# Learning to Visually Localize Sound Sources from Mixtures without Prior Source Knowledge

Dongjin Kim[1,*]     Sung Jin Um[1,*]     Sangmin Lee[2,†]     Jung Uk Kim[1,†]

[1]Kyung Hee University     [2]University of Illinois Urbana-Champaign

{rlaehdwls310, sungzin1, ju.kim}@khu.ac.kr, sangminl@illinois.edu

## Abstract

*The goal of the multi-sound source localization task is to localize sound sources from the mixture individually. While recent multi-sound source localization methods have shown improved performance, they face challenges due to their reliance on prior information about the number of objects to be separated. In this paper, to overcome this limitation, we present a novel multi-sound source localization method that can perform localization without prior knowledge of the number of sound sources. To achieve this goal, we propose an iterative object identification (IOI) module, which can recognize sound-making objects in an iterative manner. After finding the regions of sound-making objects, we devise object similarity-aware clustering (OSC) loss to guide the IOI module to effectively combine regions of the same object but also distinguish between different objects and backgrounds. It enables our method to perform accurate localization of sound-making objects without any prior knowledge. Extensive experimental results on the MUSIC and VGGSound benchmarks show the significant performance improvements of the proposed method over the existing methods for both single and multi-source. Our code is available at: https://github.com/VisualAIKHU/NoPrior_MultiSSL.*

## 1. Introduction

Humans naturally perceive various sounds and identify the origins of those sounds by using both visual sense (eyes) and auditory sense (ears) [12]. The sound source localization task aims to mimic the ability of humans to correlate auditory cues with visual information in order to identify sound-making objects. Due to this property, sound source localization is closely related to various real-world applications, including unmanned aerial vehicles [14, 30], robotics [20, 28], and speaker source localization [6, 31].

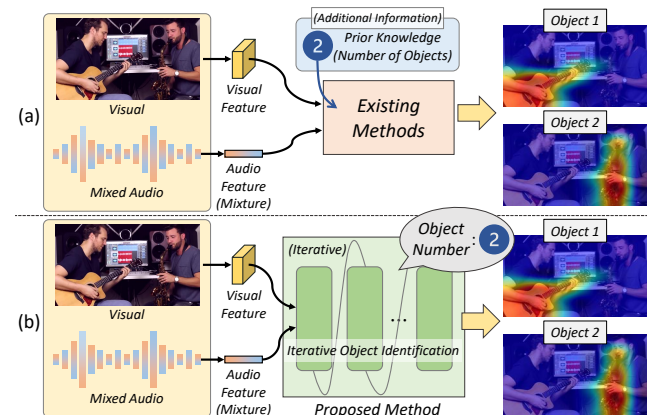The sound source localization task can be divided into



Figure 1. Conceptual comparison between (a) existing methods and (b) the proposed method. The existing methods require prior source knowledge of the number of sound-making objects. In contrast, our method can effectively localize multiple sound-making objects without the need for prior source knowledge.

two categories: (1) single sound source localization and (2) multi-sound source localization. The single sound source localization task [8, 11, 21, 22, 25, 32–36, 38, 39, 41, 44, 48] aims to find one source in a scene by utilizing cross-modal correlations [32, 40] between audio and visual cues. Various methods have been developed for the effective single sound source localization by introducing hard positive mining [8, 33], iterative learning [21], feature regularization [22], negative free learning [38], false negative aware learning [39], momentum target encoders [25], optical flow [11, 36], and spatial integration [41]. However, these methods focus primarily on locating a single sound source, which can be challenging in real-world environments where multiple sounds are often mixed together.

In response to this challenge, several multi-sound source localization methods [16, 17, 26, 27] have been developed. The main goal of multi-sound source localization is to separate and localize individual sources from complex mixtures containing different sounds, such as self-supervised audio-visual matching [16], coarse-to-fine manner [27], contrastive random walker [17], and audio-visual grouping network [26]. However, a limitation of existing methods is

---

*Equal contribution
†Corresponding author

their reliance on prior information about the number of objects that need to be separated. As shown in Figure 1(a), existing methods heavily rely on prior knowledge about the number of sound sources. If prior knowledge is incorrect, they are frequently failed to localize sound-making objects. Thus, they can only operate in constrained environments where prior source knowledge is available for sound source localization. Consequently, accurate localization of multi-sound sources becomes challenging when this prior knowledge is not available. In addition, since prior knowledge is generally not provided in real-world environments, it limits their applicability in practical scenarios.

To address the aforementioned challenges, we introduce a novel method for multi-sound source localization that can identify multiple sound sources without the need for prior knowledge of the number of sources. As shown in Figure 1(b), our method can adapt to various numbers of sound sources by automatically recognizing the number of sound-making objects without relying on any prior knowledge. To this end, we propose an Iterative Object Identification (IOI) module. The goal of our IOI is to effectively automate object separation by repeatedly recognizing objects that make sounds from the mixtures in an iterative manner. By doing so, it continuously searches for sound-making regions until it determines that all relevant objects have been identified.

To achieve this goal, we provide guidance to the IOI module for effective sound-making object identification. First, in the iterative process, we identify regions that are considered to be foreground objects and repeat this process until there are no more regions considered as foreground objects. At each iteration step, regions previously considered as foreground objects are eliminated in the subsequent iteration. Next, we employ a clustering process to merge regions belonging to the same object based on the foreground regions. To effectively carry out this process, we devise an object similarity-aware clustering (OSC) loss. This loss not only guides the IOI module to combine regions of the same object but also aids in distinguishing between different objects and backgrounds. By doing so, the proposed framework is able to distinguish between various objects with distinct sounds through the iterative process.

Consequently, our method can precisely locate the sound-making objects without the need for prior source knowledge, and can effectively distinguish between multiple sound sources. As a result, our method shows significant improvements in sound source localization performance compared to existing methods.

The major contributions of our paper are as follows:

- We introduce Iterative Object Identification (IOI) module to adaptively localize multiple sound sources without any prior source knowledge. To the best of our knowledge, this is the first attempt to perform multi-sound source localization without knowing the number of sound sources.

- We propose Object Similarity-aware Clustering (OSC) loss to guide the IOI module to merge regions belonging to the same object during the iteration process, while distinguishing between different objects and the background.

