This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Improving Sound Source Localization with Joint Slot Attention on Image and Audio

Inho Kim[1] Youngkil Song[2] Jicheol Park[2] Won Hwa Kim[1,2] Suha Kwak[1,2]
[1]Dept. of CSE, POSTECH [2]Graduate School of AI, POSTECH
{kimih, songyk, jicheol, wonhwa, suha.kwak}@postech.ac.kr

Abstract

Sound source localization (SSL) is the task of locating the source of sound within an image. Due to the lack of localization labels, the de facto standard in SSL has been to represent an image and audio as a single embedding vector each, and use them to learn SSL via contrastive learning. To this end, previous work samples one of local image features as the image embedding and aggregates all local audio features to obtain the audio embedding, which is far from optimal due to the presence of noise and background irrelevant to the actual target in the input. We present a novel SSL method that addresses this chronic issue by joint slot attention on image and audio. To be specific, two slots competitively attend image and audio features to decompose them into target and off-target representations, and only target representations of image and audio are used for contrastive learning. Also, we introduce cross-modal attention matching to further align local features of image and audio. Our method achieved the best in almost all settings on three public benchmarks for SSL, and substantially outperformed all the prior work in cross-modal retrieval.

1. Introduction

Humans easily identify where a sound comes from within their sight. This capability is a key aspect of human perception facilitating interaction with surroundings, and is also demanded in a large variety of vision applications such as autonomous driving and robotics [10, 28, 34]. Motivated by this, a large body of research [1, 3, 6, 12, 14, 15, 21, 22, 25, 26, 29–33] has studied sound source localization (SSL), the task of locating the source of a sound within the paired image in the form of segmentation mask.

Due to the prohibitively expensive cost of sound source annotation for image-audio pair data, existing methods [6, 21, 31, 33] have focused on learning SSL using only image-audio pairs without sound source labels. To this end, they consider the task as multiple instance learning [2, 13, 20]

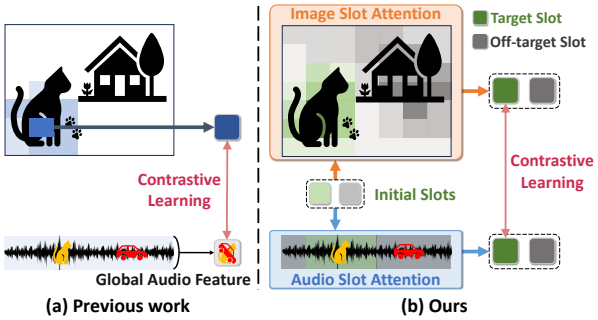

Figure 1. **Difference between previous work and ours.** For contrastive learning of SSL, our method identifies and utilizes pairs of relevant image region and audio segment, while previous work exploits a local image feature and the global audio feature.

and assume that, given an image-audio pair, at least one patch of the image is relevant to the audio; then they represent the image with its local feature that most closely resembles the global audio feature, and learn SSL through contrastive learning [24, 27] with these features (Fig. 1(a)).

While this approach has been proven to work, there is plenty of room for further improvement. First, audio may contain not only the sound of the target shared by the two modalities but also irrelevant sounds, such as noise or sounds from sources outside the image. In such cases, using the global audio feature for contrastive learning hinders accurate alignment between the image and audio features. Second, representing an image with only a single local feature in contrastive learning causes models to locate a small discriminative part of the sound source, not its whole body [35], which can degrade SSL accuracy.

To address these limitations, we propose a novel SSL method based on slot attention [17], which is conceptually illustrated in Fig. 1(b). In slot attention, slots, which are randomly sampled or learnable embeddings, compete with each other for aggregating input features progressively, and thus they decompose the features into distinct and exclusive representations. Motivated by this, we employ slot attention with two learnable slots to jointly decompose image

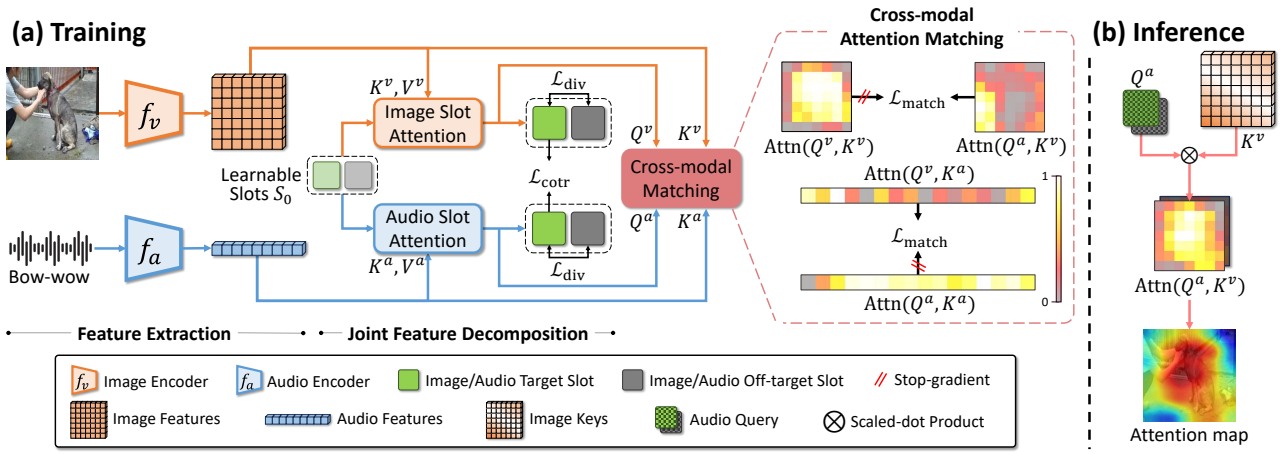

Figure 2. **Overview of the proposed method.** First, image and audio features are extracted by the image and audio encoders, and serve as keys and values for slot attention. In slot attention, two learnable slots are used as queries to draw attentions on the input features while competing with each other, decomposing the features into the target and off-target representations per modality. The model is then trained by contrastive learning using only the target representations, and by cross-modal attention matching for further feature-level alignment between the two modalities. For localization in testing, we draw the attention map between the audio target representation and image features. Some details for learning slot attention, such as decoders and loss for slot reconstruction, have been omitted for clarity.

