Audio-Visual Spatial Integration and Recursive Attention for Robust Sound Source Localization

Sung Jin Um*
sungzin1@khu.ac.kr
Kyung Hee University
Yongin-si, South Korea

Dongjin Kim*
rlaehdwls310@khu.ac.kr
Kyung Hee University
Yongin-si, South Korea

Jung Uk Kim†
ju.kim@khu.ac.kr
Kyung Hee University
Yongin-si, South Korea

ABSTRACT

The objective of the sound source localization task is to enable machines to detect the location of sound-making objects within a visual scene. While the audio modality provides spatial cues to locate the sound source, existing approaches only use audio as an auxiliary role to compare spatial regions of the visual modality. Humans, on the other hand, utilize both audio and visual modalities as spatial cues to locate sound sources. In this paper, we propose an audio-visual spatial integration network that integrates spatial cues from both modalities to mimic human behavior when detecting sound-making objects. Additionally, we introduce a recursive attention network to mimic human behavior of iterative focusing on objects, resulting in more accurate attention regions. To effectively encode spatial information from both modalities, we propose audio-visual pair matching loss and spatial region alignment loss. By utilizing the spatial cues of audio-visual modalities and recursively focusing objects, our method can perform more robust sound source localization. Comprehensive experimental results on the Flickr SoundNet and VGG-Sound Source datasets demonstrate the superiority of our proposed method over existing approaches. Our code is available at: https://github.com/VisualAIKHU/SIRA-SSL.

CCS CONCEPTS

• **Information systems → Multimedia information systems**; • **Computing methodologies → Computer vision**.

KEYWORDS

Sound source localization, audio-visual spatial integration, recursive attention, multimodal learning

*Both authors have contributed equally to this work.
†Corresponding author.

Figure 1: Conceptual comparison between (a) existing methods (red) and (b) the proposed method (blue). The existing methods use the spatial information of visual modality as the primary modality to estimate region of sound-making objects (M_v). We observe that the audio modality itself also contains spatial information for estimating regions of the sound-making object (M_a). In our work, we try to integrate the spatial knowledge of the audio-visual modalities (M_{av}) for more accurate sound source localization.

1 INTRODUCTION

Sound source localization aims to identify the location of a sounding object within a visual scene [49]. This task is similar to the innate ability of humans to find the location by correlating sounds heard with their ears and scenes seen with their eyes. Because of this property, sound source localization has a wide range of applications, such as multimodal robotics [30, 37], sound source separation [8], and indoor positioning [3].

Since the sound source localization task utilizes multimodal information (*i.e.,* audio-visual), it is essential to consider how to effectively combine the different two modal information for more accurate localization. In addition, while audio-visual data can be obtained in abundance, manually annotating object locations (*e.g.,* bounding boxes or segmentation masks) is time-consuming and labor-intensive. To address the two issues, several self-supervised approaches [9, 15, 49, 51, 60, 63] have been proposed. Senocak et al. [49] proposed the attention mechanism with unsupervised learning to match the audio-visual information. Chen et al. [9] introduced a network to explicitly mine the hard negative locations from the foreground locations by using sound information. Xuan et al. [60] proposed a proposal-based method that focuses on the region inside the bounding box of each object based on the given sound. In

[15], the optical flow information was additionally incorporated to effectively combine the audio-visual modalities.

However, the above-mentioned methods have in common that, as shown in Figure 1(a), they utilize the audio modality only as an auxiliary role (red) in comparing whether each grid region of the visual modality corresponds to the area of the sounding object. In fact, humans also have the ability to detect the location of an object just by hearing the sound. For example, even when our eyes are closed, we can still perceive the location of a car making a sound by paying our attention to the corresponding spatial area. This is because the spatial information can be inferred by relying on cues such as differences in arrival time, loudness, and spectral content of the sound [18, 42, 50]. As shown in Figure 1(b), we observe that the audio modal itself also contains valuable spatial cues for inferring the sound-making objects.

Moreover, according to [5, 26, 44], when humans receive both visual and auditory information, they naturally generate a region of interest (ROI) in each modality. These ROIs are then integrated to form a region of attention, which is an indicator of where to focus based on the combined audio-visual information. After focusing the attention region and eliminating the unnecessary areas, humans identify the sound-making object by repeatedly engaging in a recursive recognition process [25, 41]. By doing so, we can make more accurate predictions. This cognitive process enables humans to effectively utilize visual and auditory information, leading to more accurate and comprehensive understanding of the world around them.

In this paper, based on our aforementioned motivations, we propose a novel sound source localization framework that mimics the above-mentioned two cognitive psychological perspectives of humans (*i.e.*, potential of spatial cues in audio modality and the ability to recognize sound-making objects in a recursive manner). Our framework consists of two stages. First, we propose an audio-visual spatial integration network that integrates spatial knowledge from both audio-visual modalities to produce an integrated localization map. The aim of generating the integrated localization map is to contain rich spatial information about the sound-making objects. Second, we introduce a recursive attention network to mimic the human ability to recognize the objects in a recursive manner. Based on the integrated localization map, the unnecessary regions of the input image are eliminated and attentive input image is generated. Consequently, with the attentive input image, more precise localization of the sound-making object is possible in our recursive attention network. In addition, within the recursive attention network, we devise an audio-visual pair matching loss to guide the feature representation of each single modality (audio and visual) to resemble that of the attentive input image. By doing so, the features of both modalities can embed more precise spatial knowledge. Moreover, although the spatial knowledge of the audio modality contains valuable information, it may be relatively less precise than those of the visual modality. To address this issue, we introduce a spatial region alignment loss to guide the spatial representation of the audio modality to resemble that of the attentive input image. As a result, the feature representations of the audio modality are significantly enhanced, leading to a more accurate final localization map generation.

To sum up, the major contributions of this paper are summarized as follows:

- We introduce audio-visual spatial integration network that exploits the spatial knowledge of audio-visual modalities. In addition, we propose recursive attention network to refine the localization map in a recursive manner. To the best of our knowledge, it is the first work that considers the spatial knowledge of audio modality for sound source localization.

- To guide the feature representation of the single modality, we propose audio-visual pair matching loss. Also, to enhance spatial knowledge of the audio modality, we introduce spatial region alignment loss to resemble that of the attentive image.

- Comprehensive quantitative and qualitative experimental results on Flickr-SoundNet and VGG-Sound Source datasets validate the effectiveness of the proposed framework.

2 RELATED WORK

2.1 Sound Source Localization

Sound source localization aims to estimate the sound source location using visual scenes. It requires an effective combination of visual and audio data, and various algorithms have been developed over the years to optimize this multimodal integration for accurate localization [1, 9, 15, 22, 46, 49, 51, 60, 63].