- Experimental results on MUSIC and VGGSound datasets demonstrate the effectiveness of the proposed method for both single-/multi-source sound localization.

## 2. Related Work

### 2.1. Sound Source Localization

The sound source localization focuses on finding the spatial location of sound sources within a video frame. It can be divided into two main streams: (1) single sound source localization and (2) multi-sound source localization.

The single sound source localization [8, 9, 11, 21, 22, 25, 32–36, 38, 39, 41, 44, 48] leverages the cross-modal inter-relations between auditory and visual modalities to localize a sound-making object. Senocak et al. [32] introduced an unsupervised method to incorporate an audio-visual two-stream network with attention mechanisms. Chen et al. [8] proposed a framework that automatically identifies hard samples through contrastive learning. Lin et al. [21] adopted iterative learning with pseudo-labels to refine the localization process. In addition, Fedorishin et al. [11] and Singh et al. [36] leveraged optical flow as a prior and introduced a cross-attention mechanism over the relevant video frame. Sun et al. [39] introduced the False Negative Aware Contrastive (FNAC) strategy to guide similar-looking samples to be more similar. Senocak et al. [34] improved localization accuracy through enhanced cross-modal semantic understanding. Um et al. [41] proposed the spatial knowledge integration of the audio-visual modalities, promoting recursive enhancement of localization. However, in real-world scenarios, they encounter limitations due to the frequent occurrence of mixed sounds from various sources.

Recent studies focus on the multi-sound source localization [16, 17, 26, 27] to effectively find the location of sound sources from mixtures. Hu et al. [16] introduced a technique that constructs a supervisory signal to differentiate the sound-making entities within a scene. Qian et al. [27] developed a two-stage audiovisual learning approach that separates audio and visual representations of various categories from scenes in a coarse-to-fine manner. Hu et al. [17] made a graph model for robust multi-sound source localization. Mo et al. [26] proposed a way to simultaneously extract and assimilate category-specific semantic features for each sound source from auditory and visual inputs.

However, the existing multi-sound source localization methods also exhibit a limitation in real-world scenarios. They heavily rely on prior knowledge about the number of sound sources to differentiate between multi-sound sources. Thus, we propose a sound source localization framework
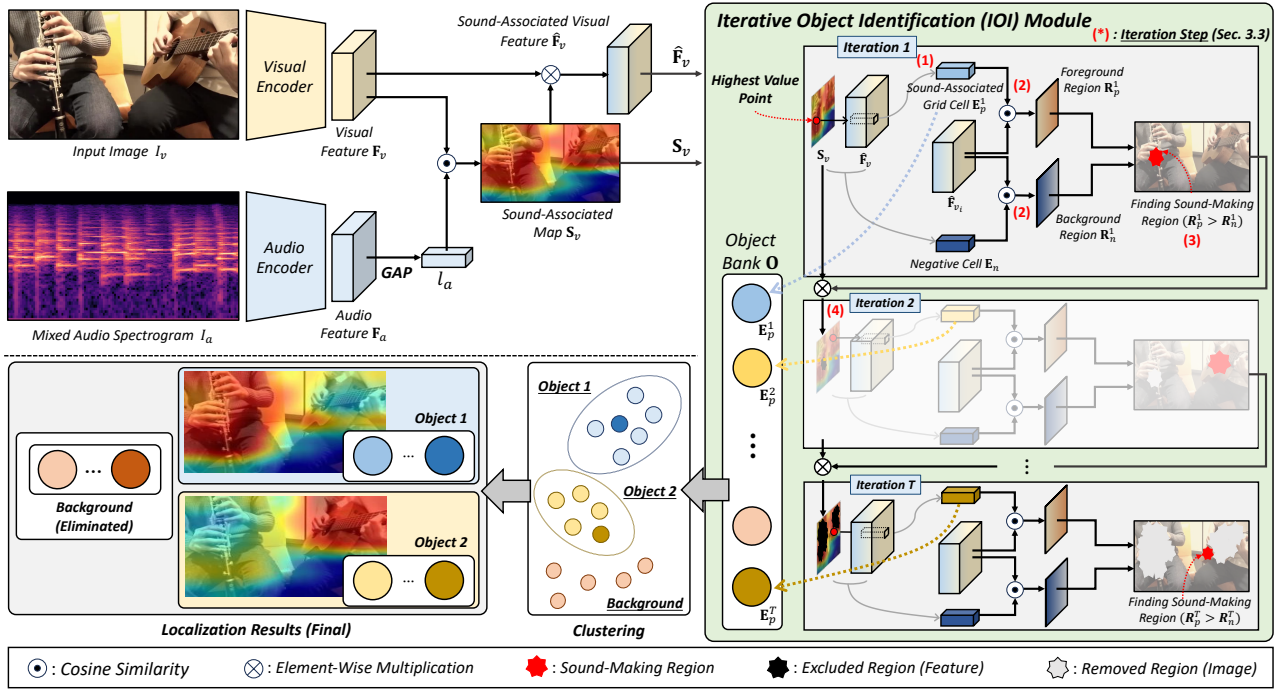
Figure 2. Network configuration of the proposed sound source localization framework. GAP denotes global average pooling.

that can autonomously recognize the number of objects within a scene and localize the objects without relying on annotations or prior knowledge.

## 2.2. Iterative Methods in Computer Vision

Iterative techniques have been applied to various research fields [1, 4, 5, 19, 23, 29, 37, 42, 43, 45, 47] because they provide a refining solutions through repetitive cycles. For example, the iterative methods have been applied to image deblurring and denoising tasks [5, 19], showing significant improvements over traditional non-iterative techniques. In addition, for the mainstrean computer vision tasks, e.g., object detection and segmentation tasks, various works [1, 37, 42, 45, 47] also adopt the iterative refinement techniques for more improved performance.

The iterative technique possesses inherent characteristic of iteratively converging towards the desired solution. In this study, we adopt this technique to perform effective multi-sound source localization. Though this approach, we automatically and iteratively identify multiple sound sources, enhancing the localization for the same objects and allowing differentiation between different objects. As a result, our method demonstrates improved performance compared to conventional methods. Moreover, due to its automatic object-finding nature, our iterative approach is not constrained by prior source knowledge.