and audio features into two distinct slot representations per modality, one for the target and the other for the rest. To be specific, we designate the first slot as the *target slot*, which is guided to focus on information related to the shared sound source across both modalities. The exclusive assignment property of slot attention directs the second slot, referred to as the *off-target slot*, to capture remaining elements, such as the image background or irrelevant sounds in the audio. Then we learn SSL by contrastive learning using the two target slot representations from the image and audio, respectively. Unlike the prior work using only a single global audio feature, our method allows to exploit only the relevant part (*i.e.*, the target slot) of the audio for more accurate SSL. Furthermore, since the target slot of the image covers the sound source area rather than a single location, our method avoids the problem of localization being reduced to a small discriminative part.

Given the target slot representations, SSL is conducted by drawing an attention map on the image feature map using the target slot representation of the audio as query (Fig. 2(b)). For accurate pixel-level localization, the target slot of the audio should be well aligned with the keys derived from relevant individual image features; in this regard, contrastive learning using only the target slots is insufficient as it does not relate the audio target slot with individual image features. We thus introduce an additional training strategy dubbed *cross-modal attention matching*. This strategy encourages the cross-modal attention map between the audio target slot and the input image features to approximate the intra-modal attention map between the image target slot and the input image features that are inherently well aligned. In other words, it provides pseudo-supervision for SSL prediction (*i.e.*, the cross-modal attention map) using the intra-modal attention map, which is drawn by the query and keys from the same modality and thus better localizes the target. To further facilitate its cross-modal alignment, we apply this strategy to the audio modality as well.

Our model was trained on two datasets, Flickr-SoundNet [4] and VGG-Sound [5], and evaluated on three benchmarks, Flickr-SoundNet-Test [4], VGG-SS [5], and AVSBench [36]. It achieved the state of the art in almost all settings across all the three evaluation benchmarks. Further, our model outperformed all the prior work by large margin in cross-modal retrieval, suggesting that it is not a simple saliency detector but properly learns to relate image and audio for SSL [31]. In summary, the contribution of our method is three-fold:

- We propose a novel SSL model that identifies and extracts features relevant to sound source jointly from image and audio modalities through slot attention.
- We also propose cross-modal attention matching, which enables effective cross-modal alignment for individual input features and thus improves localization performance.
- Our model achieved the state of the art in almost all settings across all the three SSL benchmarks and improved cross-modal retrieval performance substantially.

2. Related Work

Sound source localization aims to identify specific regions in an image that correspond to the sources of sounds from an audio input. Since constructing audio-visual datasets is

costly, sound source localization studies have evolved into self-supervised learning approaches that use only image-audio pairs for contrastive learning [1, 6, 12, 14, 16, 21, 22, 26, 29–31, 33].

An early study, [29], addresses the task through contrastive learning between the audio-aware image features and the global audio feature. The audio-aware image features are obtained by calculating a correspondence map through the normalized inner product between the global audio feature and the image feature map. Chen *et al.* [6] propose method that uses a differentiable threshold on the correspondence map to divide audio-aware and other image features, treating the latter as hard negatives in contrastive learning to improve localization.

Recent studies focus on fine-grained sound source localization through contrastive learning between the global audio feature and the most relevant feature selected from the image feature map [21, 22, 31, 33]. Mo and Morgado [21] propose a method that refines the audio-visual correspondence map through a post-processing technique leveraging a pretrained visual model. Sun *et al.* [33] introduce a regularizer that applies mean absolute error regularization to inter-modal and intra-modality similarity matrices to increase awareness of false negatives. Senocak *et al.* [31] propose a cross-modal alignment method that preserves localization performance and enhances cross-modal retrieval by leveraging multi-view augmentation and nearest neighbor search to construct multiple positive pairs for contrastive learning. These methods employ contrastive learning with the most relevant image feature to assist in coarsely identifying sound sources but often fall short of capturing a comprehensive representation, thereby hindering localization ability. Furthermore, relying solely on global audio features can mislead contrastive learning when noise exists, potentially highlighting incorrect image features.

Following a complementary direction, several studies employ external prior knowledge or even utilize off-the-shelf segmentation modules to achieve high-quality localization results [25, 26]. Also, several studies employ sound source localization with additional information such as optical flow [9], text inputs [19] or ground truth category [23]. By contrast to these studies, our method can effectively perform sound source localization without relying on external prior knowledge or off-the-shelf models.

To address the above issues, we propose a sound source localization framework based on the slot attention mechanism. Slot attention [17] is proposed for unsupervised object-centric learning, effectively decomposing input features into distinct and mutually exclusive components via iterative attention and competition among slots. Inspired by this, we adopt slot attention in both image and audio features to decompose them into features related to the sound source and those irrelevant to it without supervision.

3. Proposed Method

As shown in Fig. 2, our model consists of two parts: 1) image and audio encoders for feature extraction, and 2) joint feature decomposition module with joint slot attention mechanism. A pair of image and audio is first fed into each encoder to extract image and audio features (Sec. 3.1). The features of each modality are decomposed into two distinct slot representations by using joint slot attention: one slot for the target and the other for the rest. (Sec. 3.2). The slot representations are then used for training as input to contrastive learning and cross-modal attention matching (Sec. 3.3), while effectively handling false negatives (Sec. 3.4). Lastly, an attention on the image feature map using the target slot representation of the audio is computed to locate the source of sound (Sec. 3.5).