One such approach is the use of attention mechanisms, which allow the network to selectively focus on relevant parts of the input data. In [49], Senocak et al. propose a sound localization network that incorporates an attention mechanism to focus on relevant parts of the visual modality and audio modality, resulting in more accurate sound source localization. In [9], Chen et al. introduce tri-maps to incorporate background mining techniques for identifying positive correlation region, no correlation region (background), and ignoring region to avoid uncertain areas in the visual scene. They utilize audio-visual pairs to create a tri-map highlighting positive/negative regions. In [60], Xuan et al. adopt the selective search [57] to utilize the proposal-based paradigm. Since the proposal region contains information of sound-making objects, finding the candidate objects firstly rather than the location of the sound can be superior. In [15], Fedorishin et al., assumed that most of the sound sources in visual scenes will be moving objects. Therefore, they adopt the optical flow algorithm in the visual modality to achieve more effective sound source localization.

In many studies on sound source localization task, the visual modality is usually considered to be a crucial modality (*e.g.*, selective search, optical flow, etc.). However, the audio modality is only utilized as an auxiliary role, primarily being used for similarity measurements (*e.g.*, cosine similarity) to generate the attention region of the visual modality. Thus, we claim that the existing methods tend to give weight to visual modality rather than audio modality. However, humans use both eyes and ears as important factors to judge situations in the natural environment. Therefore, we propose a sound source localization framework that uses audio modality as well as visual modality for acquire more abundant spatial knowledge of the audio-visual modalities.

Figure 2: Network configuration of the proposed sound localization framework. ⊕ and ⊗ indicate element-wise addition and element-wise multiplication, respectively. Note that final localization map M_{final} is generated by combining M_v, M_a, and M_v^{att}.

2.2 Recursive Deep Learning Framework in Computer Vision

Recursive deep learning frameworks [2, 24, 29, 35, 52, 54] have become increasingly popular for their ability to handle complex dependencies in sequential or structured data. Many works have adopted a recursive approach and applied it to the various computer vision tasks to improve their performance, such as object detection [11, 27, 36] and recognition [6, 7, 53], image super-resolution [28, 56, 58], visual tracking [17, 23], and semantic segmentation [45, 61, 62].

For example, in the object detection, a recursive model with the multistage framework is proposed [36]. This approach uses an EM-like group recursive learning technique to iteratively refine object proposals and improve the spatial configuration of object detection. Socher et al. [53] proposed a model that combines convolutional and recursive neural networks to detect object in the RGB-D images. In addition, for image super-resolution, Kim et al. [28] proposed the deeply-recursive convolutional network (DRCN) to improve the feature representation without adding more convolution parameters. To overcome the challenges of learning a DRCN, they introduce recursive supervision and skip connection.

In the visual object tracking, Gao et al. [17] utilized recursive least-squares estimation (LSE) for online learning. By integrating fully-connected layers with LSE and employing an enhanced mini-batch stochastic gradient descent algorithm, they enhanced the performance of visual object tracking. For semantic segmentation and depth estimation tasks, Zhang et al. [62] introduced the Joint Task-Recursive Learning (TRL) framework. It uses a Task-Attentional Module (TAM) to recursively refine the results.

For designing our method, we utilize the recursively refining idea to mimic the behavior of humans that repeatedly focus sound-making object for more accurate sound source localization. By recursively refining a model, the proposed method can improve the attention region of the sound-making object, by eliminating the unnecessary regions. As a result, our method achieves the outstanding performance over the state-of-the-art sound source localization works.

3 PROPOSED METHOD

3.1 Overall Architecture

The overall architecture of our sound source localization framework is depicted in Figure 2. Our framework consists of two stages: (1) audio-visual spatial integration network and (2) recursive attention network. First, in the audio-visual spatial integration network, input image set $I_v \in \mathbb{R}^{N \times W_v \times H_v \times 3}$ (N indicates the number of batch, W_v and H_v denote width and height of I_v, respectively) and the corresponding audio spectrogram set $I_a \in \mathbb{R}^{N \times W_a \times H_a \times 1}$ (W_a and H_a denote width and height of I_a, respectively) pass through each modal encoder (*i.e.,* visual encoder and audio encoder) to generate the spatial features $\mathbf{F}_v$ and $\mathbf{F}_a$, respectively. Then, image attentive localization map $\mathbf{M}_v$ and audio attentive localization map $\mathbf{M}_a$ are generated based on $\mathbf{F}_v$ and $\mathbf{F}_a$ through the attention module. $\mathbf{M}_v$ and $\mathbf{M}_a$ are attention maps that focus on the location of a sounding object based on the spatial features encoded in each modality. $\mathbf{M}_v$ and $\mathbf{M}_a$ are integrated to generate the audio-visual integrated localization map $\mathbf{M}_{av}$.

Second, the recursive attention network takes the resized $\mathbf{M}_{av}$ and multiplies it with I_v to generate attentive input image I_v^{att}. I_v^{att} is passed through the visual encoder to generate visual attention feature $\mathbf{F}_v^{att}$. Note that the weight parameters of the visual encoder in the audio-visual spatial integration network and recursive attention network are shared. With $\mathbf{F}_v^{att}$ and I_a, the localization map $\mathbf{M}_v^{att}$ is generated. More details are in the following subsections.

3.2 Audio-Visual Spatial Integration Network

When humans see a visual scene with their eyes while listening to a sounding object, they can acquire spatial cue information not only through vision but also through sound [18, 50]. We mimic the behaviors of humans for more accurate localizing sound source objects. To this end, we propose an audio-visual spatial integration network to exploit the spatial cues of both visual modality and audio modalities.

As shown in Figure 2, our audio-visual spatial integration network consists of two streams: (1) visual stream and (2) audio stream. In the visual stream, the visual spatial feature $\mathbf{F}_v \in \mathbb{R}^{N \times w \times h \times c}$ (w, h, and c are the width, height, and channel number) is mainly used to localize sound-making object. Specifically, the audio spatial feature $\mathbf{F}_a \in \mathbb{R}^{N \times w \times h \times c}$ is subject to a global average pooling (GAP) operation to generate $l_a \in \mathbb{R}^{N \times c}$. Then, $\mathbf{F}_v$ and l_a are compared using a similarity calculation in the attention module to generate $\mathbf{S}_v = \{S_{v_{ij}}\}_{i=1,...,h,j=1,...,w} \in \mathbb{R}^{N \times w \times h}$, which is measured as:

$$S_{v_{ij}} = \frac{Sim(\mathbf{F}_{v_{ij}}, l_a)}{\sum_{i=1}^{h} \sum_{j=1}^{w} Sim(\mathbf{F}_{v_{ij}}, l_a)}, \; Sim(\mathbf{F}_{v_{ij}}, l_a) = \frac{\mathbf{F}_{v_{ij}} \cdot l_a}{||\mathbf{F}_{v_{ij}}|| \, ||l_a||}. \quad (1)$$

Then, $\mathbf{S}_v$ is normalized by the softmax to generate the image attentive localization map $\mathbf{M}_v \in \mathbb{R}^{N \times w \times h}$.