## 3. Proposed Approach

### 3.1. Overall Architecture

Figure 2 shows the overall architecture of our framework. Similar to the previous works [8, 32], we adopt a two-stream

network to extract visual and audio features. Input image set $I_v \in \mathbb{R}^{B \times W_v \times H_v \times 3}$ ($B$ indicates batch number, $W_v$ and $H_v$ denote width and height of $I_v$) and the corresponding mixed audio spectrogram $I_a \in \mathbb{R}^{B \times W_a \times H_a \times 1}$ ($W_a$ and $H_a$ denote width and height of $I_a$) pass through each modal encoder (*i.e.*, visual and audio encoders) to generate visual feature $\mathbf{F}_v$ and audio feature $\mathbf{F}_a$, respectively. First, we identify all the sound-making regions associated with mixed audios to filter the regions of interest. We measure cosine similarity between $\mathbf{F}_v$ and $l_a$, derived from the global average pooling (GAP) of $\mathbf{F}_a$, to obtain a sound-associated map $\mathbf{S}_v$.

Next, we obtain a sound-associated visual feature $\hat{\mathbf{F}}_v$ by the product of $\mathbf{S}_v$ and $\mathbf{F}_v$ to being aware of features for overall sound-making regions. The Iterative Object Identification (IOI) module receives $\hat{\mathbf{F}}_v$ to iteratively discriminate sound-associated grid cells and collect each cell in an object bank $\mathbf{O}$. Subsequently, sound-associated grid cells in $\mathbf{O}$ are allocated into object groups (*e.g.*, object 1, object 2, etc.), and the background regions are removed. Our method differentiates various objects without prior source knowledge (*i.e.*, the number of sound sources).

### 3.2. Sound-Associated Region Localization

To automatically localize sound-making objects in a visual scene without any annotation or prior knowledge, it is essential to measure the correspondence between audio and visual features. To achieve this, we try to distinguish sound-making regions from non-sound regions.

We calculate the cosine similarity between $\mathbf{F}_v \in \mathbb{R}^{B \times w \times h \times c}$ ($w$, $h$, and $c$ are the width, height, and channel) and $l_a \in \mathbb{R}^{B \times c}$ to generate sound-associated map $\mathbf{S}_v = \{S_{v_{ij}}\}_{i=1,...,h,j=1,...,w} \in \mathbb{R}^{B \times w \times h}$. To make $\mathbf{S}_v$

meaningful, our method allows positive audio-visual pairs to be attracted to each other and negative pairs to be repelled at training time. By doing so, it is possible to identify spatial regions associated with given sound (positive pairs) and regions which are not associated with that sound (negative pairs). We define $\mathbf{S}_v^{n \to m}$ as cosine similarity map between $n$-th image and $m$-th audio, and the $ij$-th element of $\mathbf{S}_v^{n \to m}$, which can be represented as:

$$Sim(A, B) = \frac{\langle A, B \rangle}{||A|| \, ||B||},$$
$$S_{v_{ij}}^{n \to m} = \frac{Sim(\mathbf{F}_{v_{ij}}^n, l_a^m)}{\sum_{i=1}^{h} \sum_{j=1}^{w} Sim(\mathbf{F}_{v_{ij}}^n, l_a^n)}, \quad (1)$$

where $\langle \cdot, \cdot \rangle$ denotes inner-product. Next, we obtain the mask of $n$-th clip $M^n$, represented as:

$$M^n = sigmoid((S_v^{n \to n} - \alpha)/\omega), \quad (2)$$

where $\alpha$ and $\omega$ are the hyper-parameters. Based on $M^n$, we aim to localize all the sound-making regions associated with mixed audios. Thus, motivated by [8], we employ an audio-visual contrastive loss $\mathcal{L}_{avc}$, which can be represented as:

$$Pos^n = \frac{1}{|M^n|} \langle M^n, S_v^{n \to n} \rangle,$$
$$Neg^n = \frac{\langle \mathbf{1} - M^n, S_v^{n \to n} \rangle}{|\mathbf{1} - M^n|} + \frac{1}{hw} \sum_{n \neq m} \langle \mathbf{1}, S_v^{n \to m} \rangle, \quad (3)$$
$$\mathcal{L}_{avc} = -\frac{1}{B} \sum_{n=1}^{B} \left[ \log \frac{\exp(Pos^n)}{\exp(Pos^n) + \exp(Neg^n)} \right].$$

$Pos^n$ represents the average vector of highly correlated regions between the audio-visual pairs in the $n$-th clip. In contrast, $Neg^n$ contains not only regions with low correlated between audio-visual pairs within the same clip but also includes correlation with other clips. Consequently, $\mathcal{L}_{avc}$ allows the $\mathbf{S}_v$ to effectively filter out the foreground regions associated with the mixed audio, separating from the background regions that do not produce sound.

### 3.3. Iterative Object Identification Module

Through Section 3.2, the sound-associated map $\mathbf{S}_v$ can contain knowledge of the sound-associated objects. We then multiply $\mathbf{S}_v$ by the original visual feature $\mathbf{F}_v$ to encode sound-associated visual feature $\hat{\mathbf{F}}_v \in \mathbb{R}^{B \times w \times h \times c}$. By doing so, background cells are removed, leaving all the sound-associated cells that represent the sound-making object.

Next, the proposed Iterative Object Identification (IOI) module takes $\hat{\mathbf{F}}_v$ to identify sound-making objects in mixed audio-visual data. Using $\hat{\mathbf{F}}_v$, the IOI module conducts a total of $T$ iterations to find all sound-associated cells, and in each iteration, it finds the area to localize around the position of the sound-associated cell identified. Each iteration consists of four steps: (1) Highest sound-associated cell

---

**Algorithm 1** Iterative Object Identification Algorithm

---

**Require:** $\mathbf{S}_v, \hat{\mathbf{F}}_v$
**Ensure:** $\mathbf{O} = [\,], t = 0$
  **while** $\max(\mathbf{S}_v) > \varepsilon$ **do**
    $\mathbf{E}_p^t \leftarrow \hat{\mathbf{F}}_v[\text{argmax}(\mathbf{S}_v)]$
    $\mathbf{R}_p \leftarrow \langle \hat{\mathbf{F}}_v, \mathbf{E}_p^t \rangle$                    ▷ Inner Product
    $\mathbf{R}_n \leftarrow \langle \hat{\mathbf{F}}_v, \mathbf{E}_n \rangle$
    $sound\_making\_region \leftarrow \mathbf{R}_p > \mathbf{R}_n$
    $\mathbf{S}_v \leftarrow \mathbf{S}_v \cdot \neg sound\_making\_region$
    $\mathbf{O}.append(\mathbf{E}_p^t)$
    $t \leftarrow t + 1$
  **end while**
**Results:** $T = t, \mathbf{O} = [\mathbf{E}_p^1, \mathbf{E}_p^2, ..., \mathbf{E}_p^T]$

---

selection, (2) foreground/background selection, (3) finding sound-making regions, and (4) identified sound-making region exclusion. After the end of $T$ iteration process, we cluster all sound-associated grid cells as each object or negative.