3.1. Feature Extraction

Let $D = \{(v_i, a_i)\}_{i=1}^{M}$ denote a training set of image-audio pairs, where $v_i \in \mathbb{R}^{3\times H_v \times W_v}$ is an image of size $H_v \times W_v$, $a_i \in \mathbb{R}^{1\times F_a \times T_a}$ is a corresponding audio spectrogram of frequency resolution F_a and temporal length T_a, and M denotes the total number of such pairs in the dataset. For feature extraction, separate image and audio encoders f_v and f_a, are adopted. The image encoder extracts the image features $\mathbf{v}_i := f_v(v_i) \in \mathbb{R}^{h\times w\times c}$ for v_i, and the audio encoder extracts the audio features $\mathbf{a}_i := f_a(a_i) \in \mathbb{R}^{t\times c}$ for a_i, where h and w are the height and width of the image feature map respectively, t is the length of the audio features, and c is the image and audio feature dimension.

3.2. Joint Feature Decomposition

Slot attention [17] is an attention mechanism where each slot progressively competes to aggregate input features, leading to the decomposition of input features into distinct and exclusive representations. Building on this idea, for joint feature decomposition, we propose *joint slot attention (JSA)*, which jointly decomposes image and audio features into two distinct slot representations for each modality. To achieve this, we employ slot attention with two initial slots $S_0 \in \mathbb{R}^{2\times D}$ shared in the two modalities. Furthermore, we designate the first slot as *target slot*, which captures features relevant to sound source, and the other as *off-target slot*, which captures features for the rest. Since the overall pipeline of JSA is identical in the two modalities, we will describe the module in terms of image only for brevity.

JSA first projects an input image feature $\mathbf{v}$ to derive its corresponding key K^v and value V^v by applying the following operations:

$$\begin{aligned} K^v &= W^k(\text{LayerNorm}(\mathbf{v})) \in \mathbb{R}^{hw\times d}, \\ V^v &= W^v(\text{LayerNorm}(\mathbf{v})) \in \mathbb{R}^{hw\times d}, \end{aligned} \tag{1}$$

where W^k, W^v are the linear projection layers to obtain

embeddings in d-dimension. For each iteration $n \in [1, N]$, the query Q_n^v is derived by applying linear projection to the slots with the initial slots $S_0^v := S_0$, which is given by

$$Q_n^v = W^q(\text{LayerNorm}(S_{n-1}^v)) \in \mathbb{R}^{2\times d}, \tag{2}$$

where W^q is a linear projection layer shared for all iterations. The slot attention matrix A is computed by taking the dot product of the given query Q_n^v and key K^v, scaled by the square root of the dimension d as follows:

$$A_{i,j} = \frac{\exp(M_{i,j})}{\sum_{l=1}^{2}\exp(M_{i,l})}, \text{ where } M = \frac{K^v(Q_n^v)^\top}{\sqrt{d}} \in \mathbb{R}^{hw\times 2}, \tag{3}$$

where i is key index, and j is slot index. Subsequently, a weighted mean is utilized to the slot attention matrix A:

$$\hat{A}_{i,j} = \frac{A_{i,j}}{\sum_{l=1}^{hw} A_{l,j}}. \tag{4}$$

Then, the updated slot $\bar{S_n}$ is computed by Gated Recurrent Unit (GRU) [7] with S_{n-1} as hidden state, and $\hat{A}^\top \cdot V^v$ as input as follows:

$$\bar{S_n^v} = \text{GRU}(S_{n-1}^v,\ \hat{A}^\top \cdot V^v). \tag{5}$$

Finally, the slot is updated by a residual connection:

$$S_n^v = \bar{S_n^v} + \text{MLP}(\text{LayerNorm}(\bar{S_n^v})) \in \mathbb{R}^{2\times d}. \tag{6}$$

The overall process encourages the input feature to bind to specific slots, ensuring that each slot forms a mutually exclusive representation. The final outputs of JSA are image slots $\{\mathbf{p}^v; \mathbf{r}^v\} := S_N^v \in \mathbb{R}^{2\times d}$ and image query $Q_N^v \in \mathbb{R}^{2\times d}$, where $\mathbf{p}^v$ is a target slot and $\mathbf{r}^v$ is an off-target slot. The same process is applied to the audio modality using audio feature $\mathbf{a}$, yielding audio slots $\{\mathbf{p}^a; \mathbf{r}^a\} := S_N^a \in \mathbb{R}^{2\times d}$ and audio query $Q_N^a \in \mathbb{R}^{2\times d}$. Since this process begins from the shared initial slots, the initial slots are able to effectively learn joint features between two modalities. For brevity, we omit the subscript N for the final query, $Q^v := Q_N^v$ and $Q^a := Q_N^a$.

3.3. Learning Objectives

Contrastive Learning with Target Slots. We adopt the InfoNCE loss [24, 27] for contrastive learning using the target slots, guiding the target slots to identify features relevant to the sound source jointly from image and audio. Let s denote a similarity function, $s(x, y) = \exp(\cos(x, y)/\tau)$, where τ is a temperature hyperparameter. The contrastive learning loss $\mathcal{L}_{\text{cotr}}$ is given by

$$\mathcal{L}_{\text{cotr}} = -\frac{1}{B}\sum_{i=1}^{B}\left\{\log\frac{s(\mathbf{p}_i^v, \mathbf{p}_i^a)}{\sum_j^B s(\mathbf{p}_i^v, \mathbf{p}_j^a)} + \log\frac{s(\mathbf{p}_i^v, \mathbf{p}_i^a)}{\sum_j^B s(\mathbf{p}_j^v, \mathbf{p}_i^a)}\right\}, \tag{7}$$

where B denotes the number of data in a mini-batch, and the negative pairs refer to the combinations of visual and audio features that do not belong to the same original pair within the mini-batch. Since the image target slot captures the whole features of the target region, it avoids focusing solely on a specific discriminative part of the image. Furthermore, the audio target slot captures only the relevant sound from the source, it prevents the issue of learning sounds that are unrelated to the source.