In the audio stream, the audio spatial feature $\mathbf{F}_a \in \mathbb{R}^{N \times w \times h \times c}$ is mainly used to localize sound-making objects. However, while the audio modality contains the spatial cues for localizing objects, it generally lacks the levels of detail compared to the visual modality. For example, if we hear an object sound with our eyes closed, we can roughly estimate its location, but it is typically less precise than if we were to open our eyes and visually locate the object. Thus, we transfer the spatial knowledge of $\mathbf{F}_v$ to $\mathbf{F}_a$ while maintaining the area that $\mathbf{F}_a$ focuses on by generating $\mathbf{F}_{av}$. $\mathbf{F}_{av}$ is obtained as:

$$\mathbf{F}_{av} = \mathbf{F}_v \circ \bar{\mathbf{F}}_a, \quad (2)$$

where $\bar{\mathbf{F}}_a$ denotes the normalized version of $\mathbf{F}_a$ (min-max normalization is conducted with the value between 0 and 1), and $\circ$ indicates the element-wise multiplication.

Next, similar to Eq. (1), the $\mathbf{S}_{av} = \{S_{av_{ij}}\}_{i=1,...,h,j=1,...,w} \in \mathbb{R}^{N \times w \times h}$ is obtained as:

$$S_{av_{ij}} = \frac{Sim(\mathbf{F}_{av_{ij}}, l_a)}{\sum_{i=1}^{h} \sum_{j=1}^{w} Sim(\mathbf{F}_{av_{ij}}, l_a)}, \; Sim(\mathbf{F}_{av_{ij}}, l_a) = \frac{\mathbf{F}_{av_{ij}} \cdot l_a}{||\mathbf{F}_{av_{ij}}|| \, ||l_a||}. \quad (3)$$

$\mathbf{S}_{av}$ is also normalized by the softmax to make the audio attentive localization map $\mathbf{M}_a$. The two localization maps, $\mathbf{M}_v$ and $\mathbf{M}_a$, generated by the proposed audio-visual spatial integration network, provide information about the spatial regions in each modality that are being focused on to localize the sounding objects. Therefore, we integrate the knowledge of the audio-visual modalities to make

$\mathbf{M}_{av} \in \mathbb{R}^{N \times w \times h}$, which can be obtained as follows:

$$\mathbf{M}_{av} = \frac{\mathbf{M}_a + \mathbf{M}_v}{2}. \quad (4)$$

Since $\mathbf{M}_{av}$ contains the spatial information of both audio and visual modalities, it provides a more precise localization map compared to using either modality alone. By combining the spatial cues from both modalities, the proposed method is able to effectively mitigate the limitations of each modality and produce a more accurate localization result.

3.3 Recursive Attention Network

Given the visual and audio modal information, humans can integrate attention regions across different modalities, such as visual and auditory information, to concentrate on a specific region [13, 14, 55, 59]. It is called multisensory integration. By doing so, humans can concentrate their attention on specific regions of the environment that correspond to the presented sensory information. This allows them to more effectively process and respond to stimuli from both modalities [16, 34, 43].

Therefore, we build the recursive attention network to mimic the above-mentioned behaviors of humans. The recursive attention network utilizes the audio-visual integrated localization map $\mathbf{M}_{av}$ derived from the audio-visual spatial integration network to produce an attentive input image I_v^{att}. Specifically, $\mathbf{M}_{av} \in \mathbb{R}^{N \times w \times h}$ is resized to be $\mathbf{M}_{av}^r \in \mathbb{R}^{N \times W_v \times H_v}$. To the next, $\mathbf{M}_{av}^r$ and I_v are conducted element-wise multiplication to focus the attention region of the image, i.e., I_v^{att}. We feed this attentive input image I_v^{att} into the visual encoder to encode visual attention feature F_v^{att}. The attention module calculates the similarity between F_v^{att} and l_a to generate $\mathbf{S}_v^{att} = \{S_{v_{ij}}^{att}\}_{i=1,...,h,j=1,...,w} \in \mathbb{R}^{N \times w \times h}$. Note that $\mathbf{S}_v^{att}$ is calculated similarly to Eq. (1) and Eq. (3). Also, $\mathbf{S}_v^{att}$ is normalized by the softmax to make the localization map $\mathbf{M}_v^{att}$.

Finally, we combine the $\mathbf{M}_v$, $\mathbf{M}_a$, and $\mathbf{M}_v^{att}$ to generate the final localization map $\mathbf{M}_{final}$, which can be represented as:

$$\mathbf{M}_{final} = w_1 \mathbf{M}_v + w_2 \mathbf{M}_a + w_3 \mathbf{M}_v^{att}, \quad (5)$$

where w_1, w_2, and w_3 are the hyper-parameters that indicate the importance of each modality in contributing to the $\mathbf{M}_{final}$. $\mathbf{M}_a$ and $\mathbf{M}_v$ contain the spatial cues of each modality (i.e., audio and visual modalities), and $\mathbf{M}_v^{att}$ contains the spatial cues of the more attentive region from the audio-visual modalities. Therefore, by combining $\mathbf{M}_a$ and $\mathbf{M}_v$, the spatial cues from both modalities can be obtained. Additionally, by combining $\mathbf{M}_v^{att}$, more interested regions can be obtained. The recursive combination of the localization maps can utilize abundant spatial information, leading to more accurate sound source localization.

3.4 Audio-Visual Pair Matching Loss

Humans can make more accurate predictions by removing unnecessary areas by focusing attention through their eyes and ears. Similarly, in our method, the attentive input image I_v^{att} concentrates the area that is generated by the audio-visual modality in the audio-visual spatial integration network. This enables us to localize the sounding objects more accurately. This is similar to the fact that the two-stage detectors [19, 31, 32, 38, 48], which first extract the region of interest (ROI) for more accurate object detection,

generally outperform the one-stage object detectors [39, 40, 47]. Therefore, compared to $\mathbf{M}_v$, $\mathbf{M}_a$ and $\mathbf{M}_v^{att}$, $\mathbf{M}_v^{att}$ usually contains more meaningful regions than $\mathbf{M}_v$ and $\mathbf{M}_a$. As a result, we propose an audio-visual pair matching loss to guide the feature representations of the visual modality $\mathbf{F}_v$ and the audio modality $\mathbf{F}_a$ to be similar to that of the visual attention feature $\mathbf{F}_v^{att}$.