Before starting iteration, we define a negative cell vector $\mathbf{E}_n \in \mathbb{R}^{B \times c}$ to represent background (non-sound-making objects). $\mathbf{E}_n$ is obtained by averaging the background cells that were excluded during the creation of $\hat{\mathbf{F}}_v$. This vector is essential for identifying regions that do not correspond to the sound, to provide contrast to the foreground.

**(1) Highest sound-associated cell selection.** In the first step in $t$-th iteration, we select the cell in $\mathbf{S}_v$ with the highest value and select this cell as a starting point. This selection is based on the understanding that the highest value in $\mathbf{S}_v$ indicates the strongest audio-visual correspondence in that cell. Therefore, we identify and mark the coordinates of this cell as $(i, j)$. The vector $\hat{\mathbf{F}}_{v_{ij}}$ is then assigned as the sound-associated grid cell $\mathbf{E}_p^t \in \mathbb{R}^{B \times c}$. Then, we store $\mathbf{E}_p^t$ in the $i$-th element of the object bank $\mathbf{O} = \{\mathbf{E}_p^t\}_{t=1}^{T}$.

**(2) Foreground/background selection.** In the second step in $t$-th iteration, we perform foreground and background selection using $\mathbf{E}_p^t$ and $\mathbf{E}_n$. The similarity measurement is carried out by calculating the inner product of $\hat{\mathbf{F}}_v$ and $\mathbf{E}_p^t$. This result is used as foreground region $\mathbf{R}_p^t \in \mathbb{R}^{B \times w \times h}$. A higher value in $\mathbf{R}_p^t$ indicates a higher likelihood of being the same object as $\mathbf{E}_p^t$. Similarly, the background region is calculated by inner product of $\hat{\mathbf{F}}_v$ and $\mathbf{E}_n$, resulting in $\mathbf{R}_n^t \in \mathbb{R}^{B \times w \times h}$.

**(3) Finding sound-making regions.** In the third step in $t$-th iteration, we compare $\mathbf{R}_p^t$ and $\mathbf{R}_n^t$ to localize sound-making regions. The choice of sound-making region sizes is based on the condition that the value of foreground region $\mathbf{R}_p^t$ is higher than that of the background $\mathbf{R}_n^t$. Using these regions, we generate the $t$-th localization map.

**(4) Identified sound-making region exclusion.** In the fourth step in $t$-th iteration, to prevent redundancy in finding sound-associated grid cell vector $\mathbf{E}_p^t$, we need to modify $\mathbf{S}_v$ for the next iteration. This involves setting the values of regions already localized to zero in $\mathbf{S}_v$, thereby excluding
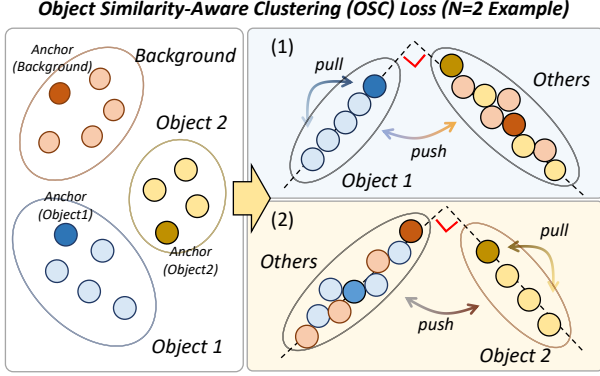
**Object Similarity-Aware Clustering (OSC) Loss (N=2 Example)**

Figure 3. Explanation of the proposed Object Similarity-Aware Clustering (OSC) loss ($N = 2$ example).

them from further iteration. After completing these steps, we move on to the next iteration. We repeat this process until we find all sound-associated grid cells. The pseudocode detailing this procedure is presented in Algorithm 1, offering a clear step-by-step guide.

### 3.4. Object Identification and Final Localization.

After finishing the iteration process, we identify all sound-associated grid cells as each object or negative. Each cell is evaluated to determine whether it represents a sound-making object or background. This classification is crucial as it ensures the accuracy of our object identification.

First, to eliminate cells that were identified as sound-associated grid cells but are actually background, we calculate cosine similarity between negative cell vector $\mathbf{E}_n$ and each sound-associated grid cell $\mathbf{E}_p^t$ in object bank $\mathbf{O}$. Cells with a similarity exceeding a threshold $\tau_1$ are considered to be a part of the background and are eliminated from $\mathbf{O}$.

Subsequently, to cluster the remaining sound-associated grid cells into each object, we calculate the cosine similarity between every pair of cells. This calculation helps us understand how similar or related these cells are to each other. If the similarity between any two cells is higher than the threshold $\tau_2$, we consider these cells to be part of the same object. To effectively cluster each object, we use the union-find algorithm [2]. It merges other cells into each group if they are determined to belong to the same object.

Each identified group is labeled as object $k$ (where $k = 1, 2, ..., K_b$). Here, $K_b$ is the number of objects in $b$-th clip. Finally, we combine the localization maps from object $k$ to create the final localization map for each object.