Cross-modal Attention Matching. As will be described in Sec. 3.5, the localization result is derived by utilizing cross-modal attention between the audio target slot and image features. For accurate sound source localization, the audio target slot should be properly aligned with keys computed from the relevant image features. Although contrastive learning encourages alignment of target slot representations between different modalities, it does not ensure alignment between the audio target slot and image features, which degrades the localization accuracy. We resolve this issue by *cross-modal attention matching*, which encourages the cross-modal attention map to resemble the intra-modal attention map for improving feature-level alignment across the two modalities. To this end, we first calculate the cross-modal and intra-modal attention matrices. Specifically, we compute the cross-modal attention matrix between the audio and visual modalities in the same manner as Eq. (3) and Eq. (4), with the audio queries Q^a and visual keys K^v. We denote this cross-modal attention matrix as $\hat{A}^{a,v} \in \mathbb{R}^{hw\times 2}$, where hw is the length of K^v. Subsequently, to obtain the *cross-modal attention* that represents the attention weights between the audio target query and visual keys, we utilize only the first column of $\hat{A}^{a,v}$. This cross-modal attention is defined as follows:

$$\mathbf{ca}^{a,v} := [\hat{A}_{i,1}^{a,v}]_{i=1}^{hw} \in \mathbb{R}^{hw}.$$

Similarly, for intra-modal attention matrix within the visual modality, we compute the intra-attention matrix $\hat{A}^{v,v}$ using the Q^v and K^v. Then, the *intra-modal attention* that represents the attention weights between the image target query and visual keys is defined as follows:

$$\mathbf{ia}^{v,v} := [\hat{A}_{i,1}^{v,v}]_{i=1}^{hw} \in \mathbb{R}^{hw}.$$

To further facilitate cross-modal alignment, we apply this strategy symmetrically to the audio modality as well. Finally we formulate the cross-modal attention matching loss as follows:

$$\mathcal{L}_{\text{match}} = \|\mathbf{ca}^{a,v} - \text{sg}(\mathbf{ia}^{v,v})\|_2^2 + \|\mathbf{ca}^{v,a} - \text{sg}(\mathbf{ia}^{a,a})\|_2^2, \tag{8}$$

where sg$(\cdot)$ denotes the stop-gradient operation. $\mathcal{L}_{\text{match}}$ minimizes the discrepancy between cross-modal and intra-modal attentions, promoting robust feature-level alignment between the two modalities for accurate localization.

Slot Divergence. In slot attention, the diversity between the target slot and off-target slot should be large enough to ensure that each slot effectively captures distinct, mutually exclusive regions. Therefore, we introduce a slot divergence loss that encourages greater diversity among the slots. The slot divergence loss $\mathcal{L}_{\text{div}}$ is given by

$$\mathcal{L}_{\text{div}} = \max\{0, \cos(\mathbf{p}^v, \mathbf{r}^v)\} + \max\{0, \cos(\mathbf{p}^a, \mathbf{r}^a)\}. \quad (9)$$

It ensures that the different slot embeddings to be sufficiently dissimilar, enabling the target slot to more effectively localize the sound source.

Slot Reconstruction. The successful feature reconstruction in slot attention implies that each slot successfully captures unique and independent parts of the input data, which means that the slot attention mechanism effectively decomposed the input into individual slot representations. It indicates that each slot accurately represents a specific object or region in the input. To encourage this condition, we apply the slot reconstruction loss $\mathcal{L}_{\text{recon}}$ as follows:

$$\mathcal{L}_{\text{recon}} = \|\mathbf{v} - g_v(\{\mathbf{p}^v; \mathbf{r}^v\})\|_2^2 + \|\mathbf{a} - g_a(\{\mathbf{p}^a; \mathbf{r}^a\})\|_2^2, \quad (10)$$

where g_v and g_a are the decoders for the image and audio features, respectively. These decoders are composed of a small number of MLP layers with ReLU activations. Note that the decoders are auxiliary modules used only in training and not required for inference.

The overall loss is a combination of $\mathcal{L}_{\text{cotr}}$, $\mathcal{L}_{\text{match}}$, $\mathcal{L}_{\text{div}}$, and $\mathcal{L}_{\text{recon}}$, which is given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cotr}} + \lambda_1 \mathcal{L}_{\text{match}} + \lambda_2 \mathcal{L}_{\text{div}} + \lambda_3 \mathcal{L}_{\text{recon}}, \quad (11)$$

where λ_1, λ_2, and λ_3 are hyperparameters weighting the loss terms.

3.4. Mitigating False Negatives

In general, unpaired images and audios are all treated as negatives for contrastive training. However, some of them are false negatives that are semantically well matched, so treating these as negatives causes adverse effect [33]. Therefore, we propose a false negative mitigation strategy using k-reciprocal nearest neighbors of target slots.

For the image target slot $\mathbf{p}_i^v$, which is derived from v_i, we define its set of k-reciprocal nearest neighbors $R_k^v(i)$ as

$$R_k^v(i) = \{(v_j, a_j) \mid (j \in N_k^v(i)) \cap (i \in N_k^v(j))\}, \quad (12)$$

where $N_k^v(i)$ is the k-nearest neighbors of $\mathbf{p}_i^v$, computed using cosine similarity. Likewise, the set of k-reciprocal nearest neighbors between audio target slots , i.e., $R_k^a(i)$, can be derived. The predicted false negative set $R_k(i)$ is defined as

$$R_k(i) = R_k^v(i) \cap R_k^a(i). \quad (13)$$

Since the samples in $R_k(i)$ have both similar image and audio target slots with (v_i, a_i), they are likely to belong to false negatives. We exclude these predicted false negatives from the contrastive learning in Eq. (7).