To this end, we first conduct global average pooling (GAP) of $\mathbf{F}_v^{att}$, $\mathbf{F}_v$, and $\mathbf{F}_a$ and normalize them to generate l_v^{att}, l_v, and l_a, respectively. Next, we adopt the triplet loss [21] for the audio-visual pair matching loss $\mathcal{L}_{avpm}$, which can be represented as:

$$T(l_{v_i}^{att}, l_{a_i}, l_{a_j}) = D(l_{v_i}^{att}, l_{a_i}) + max(\delta - D(l_{v_i}^{att}, l_{a_j}), 0),$$

$$T(l_{v_i}^{att}, l_{v_i}, l_{v_j}) = D(l_{v_i}^{att}, l_{v_i}) + max(\delta - D(l_{v_i}^{att}, l_{v_j}), 0),$$

$$\mathcal{L}_{avpm} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1(j\neq i)}^{N} T(l_{v_i}^{att}, l_{a_i}, l_{a_j}) + T(l_{v_i}^{att}, l_{v_i}, l_{v_j}),$$
$$(6)$$

where $D(\alpha, \beta) = ||(\alpha - \beta)/\tau||_2^2$ denotes the L2 norm to calculate the distance between two features with temperature parameter τ, $l_{v_i}^{att}$, l_{a_i}, and l_{a_j} are the features of anchor, positive, and negative samples, respectively, and δ is the margin.

The aim of $T(l_{v_i}^{att}, l_{a_i}, l_{a_j})$ and $T(l_{v_i}^{att}, l_{v_i}, l_{v_j})$ is to make the anchor ($l_{v_i}^{att}$) and the positive pair (l_{a_i}, l_{v_i}) similar while pushing the negative pair (l_{a_j}, l_{v_j}) apart. By doing so, $\mathcal{L}_{avpm}$ can guide the feature representation of $\mathbf{F}_v$ and $\mathbf{F}_a$ to be similar that of $\mathbf{F}_v^{att}$. As a result, the feature representation of $\mathbf{F}_v$ and $\mathbf{F}_a$ improve the performance of sound source localization (please see Section 4.5).

3.5 Spatial Region Alignment Loss

Although we can infer spatial information using sound, it is relatively less accurate than visual information. Therefore, we introduce a spatial region alignment loss in order to guide the spatial regions that audio feature $\mathbf{F}_a$ focus on to be similar to that of the $\mathbf{F}_v^{att}$. To this end, we add all c channels of $\mathbf{F}_a$ and $\mathbf{F}_v^{att}$ to normalize them to generate $\hat{\mathbf{F}}_a \in \mathbb{R}^{N \times w \times h}$ and $\hat{\mathbf{F}}_v^{att} \in \mathbb{R}^{N \times w \times h}$. After that, they are flattened to conduct softmax function to generate $\hat{\mathbf{G}}_a \in \mathbb{R}^{N \times wh}$ and $\hat{\mathbf{G}}_v^{att} \in \mathbb{R}^{N \times wh}$, respectively. Based on the $\hat{\mathbf{G}}_v^{att}$ and $\hat{\mathbf{G}}_a$, the spatial region alignment loss $\mathcal{L}_{sra}$ is represented as follows:

$$\mathcal{L}_{sra} = \frac{1}{N} \sum_{i=1}^{N} \underbrace{D_{KL}\left(\hat{\mathbf{G}}_{v_i}^{att} \| \hat{\mathbf{G}}_{a_i}\right)}_{\text{audio to attentive visual}},$$
$$(7)$$

where $D_{KL}(\cdot)$ indicates the Kullback-Leibler (KL) divergence. $\mathcal{L}_{sra}$ makes the spatial representation of $\mathbf{F}_a$ to be similar to that of $\mathbf{F}_v^{att}$ in the training phase. By doing so, when generating $\mathbf{F}_a$, our method can effectively estimate the spatial regions by hearing sounds.

3.6 Total Loss Function

To train our method, the total loss function is composed as follows:

$$\mathcal{L}_{Total} = \mathcal{L}_{SSL} + \lambda_1 \mathcal{L}_{avpm} + \lambda_2 \mathcal{L}_{sra}, \quad (8)$$

where $\mathcal{L}_{SSL}$ is the unsupervised loss function of the sound source localization that tries to impose the audio-visual feature pairs are close to each other, following [9], λ_1 and λ_2 denote the balancing parameter. Through $\mathcal{L}_{Total}$, our method can perform effective sound source localization by leveraging the spatial knowledge of

Table 1: Experimental results on Flickr test set when the training sets are Flickr10k and Flickr144k, respectively.

Method	Training Set	cIoU$_{0.5}\uparrow$	AUC$\uparrow$
Attention [49] (CVPR'18)		0.436	0.449
DMC [22] (CVPR'19)		0.414	0.450
CoarseToFine [46] (ECCV'20)		0.522	0.496
AVObject [1] (ECCV'20)		0.546	0.504
LVS [9] (CVPR'21)	Flickr10k	0.582	0.525
Zhou et al. [63] (WACV'23)		0.631	0.551
Shi et al. [51] (WACV'22)		0.734	0.576
SSPL [60] (CVPR'22)		0.743	0.587
HTF [15] (WACV'23)		0.860	0.634
Proposed Method		**0.876**	**0.641**
Attention [49] (CVPR'18)		0.660	0.558
DMC [22] (CVPR'19)		0.671	0.568
LVS [9] (CVPR'21)	Flickr144k	0.699	0.573
SSPL [60] (CVPR'22)		0.759	0.610
HTF [15] (WACV'23)		0.865	0.639
Proposed Method		**0.881**	**0.652**

the audio-visual modality and combining all attention maps in a recursive manner.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

Flickr-SoundNet. Flickr-SoundNet [4] consists more than 2 million videos from Flickr. In the training phase, to enable direct comparison with prior research, we train our models with two random subsets of 10k and 144k image-audio pairs. In the inference phase, we use Flickr-SoundNet test set. It contains 250 annotated pairs with labeled bounding box, manually annotated by the annotators [9, 49].

VGG-Sound Source. VGG-Sound dataset [10] consists of 200k video clips from 300 different sound categories. Following [15], we use a training dataset with 10k and 144k image-audio pairs. For evaluation, we use VGG-Sound Source (VGG-SS) dataset [9] with 5,000 annotated image-audio pairs from 220 classes. Compared with Flickr-SoundNet, which contains 50 audio categories, the VGG-SS dataset set offers a larger number of sound sources. Therefore, it contains more challenging scenario for sound source localization.

