8]$_{\text{CVPR21}}$	✗	26.9	33.6
	EZ-VSL (w/o OGL) [39]$_{\text{ECCV22}}$	✓	27.6	34.2
	SLAVC (w/o OGL) [38]$_{\text{NeurIPS22}}$	✓	28.1	34.6
	Ours (w/o OGL)			
	↳ NN Search w/ Supervised Pre. Encoders	✗	**29.6**	**35.9**
	↳ NN Search w/ Self-Supervised Pre. Encoders	✗	29.3	35.6

Table 5. **Quantitative results on AVS Bench S4 dataset.** All models are trained on the VGGSound 144K dataset.

		A → I			I → A		
Model	Pre. Vision	R@1	R@5	R@10	R@1	R@5	R@10
LVS [8]$_{\text{CVPR21}}$	✗	3.87	12.35	20.73	4.90	14.29	21.37
EZ-VSL [39]$_{\text{ECCV22}}$	✓	5.01	15.73	24.81	14.2	33.51	45.18
SSL-TIE [36]$_{\text{MM22}}$	✗	10.29	30.68	43.76	12.76	29.58	39.72
SLAVC [38]$_{\text{NeurIPS22}}$	✓	4.77	13.08	19.10	6.12	21.16	32.12
Ours							
↳ NN Search w/ Supervised Pre. Encoders	✗	**16.47**	36.99	49.00	**20.09**	**42.38**	**53.66**
↳ NN Search w/ Self-Supervised Pre. Encoders	✗	14.31	**37.81**	**49.17**	18.00	38.39	49.02

Table 6. **Summary of retrieval recall scores for all models.** All of the models are trained on VGGSound 144K data and retrieval is performed on entire VGG-SS dataset, containing ∼5K samples.

Open Set Audio-Visual Localization. The study by Chen et al. [8] evaluates the generalization ability of sound source localization methods in an open set scenario. This involves testing the models on categories that are both present in the training data (heard) and categories that are not present (unheard). To accomplish this, 110 randomly selected categories from the VGGSound dataset are used for training, while another disjoint set of 110 categories are reserved for evaluation to ensure the model have never seen or heard them before. It should be noted that not all previous works use the same train/test splits. While some works, including [8, 39, 38, 46], share the same splits,[36] uses a different split. Therefore, to ensure a fair comparison, we conduct experiments on both splits, evaluating on test samples from both heard and unheard categories. The results are shown in Table 3 and Table 4. Our model outperforms existing approaches on both categories, regardless of the train/test splits used. Specifically, in Table 3, our model (w/o OGL) even surpasses the other models that use OGL. Previous approaches draw different conclusions from

Method	Pre. Vision	Extended Flickr-SoundNet AP ↑	max-F1 ↑	LocAcc ↑	Extended VGG-SS AP ↑	max-F1 ↑	LocAcc ↑
CoarseToFine [47]$_{ECCV20}$	✓	0.00	38.20	47.20	0.00	19.80	21.93
LVS [8]$_{CVPR21}$	✗	9.80	17.90	19.60	5.15	9.90	10.43
Attention10k [50]$_{CVPR18}$	✓	15.98	24.00	34.16	6.70	13.10	14.04
DMC [30]$_{CVPR19}$	✓	25.56	41.80	52.80	11.53	20.30	22.63
DSOL [31]$_{NeurIPS20}$	✓	38.32	49.40	72.91	16.84	25.60	26.87
OGL [39]$_{ECCV22}$	-	40.20	55.70	77.20	18.73	30.90	36.58
EZ-VSL (w/o OGL) [39]$_{ECCV22}$	✓	46.30	54.60	66.40	24.55	30.90	31.58
SLAVC (w/o OGL) [38]$_{NeurIPS22}$	✓	51.63	59.10	**83.60**	32.95	40.00	37.79
Ours							
↳ NN Search w/ Supervised Pre. Encoders	✗	**64.43**	**66.90**	79.60	**34.73**	**40.70**	**39.94**
↳ NN Search w/ Self-Supervised Pre. Encoders	✗	62.67	66.10	79.20	33.09	40.00	39.20

Table 7. **Quantitative results on the Extended VGG-SS and Extended SoundNet-Flickr sets**. All models are trained with 144K samples from VGG-Sound. The results of the prior approaches are obtained from [38].