3.5. Inference

For inference, $\mathbf{ca}^{a,v}$, which represents the attention between the audio target slot and the input image features, is utilized. Specifically, the attention is interpolated to the original image size, and pixels whose score exceeds the hyperparameter θ are selected to generate the localization result.

Furthermore, unlike the object-guided localization [21], we propose a novel localization refinement strategy, dubbed *Image-Query based Refinement (IQR)*, to refine the localization result without relying on external prior knowledge. IQR estimates the regions that are likely to contain the target by using the attention between the image target slot and the input image features. It can be used to identify prior regions in the image that favor the target, regardless of whether these regions are actually the source of the sound. The refined result derived by IQR, denoted as R_{IQR}, is given as

$$R_{\text{IQR}} = \alpha \cdot \mathbf{ca}^{a,v} + (1 - \alpha) \cdot \mathbf{ia}^{v,v}, \quad (14)$$

where $\alpha \in (0, 1)$ balances the sound source localization result and potential sound source regions.

4. Experiment

4.1. Experiment Setup

Datasets. Flickr-SoundNet [4] and VGG-Sound [5] are used to train. Flickr-SoundNet contains about 2 million unconstrained videos from Flickr, while VGG-Sound has around 200k video clips. For training, we use subsets of 10k and 144k from each dataset, respectively. For the evaluation, Flickr-SoundNet-Test [4], VGG-SS [5], and AVS-Bench [36] are used following the previous research [6, 21, 31, 33]. Flickr-SoundNet-Test has 250 audio-visual pairs with manually labeled bounding boxes. VGG-SS has about 5k audio-visual pairs over 220 categories. Since VGG-SS provides its category ground truth, it is also used to evaluate cross-modal retrieval performance. AVSBench has about 5k videos with pixel-wise binary segmentation maps. Following the previous study [6], the middle frame of the video clip is used for the image. For the audio, the middle 5 seconds are extracted and converted into log spectrograms with 257 frequency bins at a sampling rate of 16kHz.

Network Architecture. ResNet-18 [11] is adopted for both the image encoder and the audio encoders. The image encoder is fine-tuned from an ImageNet [8] pretrained model, while the audio encoder is trained from scratch. Both encoders include an additional 1×1 convolutional layer for projection, and the audio encoder also includes a MaxPooling layer for pooling along the frequency axis.

Hyperparameter Settings. We use the AdamW optimizer [18] with a learning rate of 5e-5 and a weight decay of 1e-2. The mini-batch size during training is set to 256, and the size of image and audio feature size h, w, t and c

Trainset	Method	OGL	VGG-SS		Flickr-SoudNet	
			cIoU	AUC	cIoU	AUC
Flickr SoundNet 10k	Attention10k [29]		-	-	43.60	44.90
	CoursetoFine [26]		-	-	52.20	49.60
	AVObject [1]		-	-	54.60	50.40
	LVS [6]		-	-	58.20	52.50
	EZ-VSL [21]		19.86	30.96	62.24	54.74
	FNAC [33]		**35.27**	38.00	84.33	63.26
	Ours		35.11	**38.66**	**85.20**	**65.26**
	EZ-VSL [21]	✓	37.61	39.21	81.93	62.58
	FNAC [33]	✓	**40.97**	40.38	84.73	64.34
	Ours	✓	39.16	**40.51**	**87.60**	**66.50**
Flickr SoundNet 144k	Attention10k [29]		-	-	66.00	55.80
	DMC [12]		-	-	67.10	56.80
	LVS [6]		-	-	69.90	57.30
	HardPos [30]		-	-	75.20	59.70
	EZ-VSL [21]		30.27	35.92	72.69	58.70
	FNAC [33]		33.93	37.29	78.71	59.33
	Alignment [31]		-	-	85.20	62.20
	Ours		**37.01**	**39.42**	**86.00**	**65.16**
	EZ-VSL [21]	✓	41.01	40.23	83.13	63.06
	FNAC [33]	✓	**41.10**	40.44	83.93	63.06
	Alignment [31]	✓	-	-	84.00	64.16
	Ours	✓	40.15	**40.75**	**89.20**	**64.50**

Trainset	Method	OGL	VGG-SS		Flickr-SoudNet	
			cIoU	AUC	cIoU	AUC
VGGSound 10k	LVS [6]		-	-	61.80	53.60
	EZ-VSL [21]		25.84	33.68	63.85	54.44
	FNAC [33]		37.29	38.99	85.74	63.66
	Ours		**37.84**	**39.82**	**86.00**	**64.90**
	EZ-VSL [21]	✓	38.71	39.80	78.71	61.53
	FNAC [33]	✓	40.69	40.42	82.13	63.64
	Ours	✓	**41.02**	**41.29**	**87.60**	**65.42**
VGGSound 144k	Attention10k [29]		18.50	30.20	-	-
	DMC [12]		29.10	34.80	-	-
	AVObject [1]		29.70	35.70	-	-
	LVS [6]		34.40	38.20	73.50	59.00
	HardPos [30]		34.60	38.00	76.80	59.20
	EZ-VSL [21]		34.38	37.70	79.51	61.17
	SLAVC [22]		37.22	-	83.20	-
	FNAC [33]		39.50	39.66	**84.73**	63.76
	Alignment [31]		39.94	40.02	79.60	63.44
	NoPrior [14]		**41.4**	41.2	-	-
	Ours		40.62	**41.66**	83.60	**64.68**
	EZ-VSL [21]	✓	38.85	39.54	83.94	63.60
	SLAVC [22]	✓	39.67	-	**86.40**	-
	FNAC [33]	✓	41.85	40.80	85.14	64.30
	Alignment [31]	✓	**42.64**	41.48	82.40	64.60
	Ours	✓	42.56	**42.46**	84.80	**64.82**

Table 1. **Quantitative results of the model trained on Flickr-SoundNet-10k, 144k, and VGGSound-10k, 144k datasets.** Note that the results of previous research are obtained from [14, 31, 33]. The results of our method represent the performance after applying Image-Query based Refinement (IQR). The result with OGL denotes the performance after applying object-guided localization [21].