Evaluation Metrics. To compare our method with the existing methods, we adopt consensus Intersection over Union (cIoU) [49] and Area Under Curve (AUC) as evaluation metrics, which are the widely adopted metrics for sound source localization task [9, 15, 49]. For calculating cIoU, the IoU threshold is fixed to be 0.5 (*i.e.,* cIoU$_{0.5}$), following [9, 15, 49]. Note that, in our experiments, we additionally introduce a mcIoU metric to measure the performance by varying the IoU threshold to 0.5:0.05:0.95. More details are in Section 4.6.

4.2 Implementation Details

For both datasets, we resize the input image for the visual modality to be $W_v = 224$, $H_v = 224$. It is extracted from the middle frame of the 3-seconds video clips. For audio modality input, we resample

Table 2: Experimental results on VGG-SS test set when the training sets are VGG-Sound10k and VGG-Sound144k, respectively.

Method	Training Set	cIoU$_{0.5}$↑	AUC↑
LVS [9] (CVPR'21)		0.303	0.364
SSPL [60] (CVPR'22)	VGG-Sound10k	0.314	0.369
Zhou et al. [63] (WACV'23)		0.350	0.376
HTF [15] (WACV'23)		0.393	0.398
Proposed Method		**0.403**	**0.403**
LVS [9] (CVPR'21)		0.344	0.382
SSPL [60] (CVPR'22)	VGG-Sound144k	0.339	0.380
HTF [15] (WACV'23)		0.394	0.400
Proposed Method		**0.406**	**0.405**

the 3-seconds raw audio signal to 16kHz and transform it into a log-scale spectrogram, yielding a final shape $W_a = 257$ and $H_a = 276$. At this time, to enable a direct comparison with visual modal features, we resize $\mathbf{F}_a$ to be $7 \times 7 \times 512$ ($w = 7$, $h = 7$, and $c = 512$) using bilinear interpolation.

Following [9], we employ ResNet-18 [20] for both the visual and audio feature backbones to construct our baseline. Since the number of sound spectrogram channel is 1. we modify the input channel of ResNet-18 [20] conv1 from 3 to 1. We use the ImageNet [12] pretrained for the visual encoder. When our baseline is HTF [15], we additionally consider the optical flow [15] for the attention module (more details are in Section 4.6). Our sound source localization framework is trained using the Adam optimizer [33] with a learning rate of 10^{-4} and a batch size of 128. Following [15], we train our model for 100 epochs for Flickr and VGG-Sound datasets. We use 4 synchronized RTX 3090 GPUs. The weights for M_{final} in Eq. (5) are set as $w_1 = w_2 = w_3 = 1$. Also, we use $\lambda_1 = 1$, $\lambda_2 = 10$, $\tau = 0.03$, and $\delta = 25$ for our proposed loss functions ($\mathcal{L}_{avpm}$ and $\mathcal{L}_{sra}$).

4.3 Performance Comparison

We conduct the experiments to compare the effectiveness of our proposed method with the state-of-the-art sound source localization works [1, 9, 15, 22, 46, 49, 51, 60, 63]. Table 1 shows the performance of our method with the existing methods on Flickr-SoundNet. When the training set is Flickr10k, our method achieves 0.876 and 0.641 for cIoU$_{0.5}$ and AUC, respectively. Specifically, when compared to the HTF [15] which shows the highest performance among the existing methods, our method is 1.6% higher for cIoU and 0.7% higher for AUC metrices. Similar tendency is observed when the training set is Flickr144k training set. The experimental results on Table 1 demonstrate that our approach that considering the spatial knowledge of the audio-visual modalities and recursively refining the localization map leads to better localization performance.

The experimental results on the VGG-Sound dataset are shown in Table 2. For the VGG-Sound Source test set, our method achieved improvements of 1.0% cIoU and 0.5% AUC in the VGG-Sound10k dataset, and 1.2% cIoU and 0.5% AUC in the VGG-Sound144k dataset over the HTF [15]. The results validate that our method outperforms existing methods and achieves a state-of-the-art performance over the existing sound source localization works.

Table 3: Effect of the proposed audio-visual pair matching loss $\mathcal{L}_{avpm}$ and spatial region alignment loss $\mathcal{L}_{sra}$ on Flickr test set, where models are trained on the Flickr144k.

Method	$\mathcal{L}_{avpm}$	$\mathcal{L}_{sra}$	cIoU$_{0.5}$↑	AUC↑
Baseline	-	-	0.865	0.642
Proposed Method	✓	-	0.876	0.643
	-	✓	0.871	0.648
	✓	✓	**0.881**	**0.652**

Table 4: Experimental results on Flickr test set according to the hyper-parameters w_1, w_2, and w_3 for $\mathbf{M}_{final}$ in Eq. (5), where models are trained on the Flickr144k.

w_1 ($\mathbf{M}_v$)	w_2 ($\mathbf{M}_a$)	w_3 ($\mathbf{M}_v^{att}$)	cIoU$_{0.5}$↑	AUC↑
1	1	4	0.866	0.645
1	1	2	0.875	0.649
1	**1**	**1**	**0.881**	**0.652**
1	2	1	0.871	0.639
2	1	1	0.876	0.650
2	2	1	0.871	0.643

4.4 Ablation Study

We conduct various ablation studies to investigate (1) effect of the proposed losses (i.e., $\mathcal{L}_{avpm}$ and $\mathcal{L}_{sra}$), and (2) variation of the hyper-parameter w_1, w_2, w_3 for M_{final}. All the ablation studies are conducted using Flickr144k training set and Flickr test set.

Effect of the Proposed Losses. We measure the performance by changing two types of the proposed losses $\mathcal{L}_{avpm}$ and $\mathcal{L}_{sra}$. The results are shown in Table 3. When each loss is considered, our method shows the improved performance agains the baseline in which those losses are not considered. When all the proposed losses are taken into account, we show the highest performance. By incorporating the proposed losses in the training phase, our method is able to learn more robust and discriminative features that are better suited for the sound source localization task.

Variation of w_1, w_2, w_3 We conduct additional ablation study to investigate the effect of our method to the parameters w_1, w_2, and w_3 as described in Section 3.3. The results of Table 4 show that the optimal results are obtained when w_1, w_2, and w_3 are set to 1. However, it's important to note that our method still outperforms existing methods even with different values for these parameters. These results suggest that the model is robust to parameter changes, but there may be an optimal combination that maximizes its effectiveness. In our future work, we are planning to build a framework that considers weight of the localization maps.