### 3.5. Object Similarity-aware Clustering Loss

To effectively identify and separate multi-sound sources without any annotation and prior source knowledge in complex environments, we propose an Object Similarity-aware Clustering (OSC) loss. The OSC loss has an important role in guiding the IOI module to combine regions belonging to the same object while differentiating between distinct ob-

jects and backgrounds. In the bank $\mathbf{O}_b$, the anchor cell $A_k^b$ is identified as the most representative cell of the object $k$ in the $b$-th clip. $P_k^b$ and $N_k^b$ represent compositions of cells, where $P_k^b$ consists of cells attributed to the same object, and $N_k^b$ is from different objects or backgrounds ($P_k^b$ and $N_k^b$ are distinguished though Section 3.4). The formulation of the OSC loss is structured as follows:

$$
\begin{aligned}
C_{pk}^b = Sim(A_k^b, P_k^b), \quad C_{nk}^b = Sim(A_k^b, N_k^b), \\
\mathcal{L}_k^b = \left(1 - C_{pk}^b\right) + C_{nk}^b, \\
\mathcal{L}_{osc} = \frac{1}{B} \sum_{b=1}^{B} \sum_{k=1}^{K_b} \frac{\mathcal{L}_k^b}{K_b},
\end{aligned}
\tag{4}
$$

where $C_{pk}^b$ and $C_{nk}^b$ represent the cosine similarity between the anchor and the positive, and the anchor and the negative, respectively. Through this loss, learning is directed in such a way that similarity between the anchor and positive cells is reinforced, while divergence between the anchor and negative cells is accentuated. As shown in Figure 3, this effectively clusters cells that are considered to be part of the same object, thereby aiding in the segmentation of objects. This approach reduces the number of iterations to distinguish objects, as detailed in our supplementary experiments.

### 3.6. Training Objective

To train our proposed method, we construct the total training loss function as follows:

$$
\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{avc} + \lambda_2 \mathcal{L}_{osc},
\tag{5}
$$

where $\lambda_1$ and $\lambda_2$ denote balancing parameters for each loss function. By using $\mathcal{L}_{Total}$, our method effectively performs multi-sound source localization, identifying sound sources from mixtures without the need for prior source knowledge.

## 4. Experiments

### 4.1. Datasets and Evaluation Metrics

**MUSIC.** The MUSIC dataset [46] comprises 448 unedited YouTube music videos featuring solos and duets across 11 categories of musical instruments. To ensure a fair comparison with previous works, we use the same training/testing subset as utilized in AVGN [26]. Specifically, the MUSIC-Solo contains 358 solo videos that are applied for training, and 90 solo videos are for evaluation for single sound source localization. Similarly, the MUSIC-Duet contains 124 duet videos that are applied for training, and 17 duet videos are for evaluation for multi-sound source localization.

**VGG-Sound.** VGG-Sound dataset [7], which is denoted as VGGSound-Single consists of more than 200k videos from 221 different sound categories. We utilize 144k image-audio pairs as a training set. For single sound source localization, we use VGG-Sound Source [8] to evaluate our

| Method | MUSIC-Duet [46] | | | | VGGSound-Duet [7] | | | |
|---|---|---|---|---|---|---|---|---|
| | CAP(%) | PIAP(%) | CIoU@0.3(%) | AUC(%) | CAP(%) | PIAP(%) | CIoU@0.3(%) | AUC(%) |
| Attention10k [32] (CVPR'18) | – | – | 21.6 | 19.6 | – | – | 11.5 | 15.2 |
| OTS [3] (ECCV'18) | 11.6 | 17.7 | 13.3 | 18.5 | 10.5 | 12.7 | 12.2 | 15.8 |
| DMC [15] (CVPR'19) | – | – | 17.5 | 21.1 | – | – | 13.8 | 17.1 |
| CoarseToFIne [27] (ECCV'20) | – | – | 17.6 | 20.6 | – | – | 14.7 | 18.5 |
| DSOL [16] (NeurIPS'20) | – | – | 30.1 | 22.3 | – | – | 22.3 | 21.1 |
| LVS [8] (CVPR'21) | – | – | 22.5 | 20.9 | – | – | 17.3 | 19.5 |
| EZ-VSL [24] (ECCV'22) | – | – | 24.3 | 21.3 | – | – | 20.5 | 20.2 |
| Mix-and-Localize [17] (CVPR'22) | 47.5 | 54.1 | 26.5 | 21.5 | 16.3 | 22.6 | 21.1 | 20.5 |
| AVGN [26] (CVPR'23) | <u>50.6</u> | <u>57.2</u> | <u>32.5</u> | <u>24.6</u> | <u>21.9</u> | <u>28.1</u> | <u>26.2</u> | <u>23.8</u> |
| **Proposed Method** | **52.1** | **72.5** | **38.6** | **30.1** | **32.5** | **44.4** | **46.9** | **29.2** |

Table 1. Experimental results on Music-Duet(left) and VGGSound-Duet(right) for multi-sound source localization. **Bold**/<u>underlined</u> fonts indicate the best/second-best results.

| Method | MUSIC-Solo [46] | | | VGGSound-Single [7] | | |
|---|---|---|---|---|---|---|
| | AP(%) | IoU@0.5(%) | AUC(%) | AP(%) | IoU@0.5(%) | AUC(%) |
| Attention10k [32] (CVPR'18) | – | 37.2 | 38.7 | – | 19.2 | 30.6 |
| OTS [3] (ECCV'18) | 69.3 | 26.1 | 35.8 | 29.8 | 32.8 | 35.7 |
| DMC [15] (CVPR'19) | – | 29.1 | 38.0 | – | 23.9 | 27.6 |
| CoarseToFIne [27] (ECCV'20) | 70.7 | 33.6 | 39.8 | 28.2 | 29.1 | 34.8 |
| DSOL [16] (NeurIPS'20) | – | 51.4 | 43.7 | – | 35.7 | 37.2 |
| LVS [8] (CVPR'21) | 70.6 | 41.9 | 40.3 | 29.6 | 34.4 | 38.2 |
| EZ-VSL [24] (ECCV'22) | 71.5 | 45.8 | 41.2 | 31.3 | 38.9 | 39.5 |
| Mix-and-Localize [17] (CVPR'22) | 68.6 | 30.5 | 40.8 | 32.5 | 36.3 | 38.9 |
| AVGN [26] (CVPR'23) | <u>77.2</u> | <u>58.1</u> | <u>48.5</u> | <u>35.3</u> | <u>40.8</u> | **42.3** |
| **Proposed Method** | **77.4** | **62.1** | **59.4** | **46.2** | **41.4** | <u>41.2</u> |

Table 2. Experimental results on Music-Solo(left) and VGGSound-Single(right) for single sound source localization. **Bold**/<u>underlined</u> fonts indicate the best/second-best results.

method. In the training phase, for multi-sound source localization, we randomly concatenate two video frames to generate a single input image with dimensions of $448 \times 224$, and we combine the corresponding audio waveforms to generate a mixed audio signal. Following [26], we employ the VGGSound-Duet dataset for evaluation.