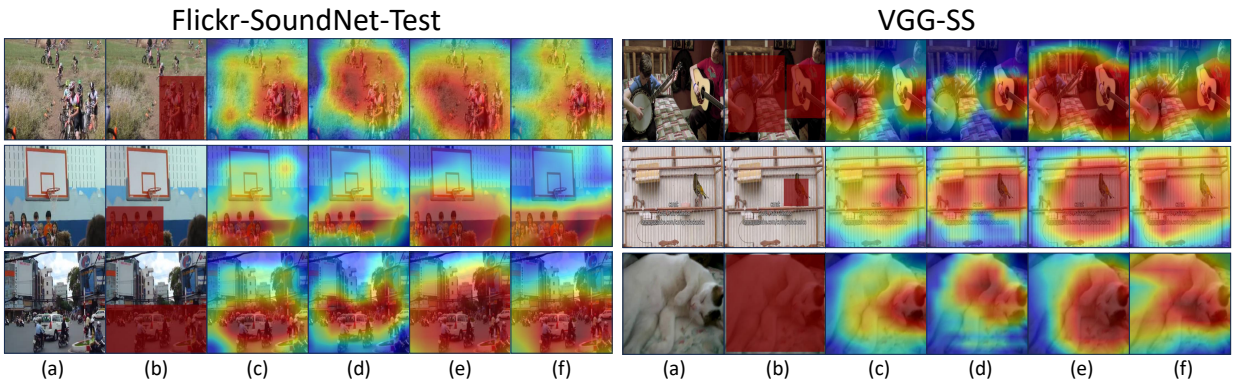

Figure 3. **Sound localization results on Flickr-SoundNet-Test [4] and VGG-SS [5].** (a) Input image. (b) Ground-Truth. (c) Ours. (d) Alignment [31]. (e) FNAC [33]. (f) EZ-VSL [21]. The qualitative results are obtained by the model trained on Flickr-144k and the model trained on VGGSound-144k, respectively. Note that all visualizations are obtained without refinement.

are set to 7, 7, 16, and 512 respectively. τ in Eq. (7) is set to 0.03. λ_1, λ_2, and λ_3 in Eq. (11) are set to 100.0, 0.1, and 0.1, respectively. α in Eq. (14) is set to 0.6. The value k in Eq. (13) is set to 20. The number of iterations N for slot attention mechanism is set to 5.

Evaluation metric. We measure the average precision at a Consensus Intersection over Union (cIoU) threshold of 0.5 to evaluate localization accuracy. Additionally, the AUC score is used to provide an overall performance measure across different thresholds.

4.2. Quantitative Results

Flickr-SoundNet and VGGSound. In this section, we compare our method with previous self-supervised SSL approaches. We trained our model on subsets of Flickr-SoundNet and VGGSound with dataset sizes of 10k and 144k, and evaluated the performance on Flickr-SoundNet-Test and VGG-SS. As shown in Table 1, our model achieves the state-of-the-art performance in most settings. Specifically, our method yields AUC improvements across all evaluated settings. Moreover, even without using Object-Guided Localization (OGL), our model performs competitively or better compared to previous methods, demonstrating the robustness of our architecture in focusing on the sound source without additional refinement steps.

Heard 110 and Unheard 110. To assess the generalization ability of SSL, we evaluate the performance with Heard 110 and Unheard 110. Heard 110 and Unheard 110 are sub-

Method	OGL	Heard 110		Unheard 110	
		cIoU	AUC	cIoU	AUC
LVS [6]		28.90	36.20	26.30	34.70
EZ-VSL [21]		31.86	36.19	32.66	36.72
SLAVC [22]		35.84	-	36.50	-
Alignment [31]		38.31	39.05	39.11	39.80
Ours		**38.46**	**40.15**	**40.35**	**41.25**
EZ-VSL [21]	✓	37.25	38.97	39.57	39.60
SLAVC [22]	✓	38.22	-	38.87	-
FNAC [33]	✓	39.54	39.83	42.91	41.17
Alignment [31]	✓	**41.85**	40.93	42.94	41.54
Ours	✓	41.50	**41.60**	**43.11**	**42.04**

Table 2. **Quantitative results on Heard 110 and Unheard 110.** Note that the results of previous research are obtained from [31, 33]. The results of our method represent the performance after applying Image-Query based Refinement (IQR). The result with OGL denotes the performance after applying object-guided localization.

Testset	Method	mIoU	F-score
S4	LVS [6]	26.9	33.6
	EZ-VSL [21]	27.6	34.2
	SLAVC [22]	28.1	34.6
	FNAC [33]	27.15	31.4
	Alignment [31]	29.6	35.9
	Alignment†	29.3	35.6
	Ours	**31.33**	**44.30**

Table 3. **Quantitative results on AVSBench.** Note that the results of previous research are obtained from [31, 33]. † denotes that the self-supervised pretrained encoder is utilized for their work.