4.5 Visualization Results

We compare our method with the current state-of-the-art approach, HTF [15], by visualizing their sound source localization results on the Flickr-SoundNet and VGG-SS test set. The results are shown in Figure 3. Through the visualization results, our method can accurately localize the sound-making objects (GT annotation indicates

(a) Flickr-Soudnet Test Set

(b) VGG-SS Test Set

Figure 3: Visualization results for both (a) Flickr-SoundNet test set and (b) VGG-SS test set. Each result is obtained from models trained on the Flickr144k and VGG-Sound144k training sets. For the Flickr-SoundNet test set, annotation is created by combining bounding boxes from different annotators.

Figure 4: Visualization results of M_v, M_a, M_v^{att}, and the final localization map M_{final} of our method on (a) Flickr-SoundNet test set and (b) VGG-SS test set.

the region of the sound-making objects). Since our method considers the spatial information of both audio and visual modalities and recursively updates the localization map, more precise attention maps are obtained.

Furthermore, Figure 4 shows visualization results of the various localization maps M_v, M_a, M_v^{att}, and M_{final} of our method. The visualization results show that considering audio-visual localization map and recursively updating them contributes to M_{final} for concentrating on a more accurate location. By doing so, our method shows the improving performance.

4.6 Discussions

Cross Dataset Evaluation. To show the effectiveness of our method on the cross dataset environment, we train our model on the VGG-Sound training set and evaluate it on the Flickr-SoundNet test set. This cross-dataset evaluation enables us to assess the ability of model to generalize and to check to new and diverse data sources. The results of Table 5 show the results when the training sets are VGG-Sound10k and VGG-Sound144k, respectively. The results show that the performances of our method are significantly higher than the existing methods. As a result, our model demonstrates

Table 5: Experimental results on the cross-dataset evaluation. Note that we trained the model on the VGG-Sound10k and VGG-Sound144k and evaluated on the Flickr test set. '∗' denotes our faithful reproduction of the method.

Method	Training Set	cIoU$_{0.5}$↑	AUC↑
LVS [9] (CVPR'21)	VGG-Sound10k	0.618	0.536
SSPL [60] (CVPR'22)		0.763	0.591
Zhou et al. [63] (WACV'23)		0.775	0.596
HTF [15]* (WACV'23)		0.842	0.628
Proposed Method		**0.875**	**0.640**
LVS [9] (CVPR'21)	VGG-Sound144k	0.719	0.582
SSPL [60] (CVPR'22)		0.767	0.605
HTF [15] (WACV'23)		0.848	0.640
Proposed Method		**0.881**	**0.651**

Table 6: Experimental results on Flickr test set with respect to various sound source localization frameworks.

Method	Training Set	cIoU$_{0.5}$↑	AUC↑
LVS [9] (CVPR'21)	Flickr144k	0.699	0.573
Proposed Method (LVS)		**0.718**	**0.577**
HTF [15] (WACV'23)	Flickr144k	0.865	0.639
Proposed Method (HTF)		**0.881**	**0.652**

Table 7: Experimental results on Flickr test set using mcIoU metric.

Method	Training Set	cIoU$_{0.5}$↑	mcIoU↑
LVS [9] (CVPR'21)	Flickr144k	0.699	0.231
HTF [15] (WACV'23)		0.865	0.363
Proposed Method		**0.881**	**0.381**

the potential to demonstrate sufficient generalization capabilities essential for real-world applications involving diverse data sources.

Generalization Ability of Our Method. In this subsection, we conduct experiments to see the generalization ability of our method by varying the baseline. To this end, we adopt the two baselines: LVS [9] and HTF [15]. The results are shown in Table 6. All the methods are trained with Flickr144k and evaluated on Flickr test set. As shown in the table, when our baseline is LVS [9], we achieve 1.19% cIoU and 0.4% AUC improvement compared to the original LVS. The results on HTF [15] also show a similar tendency. The results indicate that our method has broader applicability and can be integrated with various sound source localization frameworks.

Evaluation on the Proposed mcIoU Metric. Note that consensus intersection over union (cIoU) [49] metric has been widely used for comparing sound source localization methods. In this subsection, we additionally introduce a new metric called mcIoU (mean cIoU) to investigate the performance while varying the IoU threshold. For calculating mcIoU metric, we take the average cIoU across IoU threshold 0.5:0.05:0.95. The results are shown in Table 7. Compared to the existing methods [9, 15], the performances of our method

Table 8: The comparisons of training time, inference time, and the number of parameters.

Method	Training (s) (*per iter*)	Inference (s) (*per image*)	#params
HTF [15] (WACV'23)	0.385	0.039	33.85M
Proposed Method	0.398	0.042	34.50M

consistently improved. The results demonstrate that the effectiveness of our method considers spatial cues of audio modality and performs sound source localization in a recursive manner.

Computational Costs. In this section, we compare training time, inference time, and the number of parameters. It is shown in Table 8. We compare our method with HTF [15] which shows the highest performance among the existing methods. Since our method adopts the recursive method, the training time, inference time, and the number of parameters of our method are slightly increased (3.38%, 7.69% and 1.92% for training, inference, and parameters, respectively). Nevertheless, we claim that the increased times of training and inference time and the number of parameters are marginal compared to the HTF [15].

5 CONCLUSION

In this paper, we propose a novel sound source localization framework that considers the inherent spatial information of the audio modality as well as the visual modality for exploiting more abundant spatial knowledge. To this end, our framework consists of two stages: (1) audio-visual spatial integration network and (2) recursive attention network. The audio-visual spatial integration network is designed to incorporate the spatial information of the audio-visual modalities. By focusing on the attention region generated by the audio-visual spatial integration network, the recursive attention network aims to perform more precise sound source localization. At this time, we devise audio-visual pair matching loss and spatial region alignment loss to effectively guide the features from the audio-visual modalities to resemble the features of the attentive information. We believe that our approach, integrating spatial knowledge of audio-visual modalities and recursively refining the results leads to more improved accuracy and it can be utilized in various practical applications.

ACKNOWLEDGMENTS

REFERENCES

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. 2020. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision (ECCV)*. Springer, 208–224.

[2] Ahmad Al-Sallab, Ramy Baly, Hazem Hajj, Khaled Bashir Shaban, Wassim El-Hajj, and Gilbert Badaro. 2017. Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing* 16, 4 (2017), 1–20.

[3] Sebastià V Amengual Garí, Winfried Lachenmayr, and Eckard Mommertz. 2017. Spatial analysis and auralization of room acoustics using a tetrahedral microphone. *The Journal of the Acoustical Society of America* 141, 4 (2017), EL369–EL374.