**Evaluation Metrics.** Following the prior works [17, 26], we adopt metrics for a fair and comprehensive comparison. For single sound source localization, we employ Average Precision (AP), Intersection over Union (IoU), and Area Under Curve (AUC). Regarding multi-sound source localization, we employ Class-aware Average Precision (CAP), Permutation-Invariant Average Precision (PIAP), Class-aware IoU (CIoU), and Area Under Curve (AUC). Our method is self-supervised and does not use class labels, we follow [17] for the modified version of CAP and CIoU.

### 4.2. Implementation Details

For the visual modality input, we resize images to dimensions of $W_v = 224$, $H_v = 224$, extracting them from the central frame of 3-second video clips. For the audio modality, we resample the raw 3-second audio signal to 16kHz and convert it into a log-scale spectrogram.

In accordance with the methods presented in [8], we utilize a ResNet-18 [13] for both the visual and audio feature

backbones to establish our model. Due to the sound spectrogram having only one channel, we adapt the first convolutional layer of the ResNet-18 [13] from 3 channels to 1. The visual encoder is pretrained using ImageNet [10]. Our method employs the Adam optimizer [18] with a learning rate of $10^{-4}$ and a batch size of 128. We train our model for 50 epochs on all experiments. The training is conducted using 1 RTX 4090 GPU. We use the weights of $\tau_1 = 0.7$, $\tau_2 = 0.6$. For our proposed loss function $\mathcal{L}_{Total}$, we set $\lambda_1 = 1$ and $\lambda_2 = 1$. Following [8, 11], we utilize the implementation code and adopt the same other hyper-parameters.

### 4.3. Comparison to Prior Works

**Multi-Sound Source Localization.** First, we conduct experiments on the multi-sound source localization scenarios. We compare our method with the state-of-the-art multi-sound source localization methods [3, 8, 15–17, 24, 26, 27, 32]. Table 1 shows the results of our method on MUSIC-Duet [46] and VGGSound-Duet [7]. For the MUSIC-Duet test set, our method demonstrates superior performance, achieving 1.5%, 15.3%, 6.1%, and 5.5% higher for CAP, PIAP, CIoU@0.3, and AUC, respectively. Furthermore, in the VGGSound-Duet dataset, our method significantly outperforms existing methods. It indicates that our OSC loss enables our method to effectively localize sound-making
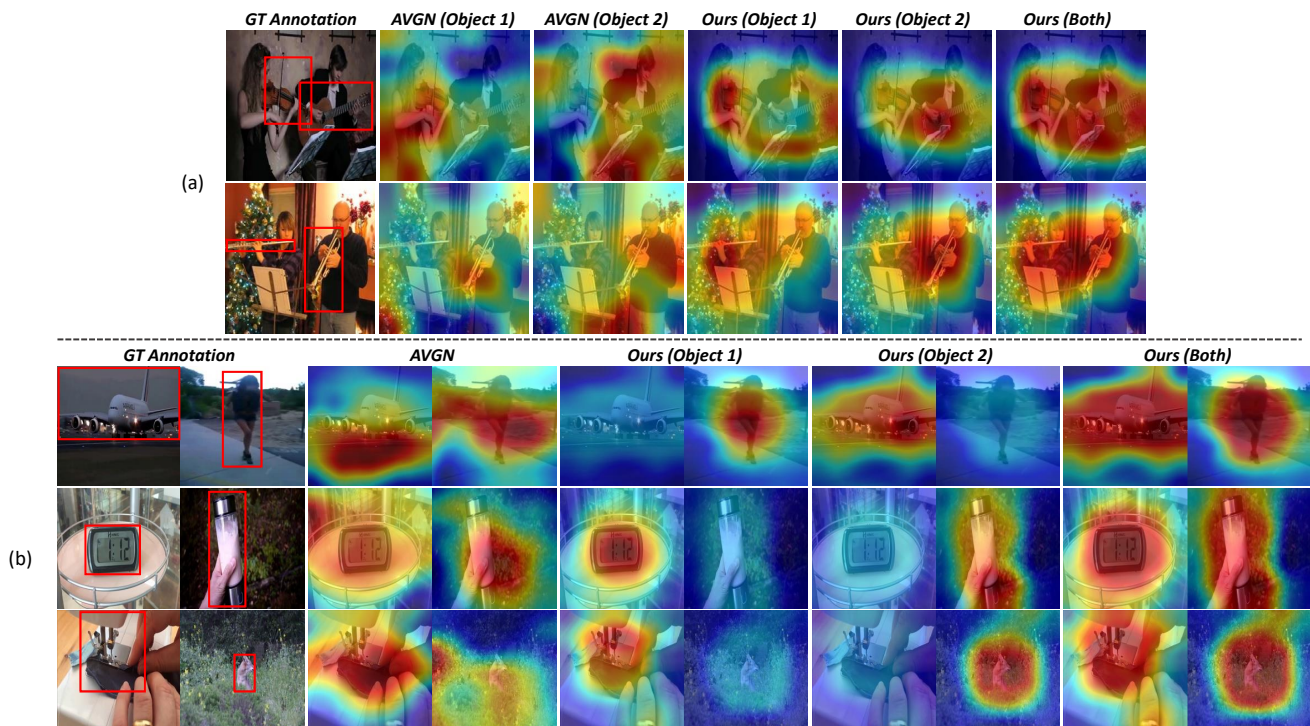
Figure 4. Visualization results for (a) MUSIC-Duet, (b) VGGSound-Duet test set. We compare our method with AVGN [26].
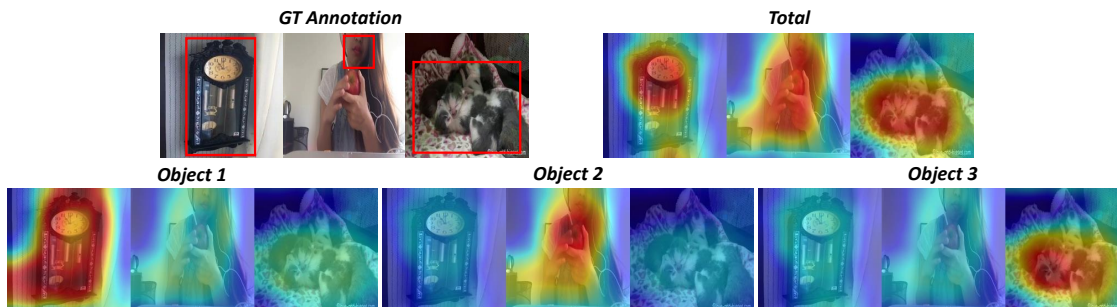


Figure 5. Visualization results of our method on VGGSound-Trio test set.

objects through an iterative process.