Method	Audio → Image			Image → Audio		
	R@1	R@5	R@10	R@1	R@5	R@10
LVS [6]	3.87	12.35	20.73	4.90	14.29	21.37
EZ-VSL [21]	5.01	15.73	24.81	14.20	33.51	45.18
SSL-TIE [16]	10.29	30.68	43.76	12.76	29.58	39.72
SLAVC [22]	4.77	13.08	19.10	6.12	21.16	32.12
Alignment [31]	16.47	36.99	49.00	20.09	42.38	53.66
Alignment†	14.31	37.81	49.17	18.00	38.39	49.02
Ours	**30.61**	**57.87**	**69.37**	**31.70**	**58.43**	**69.81**

Table 4. **Cross-modal retrieval on the VGG-SS.** All of the models are trained on VGGSound-144K. Note that the results of previous research are obtained from [31]. † denotes that the self-supervised pretrained encoder is utilized for their work.

sets of the VGG-SS, each containing 110 non-overlapping categories. During training, the model is trained only to the category set of one group. The subset comprising categories seen during training is called Heard 110, while the subset with unseen categories is Unheard 110. As shown in Table 2, our model achieves the state-of-the-art performance on Unheard 110 both without additional refinement

Trainset	$\mathcal{L}_{\text{match}}$	$\mathcal{L}_{\text{div}}$	FN Mitigate	VGG-SS	
				cIoU	AUC
VGGSound 144k				14.95	27.02
	✓			36.23	39.13
		✓		18.69	30.03
			✓	15.03	28.18
	✓	✓		40.56	41.35
	✓		✓	36.14	39.31
		✓	✓	20.37	30.81
	✓	✓	✓	**40.71**	**41.62**

Table 5. **Detailed ablation study for losses.**

Trainset	Shared Initial Slots	VGG-SS		Flickr	
		cIoU	AUC	cIoU	AUC
Flickr-144k		34.80	38.72	84.00	61.70
	✓	**36.64**	**39.24**	**86.80**	**65.44**
VGGSound-144k		33.11	37.92	81.20	64.36
	✓	**40.71**	**41.62**	**84.40**	**64.50**

Table 6. **Ablation study on shared initial slots.**

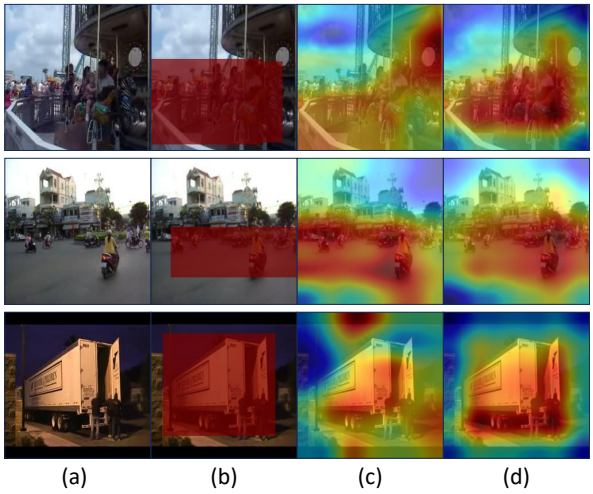

Figure 4. **Qualitative results to show the impact of $\mathcal{L}_{\text{match}}$ on Flickr-SoundNet.** (a) Input image. (b) Ground-Truth. (c) Results without cross-modal attention matching. (d) Results with cross-modal attention matching.

and with OGL applied. These results demonstrate that our model has superior generalization ability compared to previous models.

AVSBench. To more precisely evaluate the performance of the localization mask, we measured the performance on AVSBench, which provides segmentation masks. Our model was trained on VGGSound-144k and evaluated in a zero-shot manner on AVSBench. As shown in Table 3, we achieved the state-of-the-art results, specifically with improvements of 2.03%p in mIoU and 8.7%p in F-score.

Cross-modal Retrieval. As discussed in the previous study [31], the benchmark evaluation metrics, cIoU and

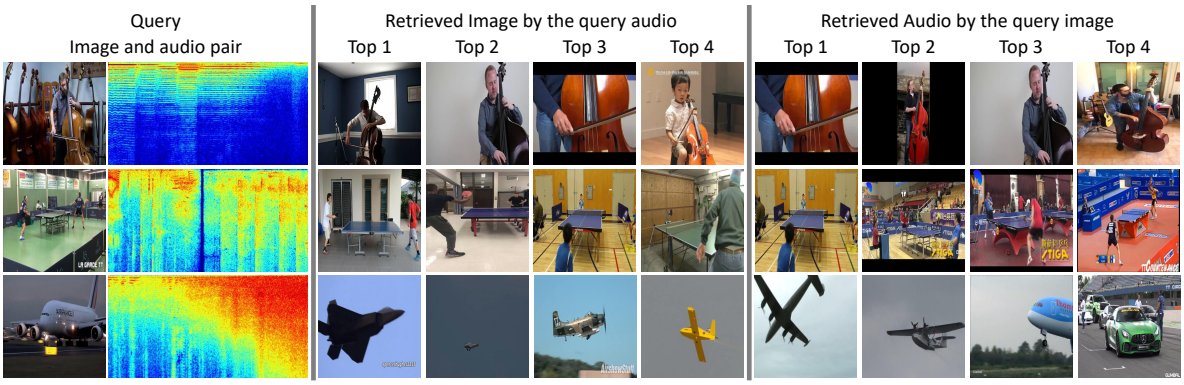

Figure 5. **Qualitative results of cross-modal retrieval on VGG-SS.**

AUC, do not guarantee better audio-visual semantic understanding. Therefore, we also evaluate cross-modal retrieval performance on VGG-SS. We extract the image and audio target slot representations and compute the cosine similarity between them. The similarities are ranked in descending order, and we calculate Recall@K performance using the category labels from the top-K retrieved audio or images. In Table 4, our model demonstrates a significant performance improvement over the previous state of the art in cross-modal retrieval. Specifically, our model improves Recall@1 by 16.3%p in audio-to-image retrieval and 13.7%p in image-to-audio retrieval. This result indicates that our model does not merely detect salient objects likely to produce sound within the image, but rather effectively locates the target specified by the audio input.