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

[5] Robert S Bolia, William R D'Angelo, and Richard L McKinley. 1999. Aurally aided visual search in three-dimensional space. *Human factors* 41, 4 (1999), 664–669.

[6] Hieu Minh Bui, Margaret Lech, Eva Cheng, Katrina Neville, and Ian S Burnett. 2016. Object recognition using deep convolutional features transformed by a recursive network structure. *IEEE Access* 4 (2016), 10059–10066.

[7] Hieu Minh Bui, Margaret Lech, Eva Cheng, Katrina Neville, and Ian S Burnett. 2016. Using grayscale images for object recognition with convolutional-recursive neural network. In *International Conference on Communications and Electronics (ICCE)*. 321–325.

[8] Shlomo E Chazan, Hodaya Hammer, Gershon Hazan, Jacob Goldberger, and Sharon Gannot. 2019. Multi-microphone speaker separation based on deep DOA estimation. In *European Signal Processing Conference (EUSIPCO)*. 1–5.

[9] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2021. Localizing visual sounds the hard way. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16867–16876.

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

[11] Zhe Chen, Jing Zhang, and Dacheng Tao. 2021. Recursive context routing for object detection. *International Journal of Computer Vision* 129, 1 (2021), 142–160.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[13] Jon Driver and Charles Spence. 1998. Cross–modal links in spatial attention. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 353, 1373 (1998), 1319–1331.

[14] Martin Eimer and Erich Schröger. 1998. ERP effects of intermodal attention and cross-modal links in spatial attention. *Psychophysiology* 35, 3 (1998), 313–327.

[15] Dennis Fedorishin, Deen Dayal Mohan, Bhavin Jawade, Srirangaraj Setlur, and Venu Govindaraju. 2023. Hear The Flow: Optical Flow-Based Self-Supervised Visual Sound Source Localization. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2278–2287.

[16] John J Foxe, Istvan A Morocz, Micah M Murray, Beth A Higgins, Daniel C Javitt, and Charles E Schroeder. 2000. Multisensory auditory–somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Cognitive Brain Research* 10, 1-2 (2000), 77–83.

[17] Jin Gao, Weiming Hu, and Yan Lu. 2020. Recursive least-squares estimator-aided online learning for visual tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7386–7395.

[18] William W Gaver. 1993. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology* 5, 1 (1993), 1–29.

[19] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. 2019. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7036–7045.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[21] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition (SIMBAD)*. Springer, 84–92.

[22] Di Hu, Feiping Nie, and Xuelong Li. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9248–9257.

[23] Zhiyong Huang, Yuanlong Yu, and Miaoxing Xu. 2019. Bidirectional tracking scheme for visual object tracking based on recursive orthogonal least squares. *IEEE Access* 7 (2019), 159199–159213.

[24] Ozan Irsoy and Claire Cardie. 2014. Deep recursive neural networks for compositionality in language. *Advances in Neural Information Processing Systems (NeurIPS)*.

[25] Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience* 2, 3 (2001), 194–203.

[26] Bill Jones and Boris Kabanoff. 1975. Eye movements in auditory space perception. *Perception & Psychophysics* 17 (1975), 241–245.

[27] Yun Yi Ke and Takahiro Tsubono. 2022. Recursive contour-saliency blending network for accurate salient object detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2940–2950.

[28] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Deeply-recursive convolutional network for image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1637–1645.

[29] Jung Uk Kim, Hak Gu Kim, and Yong Man Ro. 2017. Iterative deep convolutional encoder-decoder network for medical image segmentation. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 685–688.

[30] Jung Uk Kim and Seong Tae Kim. 2023. Towards Robust Audio-Based Vehicle Detection Via Importance-Aware Audio-Visual Learning. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[31] Jung Uk Kim, Sungjune Park, and Yong Man Ro. 2021. Robust small-scale pedestrian detection with cued recall via memory learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 3050–3059.

[32] Jung Uk Kim, Sungjune Park, and Yong Man Ro. 2021. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 3 (2021), 1510–1523.

[33] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[34] Sangmin Lee, Sungjune Park, and Yong Man Ro. 2022. Audio-Visual Mismatch-Aware Video Retrieval via Association and Adjustment. In *European Conference on Computer Vision (ECCV)*. Springer, 497–514.

[35] Changliang Li, Bo Xu, Gaowei Wu, Saike He, Guanhua Tian, and Hongwei Hao. 2014. Recursive deep learning for sentiment analysis over social data. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. 180–185.

[36] Jianan Li, Xiaodan Liang, Jianshu Li, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. 2017. Multistage object detection with group recursive learning. *IEEE Transactions on Multimedia* 20, 7 (2017), 1645–1655.

[37] Xiaofei Li, Laurent Girin, Fabien Badeig, and Radu Horaud. 2016. Reverberant sound localization with a robot head based on direct-path relative transfer function. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2819–2826.

[38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2117–2125.

[39] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2980–2988.

[40] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*. Springer, 21–37.

[41] Ren C Luo and Michael G Kay. 1989. Multisensor integration and fusion in intelligent systems. *IEEE Transactions on Systems, Man, and Cybernetics* 19, 5 (1989), 901–931.

[42] Piotr Majdak, Matthew J Goupell, and Bernhard Laback. 2010. 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Attention, Perception, & Psychophysics* 72, 2 (2010), 454–469.

[43] M-Marchsel Mesulam. 1981. A cortical network for directed attention and unilateral neglect. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 10, 4 (1981), 309–325.

[44] David R Perrott, John Cisneros, Richard L Mckinley, and William R D'Angelo. 1996. Aurally aided visual search under virtual and free-field listening conditions. *Human Factors* 38, 4 (1996), 702–715.

[45] Pedro Pinheiro and Ronan Collobert. 2014. Recurrent convolutional neural networks for scene labeling. In *International Conference on Machine Learning (ICML)*. 82–90.

[46] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. 2020. Multiple sound sources localization from coarse to fine. In *European Conference on Computer Vision (ECCV)*. Springer, 292–308.

[47] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 779–788.

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)*.

[49] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. 2018. Learning to localize sound source in visual scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4358–4366.

[50] BR Shelton and CL Searle. 1980. The influence of vision on the absolute identification of sound-source position. *Perception & Psychophysics* 28 (1980), 589–596.

[51] Jiayin Shi and Chao Ma. 2022. Unsupervised Sounding Object Localization with Bottom-Up and Top-Down Attention. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1737–1746.

[52] Richard Socher. 2014. *Recursive deep learning for natural language processing and computer vision.* Stanford University.