**Single Sound Source Localization.** For single sound source localization, we compare our method with other existing works [3, 8, 15–17, 24, 26, 27, 32]. Table 2 presents the performances on MUSIC-Solo [46] and VGGSound-Single datasets [7]. For the MUSIC-Solo test set, ours shows 0.2% higher for AP. 4.0% higher for IoU@0.5, and 10.9% higher for AUC, respectively. For the VGGSound-Single test set, ours outperforms most of the others. For AUC, ours achieves the second-best performance.

The experimental results in Table 1 and Table 2 clearly demonstrate the significant advancements our method brings to single and multi-sound source localization. This progress is largely attributed to the capabilities of the IOI module and OSC loss in effectively distinguishing objects.

## 4.4. Visualization Results

We compare ours with the state-of-the-art approach, AVGN [26], in multi-sound source localization. This comparison

utilizes the MUSIC-Duet and VGGSound-Duet test sets for visualizing the localization results. The results are shown in Figure 4. The visualization highlights the efficacy of our method in accurately localizing sound-making objects within a mixture. Ground Truth (GT) annotations are used to denote regions where sound-making objects are present.

Furthermore, Figure 5 offers additional visualization results from our method, showcasing its efficacy in differentiating objects in scenarios with three-source mixtures. We utilize VGGSound-Trio test set comprising a mixture of three sound sources from VGGSound-Single [7], following [26]. Our IOI module facilitates the repeated detection and distinction of objects in the audio-visual scene, culminating in highly accurate and refined localization maps.

## 4.5. Ablation Study

**Effect of the Proposed Losses.** We evaluate the performance impact of our two proposed losses $\mathcal{L}_{avc}$ and $\mathcal{L}_{OSC}$. The results are shown in Table 3. When each loss is consid-

| Method | $\mathcal{L}_{avc}$ | $\mathcal{L}_{osc}$ | PIAP(%) | CloU@0.3(%) | AUC(%) |
|---|---|---|---|---|---|
| AVGN [26] | - | - | 28.1 | 26.2 | 23.8 |
| Proposed Method | ✓ | - | 44.1 | 25.3 | 28.3 |
|  | - | ✓ | 44.3 | 43.0 | 27.0 |
|  | ✓ | ✓ | **44.4** | **46.9** | **29.2** |

Table 3. Effect of the proposed $\mathcal{L}_{avc}$ and $\mathcal{L}_{osc}$ loss on VGGSound-Duet test set.

| $\tau_1$ | $\tau_2$ | PIAP(%) | CloU@0.3(%) | AUC(%) |
|---|---|---|---|---|
|  | 0.6 | **44.4** | **46.9** | **29.2** |
| 0.7 | 0.7 | 44.1 | 43.9 | 27.5 |
|  | 0.8 | 44.3 | 45.0 | 28.2 |
| 0.6 |  | 44.3 | 46.3 | 28.9 |
| 0.7 | 0.6 | **44.4** | **46.9** | **29.2** |
| 0.8 |  | 44.2 | 46.8 | 29.1 |

Table 4. Experimental results on VGGSound-Duet test set according to the hyper-parameters $\tau_1$ and $\tau_2$ for IOI module in Sec 3.4.

ered, our method shows the improved performance against AVGN [26] which is currently a state-of-the-art method. When all the proposed losses are taken into account, we show the highest performance. By integrating the proposed losses into our training, ours achieves increased capacity for learning robust and discriminative features.

**Variation of $\tau_1$ and $\tau_2$.** We conduct an additional ablation study to investigate the effect of our method on the hyper-parameters $\tau_1$ and $\tau_2$ as described in Section 3.3. The results in Table 4 indicate that we obtain the best results when $\tau_1 = 0.7$ and $\tau_2 = 0.6$. Importantly, even when $\tau_1$ and $\tau_2$ are varied, our method still outperforms the existing methods. These results suggest that the proposed model is robust to the variations of hyperparameters.

## 4.6. Discussions

**Sound Source Counting Accuracy.** To validate the accuracy of counting the correct sound source for our model, we employed each VGGSound dataset, which comprises images containing varying sound sources (ranging from one to three per image). Respectively, our model achieved an impressive accuracy of 90.79%, 83.21%, and 68.94% for VGGSound-Single, Duet, and Trio test sets. These results not only highlight our performance in handling various numbers of sound sources but also demonstrate its potential applicability in diverse and dynamically changing audio-visual environments.

**Adaptability of Our Method to Varying Sound Sources.** We present experiments designed to evaluate the adaptability of our method in scenarios with varied sound sources. For this purpose, we utilized two distinct test sets: VGGSound-Trio [26] and VGGSound-Mixed. VGGSound-Mixed is a resampled dataset encompassing sounds from VGGSound-Single, VGGSound-Duet, and VGGSound-Trio. VGGSound-Mixed allows us to demonstrate the capability of our method to localize sound sources effectively, regardless of their quantity. The results are shown in Ta-

| Method | Test Set | CAP(%) | PIAP(%) | CloU@0.3(%) | AUC(%) |
|---|---|---|---|---|---|
| AVGN [26] | VGGSound-Trio | 18.5 | 23.7 | 22.7 | 21.8 |
| **Ours** |  | **29.0** | **42.1** | **34.0** | **29.3** |
| AVGN [26] | VGGSound-Mixed | N/A | 22.9 | N/A | N/A |
| **Ours** |  | **27.9** | **36.0** | **35.2** | **22.7** |

Table 5. Experimental results on VGGSound-Trio and VGGSound-Mixed test set. '$N/A$' denotes Not Available.

| Method | Training (s) (*per iter*) | Inference (s) (*per image*) | #params |
|---|---|---|---|
| AVGN [26] (CVPR'23) | 0.42 | 0.034 | 61.2M |
| **Proposed Method** | 0.45 | 0.027 | 38.5M |

Table 6. The comparisons of training time, inference time, and the number of parameters.

ble 5. For the VGGSound-Trio test set, our method outperforms AVGN [26] across all evaluation metrics. For the VGGSound-Mixed test set, our approach achieves scores of 27.9%, 36.0%, 35.2%, and 22.7% for CAP, PIAP, CIoU@0.3, and AUC, respectively. Note that AVGN requires fixed prior source information, which limits its applicability. These results highlight our effectiveness in object localization, especially without prior source information, attributable to the efficacy of IOI module and loss function.