	Semantic	Multi-View	Feature Alignment	cIoU ↑	AUC ↑
(A)	✓	✓	✓	**39.94**	**40.02**
(B)	✓	✓	✗	39.10	39.44
(C)	✓	✗	✓	38.75	39.34
(D)	✓	✗	✗	38.24	38.90
(E)	✗	✓	✓	38.30	39.38
(F)	✗	✓	✗	37.72	39.19
(G)	✗	✗	✓	34.93	37.94
(H)	✗	✗	✗	34.22	37.67

Table 8. **Ablation studies on our proposed method to see the impact of each main component.**

these open set experiments. While some conclude that their models have strong generalization ability because their performance in unheard categories is higher than heard categories [39, 38, 46], the other works that cannot achieve the same trend discuss that this is expected since their models are dealing with unseen categories [36]. However, our results show that these conclusions are highly dependent on the chosen train/test splits. Our model performs better than existing works in both splits, but there is no uniform trend in between two splits. While our method performs better on unheard categories in the splits of [8, 39, 38, 46], it performs worse on unheard categories in the split of [36]. Therefore, we conclude that the observed trends are highly dependent on the randomly selected train/test splits.

AVSBench [66]. To demonstrate the precise sound localization ability of our model, we conduct experiments on the AVSBench S4 dataset. The dataset's objective is to detect audio-visual correspondence and correlation at the pixel level. To make a fair comparison, we use some of the self-supervised sound source localization methods mentioned earlier. All models are trained on VGGSound-144K and directly assessed on the AVSBench S4 dataset without any further fine-tuning (zero-shot setting). Our results, which are presented in Table 5, indicate that our method achieves the highest performance, as in the previous experiments.

Retrieval. We evaluate sound localization models on the VGG-SS dataset for cross-modal retrieval. As shown in Table 6, our method clearly outperforms other state-of-the-art methods. One interesting observation is that EZ-VSL [39] notably performs better than SLAVC [38] on cross-modal retrieval, while SLAVC performs better on sound source localization in Table 1. This shows that with the current benchmark evaluations, better sound localization performance does not guarantee better audio-visual semantic understanding, thereby we need to additionally evaluate sound source localization methods on cross-modal understanding tasks. Another observation is that the performance gap between our method and the strongest competitor SSL-TIE [36] is notably larger on cross-modal retrieval than sound source localization. This is due to the strong cross-modal feature alignment of our method that is overlooked in the sound source localization benchmarks.

Extended Flickr and VGG-SS datasets. The prior study [38] points out that the current sound source localization benchmarks overlook false positive detection. It is because the evaluation samples always contain at least a sounding object in a scene; thus cannot capture false positive outputs, *e.g.*, silent objects or off-screen sounds. To analyze false positive detection, Mo and Morgado [38] extended the benchmarks with non-audible, non-visible, and mismatched audio-visual samples. The expectation is that a sound source localization model should not localize any objects when audio-visual semantics do not match.

The experiment with the extended datasets in Table 7 shows that our method performs favorably against state-of-the-art competitors. Our method performs better than the competing methods in false positive detection measured by **AP** and **max-F1**, while SLAVC [38] achieves better localization performance on Extended Flickr-SoundNet. As both false positive detection and cross-modal retrieval require cross-modal interaction, our method shows strong performance on both tasks.

4.3. Ablation Results

We conduct a series of experiments in order to verify our design choices and make further analysis. To save computational time and resources, we primarily perform ablation studies by training our model on VGGSound-144K with NN Search w/ Supervised Pre. Encoders setup and evaluating it on VGG-SS. Results are in Table 8.