4.3. Ablation Study

Impact of Losses and False Negative Mitigating in Sec. 3.4. In Table 5, we conduct an ablation study to verify the effectiveness of each loss function and the strategy outlined in Sec. 3.4. The overall results show that the designed losses and the false negative mitigating strategy each contribute to improve the performance. Notably, the experiment with $\mathcal{L}_{\text{match}}$ demonstrates a drastic increase in performance, highlighting the importance of cross-modality attention matching. Additionally, the ablation study on $\mathcal{L}_{\text{div}}$ shows that explicitly separating the slots enables each slot to perform its designated role more effectively.

Impact of Shared Initial Slots. We conduct an ablation study to examine the impact of shared initial slots by using different initial slots for each modality. As shown in Table 6, sharing initial slots resulted in better performance, indicating that this design choice facilitates the ability to jointly capture sound source-related information across different modalities.

4.4. Qualitative Results

Quality of SSL results. Fig. 3 visualizes the qualitative results of our model and previous studies on Flickr-SoundNet [4] and VGG-SS [5] test sets. These results demonstrate that our model has the ability to effectively localize sound source objects. Notably, it produces compact localization results even for challenging samples that other models struggle with, such as small birds, a gun, and cars surrounded by other objects in street scenes.

Qualitative Analysis of $\mathcal{L}_{\text{match}}$. Fig. 4 shows the qualitative result that shows the effect of $\mathcal{L}_{\text{match}}$ on Flickr-SoundNet-Test. The attention maps is generated between the audio query of the audio target slot and the image keys derived from image features. Without $\mathcal{L}_{\text{match}}$, the alignment between the two is insufficient for compact localization, whereas with $\mathcal{L}_{\text{match}}$, the alignment improves, resulting in more accurate localization of the sound source.

Cross-modal Retrieval. Fig. 5 shows the qualitative results of our model for cross-modal retrieval. Consistent with the impressive performance shown in Table 4, our model demonstrates a strong ability to retrieve semantically similar samples across modalities that correspond to the sound source target.

5. Conclusion

This paper presented a novel self-supervised sound source localization method leveraging joint slot attention to align image and audio features. Our approach decomposes these features into target and off-target slots, using only target slots for contrastive learning to capture sound-related features. In addition, we introduced cross-modal attention matching and slot divergence loss which significantly improves localization accuracy and and we demonstrated the effectiveness of our model through comprehensive analysis. For future work, we aim to extend our framework to multi-source sound source localization.

Acknowledgement. This work was supported by the NRF grant (RS-2021-NR059830–40%) and the IITP grants (RS-2024-00509258–40%, RS-2021-II212068–10%, RS-2019-II191906–10%) funded by Ministry of Science and ICT, Korea.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 6

[2] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15, 2002. 1

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proc. European Conference on Computer Vision (ECCV)*, pages 435–451, 2018. 1

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Proc. Neural Information Processing Systems (NeurIPS)*, 29, 2016. 2, 5, 6, 8

[5] 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 (ICASSP)*, pages 721–725. IEEE, 2020. 2, 5, 6, 8

[6] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 3, 5, 6, 7

[7] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, 2014. Association for Computational Linguistics. 4

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 5

[9] 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 *Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 2278–2287, 2023. 3

[10] Yury Furletov, Volker Willert, and Jürgen Adamy. Auditory scene understanding for autonomous driving. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 697–702. IEEE, 2021. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[12] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3, 6

[13] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proc. International Conference on Machine Learning (ICML)*, pages 2127–2136. PMLR, 2018. 1

[14] Dongjin Kim, Sung Jin Um, Sangmin Lee, and Jung Uk Kim. Learning to visually localize sound sources from mixtures without prior source knowledge. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26467–26476, 2024. 1, 3, 6

[15] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proc. ACM Multimedia Conference (ACMMM)*, pages 7590–7598, 2023. 1

[16] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *Proc. ACM Multimedia Conference (ACMMM)*, pages 3742–3753, 2022. 3, 7

[17] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Proc. Neural Information Processing Systems (NeurIPS)*, 33:11525–11538, 2020. 1, 3

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. International Conference on Learning Representations (ICLR)*, 2019. 5

[19] Tanvir Mahmud, Yapeng Tian, and Diana Marculescu. T-vsl: Text-guided visual sound source localization in mixtures. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26742–26751, 2024. 3

[20] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 1

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

[22] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3, 6, 7

[23] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10565–10574, 2023. 3

[24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 4

[25] Sooyoung Park, Arda Senocak, and Joon Son Chung. Can clip help sound source localization? In *Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2024. 1, 3

[26] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 6

[27] 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*, pages 8748–8763. PMLR, 2021. 1, 4

[28] Yoko Sasaki, Ryo Tanabe, and Hiroshi Takernura. Online spatial sound perception using microphone array on mobile robot. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2478–2484. IEEE, 2018. 1

[29] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 6

[30] 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 (ICASSP)*, 2022. 6

[31] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 5, 6, 7

[32] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3222–3231, 2022.

[33] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3, 5, 6, 7

[34] Jun Yin, Stefano Damiano, Marian Verhelst, Toon van Waterschoot, and Andre Guntoro. Real-time acoustic perception for automotive applications. In *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–6. IEEE, 2023. 1

[35] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[36] 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 *Proc. European Conference on Computer Vision (ECCV)*, pages 386–403, 2022. 2, 5