**Computational Costs.** Table 6 presents comparisons of training time, inference time, and the number of parameters. We compare our method with AVGN [26] which shows the highest performance among the existing methods. Due to the iterative method, our training time is marginally increased (7.14%). However, inference time decreased by 28.3%, and the number of parameters significantly decreased compared to the existing method with transformers.

## 5. Conclusion

In this paper, we propose an innovative approach to multi-sound source localization that does not rely on prior knowledge. The core of our method is the Iterative Object Identification (IOI) module, which effectively identifies sound-making objects through iterative processes. Our object similarity-aware clustering (OSC) loss function successfully guides the IOI module not only to merge regions associated with the same object but also to discern distinct objects from the background. We believe that our approach, the iterative object identification method enhances accuracy and has diverse practical applications.

# References

[1] Bishwo Adhikari and Heikki Huttunen. Iterative bounding box annotation for object detection. In *ICPR*, 2021. 3

[2] Alfred V Aho and John E Hopcroft. *The design and analysis of computer algorithms*. Pearson Education India, 1974. 5

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 6, 7

[4] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty. *arXiv preprint arXiv:2210.03676*, 2022. 3

[5] Jan Biemond, Reginald L Lagendijk, and Russell M Mersereau. Iterative methods for image deblurring. *Proceedings of the IEEE*, 1990. 3

[6] Carlos Busso, Sergi Hernanz, Chi-Wei Chu, Soon-il Kwon, Sung Lee, Panayiotis G Georgiou, Isaac Cohen, and Shrikanth Narayanan. Smart room: Participant and speaker localization and identification. In *ICASSP*, 2005. 1

[7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 5, 6, 7

[8] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[11] 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 *WACV*, 2023. 1, 2, 6

[12] Monica L Hawley, Ruth Y Litovsky, and H Steven Colburn. Speech intelligibility and localization in a multi-source environment. *JASA*, 1999. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[14] Kotaro Hoshiba, Kai Washizaki, Mizuho Wakabayashi, Takahiro Ishiki, Makoto Kumon, Yoshiaki Bando, Daniel Gabriel, Kazuhiro Nakadai, and Hiroshi G Okuno. Design of uav-embedded microphone array system for sound source localization in outdoor environments. *Sensors*, 2017. 1

[15] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, 2019. 6, 7

[16] 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. In *NeurIPS*, 2020. 1, 2, 6

[17] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *CVPR*, 2022. 1, 2, 6, 7

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Taihao Li, Huai Chen, Min Zhang, Shupeng Liu, Shunren Xia, Xinhua Cao, Geoffrey S Young, and Xiaoyin Xu. A new design in iterative image deblurring for improved robustness and performance. *Pattern Recognition*, 2019. 3

[20] Xiaofei Li, Laurent Girin, Fabien Badeig, and Radu Horaud. Reverberant sound localization with a robot head based on direct-path relative transfer function. In *IROS*, 2016. 1

[21] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Unsupervised sound localization via iterative contrastive learning. *CVIU*, 2023. 1, 2

[22] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *ACM MM*, 2022. 1, 2

[23] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *CVPR*, 2020. 3

[24] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *ECCV*, 2022. 6, 7

[25] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *NeurIPS*, 2022. 1, 2

[26] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *CVPR*, 2023. 1, 2, 5, 6, 7, 8

[27] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *ECCV*, 2020. 1, 2, 6, 7

[28] Caleb Rascon and Ivan Meza. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 2017. 1

[29] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, 2022. 3

[30] Daniele Salvati, Carlo Drioli, Giovanni Ferrin, and Gian Luca Foresti. Acoustic source localization from multi-rotor uavs. *TIE*, 2019. 1

[31] Halim Sayoud, Siham Ouamour, and Salah Khennouf. Speaker localization using stereo-based sound source localization. In *WOSSPA*, 2011. 1

[32] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 1, 2, 3, 6, 7

[33] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *ICASSP*, 2022. 1

[34] 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 *ICCV*, 2023. 2

[35] Jiayin Shi and Chao Ma. Unsupervised sounding object localization with bottom-up and top-down attention. In *WACV*, 2022.

[36] Rajsuryan Singh, Pablo Zinemanas, Xavier Serra, Juan Pablo Bello, and Magdalena Fuentes. Flowgrad: Using motion for visual sound source localization. In *ICASSP*, 2023. 1, 2

[37] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *ICIP*, 2022. 3

[38] 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 *CVPR*, 2022. 1, 2

[39] 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 *CVPR*, 2023. 1, 2

[40] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 1

[41] Sung Jin Um, Dongjin Kim, and Jung Uk Kim. Audio-visual spatial integration and recursive attention for robust sound source localization. In *ACM MM*, 2023. 1, 2

[42] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *CVPR*, 2019. 3

[43] Xingkui Wei, Yinda Zhang, Xinlin Ren, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Robust deep iterative refinement for structure from motion. *TPAMI*, 2023. 3

[44] 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 *CVPR*, 2022. 1, 2

[45] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *CVPR*, 2019. 3

[46] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 5, 6, 7

[47] Mingmin Zhen, Jinglu Wang, Lei Zhou, Shiwei Li, Tianwei Shen, Jiaxiang Shang, Tian Fang, and Long Quan. Joint semantic segmentation and boundary detection using iterative pyramid contexts. In *CVPR*, 2020. 3

[48] Xinchi Zhou, Dongzhan Zhou, Di Hu, Hang Zhou, and Wanli Ouyang. Exploiting visual context semantics for sound source localization. In *WACV*, 2023. 1, 2