Impact of Semantic and Multi-View Invariance. In order to understand the impact of each type of invariance (consistency), we analyze the performance of our model with different type of invariance methodologies in Table 8. As the results of (C *vs.* E) and (D *vs.* F) reveal, using semantically similar samples (semantic invariance) produces better performance (+0.45% and +0.5% on cIoU respectively) compared to augmented multi-view invariance. Moreover, as the results of (A *vs.* C) and (A *vs.* E) depict, the combination of these two different types of invariance complement each other and and further enhances the model's performance. Using pair combination of these two different

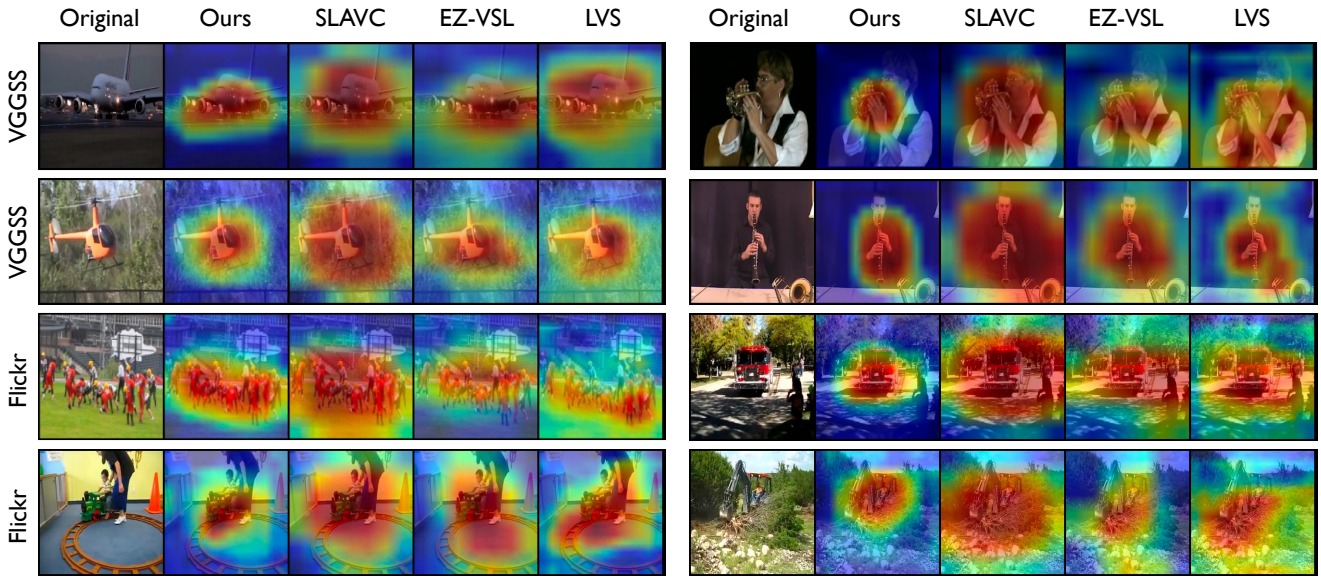

Figure 3. **Sound Localization Results on VGG-SS (top) and SoundNet-Flickr (bottom).**

k in *k*-NN	10	30	100	500	1000
cIoU ↑	38.80	38.82	39.46	39.90	**39.94**
AUC ↑	39.51	39.67	39.93	40.00	**40.02**

Table 9. **Varying k in conceptually similar sample selection.**

types of consistency elements provides additional supervisions, invariance and alignments, leading to a more robust representation space and improve sound localization performance.

Impact of Feature Alignment. We perform controlled experiments to verify the effect of the feature alignment strategy, and the results are presented in Table 8. Comparing the performance of the proposed model with and without feature alignment, (A *vs.* B), highlights the importance of this strategy to boost the performance. Further, examining the results of experiments (C *vs.* D) and (E *vs.* F) reveals that feature alignment provides additional gains irrespective of the consistency types. These findings indicate that global feature-based alignment helps the optimization of audio-visual correspondence.

Impact of k in conceptually similar sample selection. Selecting an appropriate k value for sampling nearest neighbors is crucial. If this value is set too high, it may result in noisy samples that could disrupt the learning phase. Conversely, if the value is set too low, only very similar samples to the anchor will be provided and it limits semantic invariance. Nevertheless, when compared to Table 8 (E), we observe performance gain throughout the range of k used for the ablation study. Table 9 shows an ablative evaluation of the effect of k value used to select neighborhood samples. The results indicate that an optimal choice is k=1000. This

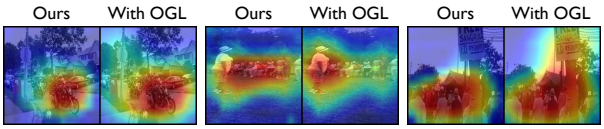

Figure 4. **OGL degrades our sound localization results on SoundNet-Flickr.**

choice of k can be explained by the fact that it provides a balance between semantic similarity and sufficient diversity.