[53] Richard Socher, Brody Huval, Bharath Bath, Christopher D Manning, and Andrew Ng. 2012. Convolutional-recursive deep learning for 3d object classification. *Advances in Neural Information Processing Systems (NeurIPS).*

[54] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *International Conference on Machine Learning (ICML).* 129–136.

[55] Barry E Stein and Terrence R Stanford. 2008. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience* 9, 4 (2008), 255–266.

[56] Ying Tai, Jian Yang, and Xiaoming Liu. 2017. Image super-resolution via deep recursive residual network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 3147–3155.

[57] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *International Journal of Computer Vision* 104 (2013), 154–171.

[58] Wei Wei, Jiangtao Nie, Yong Li, Lei Zhang, and Yanning Zhang. 2020. Deep recursive network for hyperspectral image super-resolution. *IEEE Transactions on Computational Imaging* 6 (2020), 1233–1244.

[59] Eric Zhongcong Xu, Zeyang Song, Satoshi Tsutsui, Chao Feng, Mang Ye, and Mike Zheng Shou. 2022. Ava-avd: Audio-visual speaker diarization in the wild. In *ACM International Conference on Multimedia (ACM MM).* 3838–3847.

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

[61] Yue Zhang, Xianrui Li, Mingquan Lin, Bernard Chiu, and Mingbo Zhao. 2020. Deep-recursive residual network for image semantic segmentation. *Neural Computing and Applications* 32 (2020), 12935–12947.

[62] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. 2018. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV).* 235–251.

[63] Xinchi Zhou, Dongzhan Zhou, Di Hu, Hang Zhou, and Wanli Ouyang. 2023. Exploiting Visual Context Semantics for Sound Source Localization. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).* 5199–5208.

Supplementary Material

This manuscript provides the additional results of the proposed method. Section A shows our additional implementation details, Section B indicates the additional experimental results to show the effectiveness of the cross dataset evaluation and proposed two modules (spatial integration and recursive attention), and Section C shows the additional visualization results, respectively. Please note that [PXX] indicates the reference in the main paper.

A   ADDITIONAL IMPLEMENTATION DETAILS

As mentioned in the main paper, we adopt the ResNet-18[P20] as the audio encoder. However, to match the input size, we adjusted the input channel of the first convolutional network to 1 and the output channel to 64, employing a kernel size of 7, stride 2, and padding 3. In addition, we used the Adam optimizer for training, with parameters $(\beta_1, \beta_2) = (0.9, 0.999)$, and a learning rate of 0.001. These are standard values for Adam and provided stable training dynamics in our experiments.

B   ADDITIONAL EXPERIMENTS

Cross Dataset Evaluation (Train: Flickr, Test: VGG-SS). To further validate the generalization ability of our method, we conducted an experiment where we trained our model on the Flickr dataset [P4] and evaluated it on the VGG-SS dataset [P9]. This is in contrast to our main paper, where we trained on the VGG-SS dataset and evaluated it on the Flickr dataset. As depicted in Table 1, the results demonstrate that our model still outperforms the HTF [P15], which shows state-of-the-art performance. These results confirm the ability of our method to generalize to new and diverse sources, further supporting the robustness of our approach.

Performance of proposed two modules (spatial integration and recursive attention). We conduct the additional ablation study on Flicker144k [P4] and VGG-Sound144k [P9] datasets when the proposed two modules (spatial integration and recursive attention) are considered or not. We compare the performances with two state-of-the-art methods, SSPL [P60] and HTF [P15]. The results are shown in Table 2. When the spatial integration network and the recursive attention network are considered individually, our method already exhibits improved performances compared to the existing methods. It demonstrates that each module contributes to the sound source localization task. Moreover, when the two modules are considered together, our method obtains further performance improvement.

C   ADDITIONAL VISUALIZATIONS

Sound Source Localization Result Comparisons of HTF [P15] and Ours. We additionally present a comparison between our method and HTF [P15] in sound source localization results. We used the codes provided by the authors to obtain the HTF results, which are presented in Figure 1. The fourth row of Figure 1(a) and Figure 1(b) show that the HTF model failed to localize the sound-making regions while our method was able to focus on them. The results demonstrate that the proposed method, which considers the spatial information of the audio-visual modalities and improves the

Table 1: Experimental results on the cross dataset evaluation. Note that we trained the model on the Flickr10k [P4] and Flickr144k [P4] and evaluated on the VGG-SS test set [P9]. '∗' denotes our faithful reproduction of the method.

Method	Training Set	cIoU$_{0.5}$↑	AUC↑
HTF [P15]∗ (WACV'23)	**Flickr10k**	0.384	0.396
Proposed Method		**0.396**	**0.400**
HTF [P15]∗ (WACV'23)	**Flickr144k**	0.385	0.396
Proposed Method		**0.399**	**0.401**

Table 2: Experiments result for performance comparison of two modules. Note that 'SI' represents Spatial Integration, and 'RA' represents Recursive Attention.

Method	Training Set	cIoU$_{0.5}$↑	AUC↑
SSPL [P60] (CVPR'22)		0.759	0.610
HTF [P15] (WACV'23)		0.865	0.639
Ours (w/ SI, w/o RA)	**Flickr144k**	**0.870**	**0.645**
Ours (w/o SI, w/ RA)		**0.870**	**0.649**
Ours (w/ SI, w/ RA)		**0.881**	**0.652**
SSPL [P60] (CVPR'22)		0.339	0.380
HTF [P15] (WACV'23)		0.394	0.400
Ours (w/ SI, w/o RA)	**VGG-Sound144k**	**0.403**	**0.402**
Ours (w/o SI, w/ RA)		**0.404**	**0.403**
Ours (w/ SI, w/ RA)		**0.406**	**0.405**

localization in a recursive manner, outperforms the HTF.

Spatial Information of the Audio Modality. Furthermore, we provide the sound source localization results with only audio features to investigate the effect of the spatial information of the audio modality. The outcomes are shown in Figure 2, revealing that the audio modal itself also contains valuable spatial cues for inferring the sound-making objects.

Video Results of the Proposed Sound Source Localization. In addition to the visualizations provided in the paper, we include supplementary video material to show the results of our method. The video displays the outcomes of our method applied to some samples from both the Flickr-Soundnet and VGG-SS datasets. Please see https://github.com/VisualAIKHU/SIRA-SSL.

Figure 1: Expanded visualization results for both (a) Flickr-SoundNet test set and (b) VGG-SS test set. Each result is obtained from models trained on the Flickr144k and VGG-Sound144k training sets.

Figure 2: Visualization results for both (a) Flickr-SoundNet test set and (b) VGG-SS test set with only audio. Each result is obtained from models trained on the Flickr144k and VGG-Sound144k training sets.