4.4. Qualitative Results

In this section, we visualize and compare our sound localization results with the recent prior works on standard benchmarks, namely on VGG-SS and SoundNet-Flickr. The visualized samples in Figure 3 show that localized regions of the proposed method are more compact and accurately aligns with the sounding objects than the other methods. For instance, small size musical instrument is localized accurately compared to the recent methods in the top right column.

We also compare our localization results with and without object-guided localization (OGL). As shown in Figure 4, OGL deteriorates our sound localization outputs. OGL captures objectness in a scene, thereby tending to attend to any distinctive objects regardless of whether it is the sound source or not. Therefore, OGL can be helpful when localization totally fails because of the objectness bias in the benchmarks, but it is harmful when the localization is accurate which is the case for the examples shown. This result is consistent with the quantitative result in Table 2, showing that our method with OGL performs worse.

Throughout the paper, we discuss the importance of

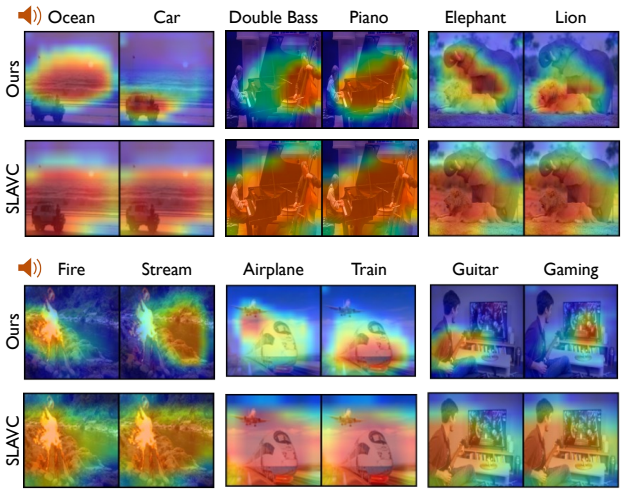

Figure 5. **Interactive Sound Localization of Ours and SLAVC [38].** Our model correctly follows the cross-modal interaction for given different sounds.

cross-modal semantic understanding. We demonstrate interactiveness of our method across modalities in Figure 5. Genuine sound source localization should be able to localize objects that are correlated with the sound. To visualize cross-modal interaction, we synthetically pair up the same image with different sounds of objects that are visible in a scene. The examples demonstrate that the proposed method can localize different objects depending on the contexts of sounds, while the competing method can not.

5. Conclusion

In this work, we investigate cross-modal semantic understanding that has been overlooked in sound source localization studies. We observe that higher sound source localization performance on the current benchmark does not necessarily show higher performance in cross-modal retrieval, despite its causal relevance in reality. To enforce strong understanding of audio-visual semantic matching while maintaining localization capability, we propose semantic alignment with multi-views of audio-visual pairs in a simple yet effective way. The ablation study shows that strong semantic alignment is achieved when both semantic alignment loss and enriched positive pairs are used. We extensively evaluate our method on sound source localization benchmarks including cross-dataset and open-set settings. Moreover, our analyses on cross-modal retrieval and false positive detection verify that the proposed method has strong capability in cross-modal interaction. Our study suggests that sound localization methods should be evaluated not only on localization benchmarks but also on cross-modal understanding tasks.

6. Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00212845, Multimodal Speech Processing for Human-Computer Interaction). H. Pfister and J. Kim were partially supported by NIH grant R01HD104969. T.-H. Oh was partially supported by IITP grant funded by the Korea government (MSIT) (No. 2021-0-02068, Artificial Intelligence Innovation Hub; No. 2022-0-00290, Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense).

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *INTERSPEECH*, pages 3244–3248, 2018. 2

[2] 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, 5

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

[4] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision*, 2018. 1

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems, NeurIPS*, 2016. 2, 4

[6] Mehdi Azabou, Mohammad Gheshlaghi Azar, Ran Liu, Chi-Heng Lin, Erik C Johnson, Kiran Bhaskaran-Nair, Max Dabagia, Bernardo Avila-Pires, Lindsey Kitchell, Keith B Hengen, et al. Mine your own view: Self-supervised learning through across-sample prediction. *arXiv preprint arXiv:2102.10106*, 2021. 2

[7] 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. In *European Conference on Computer Vision*, 2020. 2

[8] 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. 1, 2, 3, 4, 5, 6, 7

[9] 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. 4

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020. 2, 3, 5

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[13] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 2

[14] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asia Conference on Computer Vision*, pages 251–263. Springer, 2017. 2

[15] Soo-Whan Chung, Hong Goo Kang, and Joon Son Chung. Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision. In *INTERSPEECH*, 2020. 2

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2

[17] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *IEEE International Conference on Computer Vision*, 2021. 2, 4

[18] Mohamed El Banani, Karan Desai, and Justin Johnson. Learning Visual Representations via Language-Guided Sampling. In *CVPR*, 2022. 2, 4

[19] 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. 2

[20] Dennis Fedorishin, Deen Dayal Mohan, Bhavin Jawade, Srirangaraj Setlur, and Venu Govindaraju. Hear the flow: Optical flow-based self-supervised visual sound source localization. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2023. 1, 5, 6

[21] 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. 2

[22] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B. Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *IEEE International Conference on Robotics and Automation*, 2020. 2

[23] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *IEEE International Conference on Computer Vision*, 2019. 2

[24] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[25] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017. 2

[26] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020. 2, 4

[27] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005. 2

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[29] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, 2015. 2

[30] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 5, 7

[31] 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, NeurIPS*, 2020. 1, 7

[32] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[33] Sizhe Li, Yapeng Tian, and Chenliang Xu. Space-time memory network for sounding object localization in videos. In *British Machine Vision Conference*, 2021. 1, 2

[34] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. 2, 4

[35] 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. 1, 2

[36] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *ACM International Conference on Multimedia*, 2022. 1, 2, 3, 5, 6, 7

[37] Josh H McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009. 2

[38] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems, NeurIPS*, 2022. 1, 2, 3, 4, 5, 6, 7, 9

[39] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, 2022. 1, 2, 3, 5, 6, 7

[40] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[41] 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. 2

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[43] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision*, 2018. 2

[44] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 801–816. Springer, 2016. 2

[45] Takashi Oya, Shohei Iwase, Ryota Natsume, Takahiro Itazuri, Shugo Yamaguchi, and Shigeo Morishima. Do we need sound for sound source localization? In *Asia Conference on Computer Vision*, 2020. 1

[46] Sooyoung Park, Arda Senocak, and Joon Son Chung. MarginNCE: Robust sound localization with a negative margin. In *arXiv preprint arXiv:2211.01966*, 2022. 6, 7

[47] 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. 1, 2, 5, 7

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 3, 5

[49] Hyeonggon Ryu, Arda Senocak, In So Kweon, and Joon Son Chung. Hindi as a second language: Improving visually grounded speech with semantically similar samples. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2023. 4

[50] 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. 1, 2, 4, 5, 7

[51] 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. 1, 2, 4

[52] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022. 1, 2, 4, 5

[53] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2022. 1, 2, 5

[54] 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. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 5, 6

[55] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[56] 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. 2

[57] 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. 2, 5

[58] 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. 2

[59] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2CLIP: Learning robust audio representations from clip. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022. 5

[60] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[61] Haohang Xu, Xiaopeng Zhang, Hao Li, Lingxi Xie, Wenrui Dai, Hongkai Xiong, and Qi Tian. Seed the views: Hierarchical semantic alignment for contrastive representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 4

[62] Xudong Xu, Bo Dai, and Lin Dahua. Recursive visual sound separation using minus-plus net. In *IEEE International Conference on Computer Vision*, 2019. 2

[63] Hanyu Xuan, Zhiliang Wu, Jian Yang, Yan Yan, and Xavier Alameda-Pineda. A proposal-based paradigm for self-supervised sound source localization in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1038, 2022. 2

[64] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *IEEE International Conference on Computer Vision*, 2019. 2

[65] 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. 2

[66] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *European Conference on Computer Vision*, 2022. 4, 7