Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Unsupervised Sound Localization via Iterative Contrastive Learning

Yan-Bo Lin[a,b], Hung-Yu Tseng[c], Hsin-Ying Lee[d], Yen-Yu Lin[a,**], Ming-Hsuan Yang[e,f,g]

[a] *Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan*
[b] *Department of Computer Science, University of North Carolina at Chapel Hill, NC, US.*
[c] *Facebook, WA, US.*
[d] *Snap Research, CA, US.*
[e] *School of Engineering, University of California, Merced, CA, US*
[f] *Google Research*
[g] *Yonsei University, Seoul, South Korea*

ABSTRACT

Sound localization aims to find the source of the audio signal in the visual scene. However, it is labor-intensive to annotate the correlations between the signals sampled from the audio and visual modalities, thus making it difficult to supervise the learning of a machine for this task. In this work, we propose an iterative contrastive learning framework that requires no data annotations. At each iteration, the proposed method takes the 1) localization results in images predicted in the previous iteration, and 2) semantic relationships inferred from the audio signals as the pseudo-labels. We then use the pseudo-labels to learn the correlation between the visual and audio signals sampled from the same video (intra-frame sampling) as well as the association between those extracted across videos (inter-frame relation). Our iterative strategy gradually encourages the localization of the sounding objects and reduces the correlation between the non-sounding regions and the reference audio. Quantitative and qualitative experimental results demonstrate that the proposed framework performs favorably against existing unsupervised and weakly-supervised methods on the sound localization task.

1. Introduction

Multisensory signals (e.g., vision, hearing, and touching) provide rich information for human beings to perceive the surrounding environments. These cues from different modalities are usually closely related and thus enable human beings to perform complicated tasks in our daily lives. Take vision and audio as an example, one can easily imagine a lightning scene upon hearing thunders, associate multiple objects with their sources on a noisy street, and identify and converse with friends in a crowded cocktail party. In this work, we target the *sound localization* task (Hu et al., 2020a; Qian et al., 2020; Senocak et al., 2018, 2019) that aims to identify the sounding region in the image, as the example shown in Figure 1. Sound localization is an emerging research topic since it is the nexus of various audio-visual applications such as audio-visual source separation (Gan et al., 2020; Gao et al., 2018; Gao and Grauman, 2019b; Xu et al., 2019; Zhao et al., 2019, 2018; Tzinis et al., 2021; Tian et al., 2021; Gao and Grauman, 2021; Gao et al., 2020) and audio-visual event localization/parsing/recognition (Tian et al., 2020, 2018; Wu et al., 2019; Lin et al., 2019; Wu and Yang, 2021; Lin and Wang, 2020; Lee et al., 2021; Xuan et al., 2021; Mademlis et al., 2016; Li et al., 2016; Stafylakis et al., 2018).

Sound localization methods based on supervised learning entail a large amount of training data with the annotated sound-visual associations. Although Senocak et al. (Senocak et al., 2018, 2019) collect 5000 audio-image pairs from the Flickr-Sound database (Aytar et al., 2016) with bounding box annotations of the sounding regions, the amount of labeled data is not sufficient to train a deep learning model in a fully-supervised fashion. Moreover, it is challenging to scale up the efforts to collect a large labeled dataset since the annotators need to meticulously observe visual and audio signals simultaneously.

Semi-supervised (Senocak et al., 2018, 2019), weakly-supervised (Qian et al., 2020), and self-supervised learning frameworks (Hu et al., 2020a, 2019, 2020b) are proposed to overcome the limited data issue. The weakly-supervised meth-

[**] Corresponding author:
e-mail: lin@cs.nctu.edu.tw (Yen-Yu Lin)

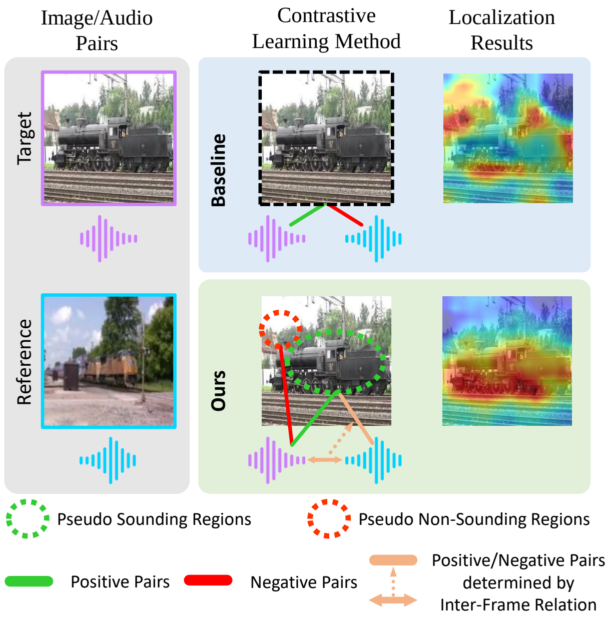


Fig. 1: **Unsupervised sound localization via iterative contrastive learning.** (*Baseline*) Existing contrastive learning usually takes the image-audio pairs sampled from the same video frame as the positive pairs, and those extracted from different videos as the negative pairs. (*Ours*) The proposed iterative approach exploits the **intra-frame sampling** that takes the sounding and non-sounding regions predicted in the previous training epoch as the *pseudo*-labels (green and red dashed circles), and the **inter-frame relation** that provides additional positive or negative correlations between the image and audio sampled across different videos where the correlations are determined by observing the relationships in the audio modality (positive in this example).

ods (Qian et al., 2020) require audio-visual event labels, and existing self-supervised methods rely on a pre-defined number of clusters (Hu et al., 2019, 2020b) or require videos of single sounding sources (Hu et al., 2020a). Furthermore, the semi-supervised methods (Senocak et al., 2018, 2019) using audio-visual correspondences alone as the supervision is less effective since a scene may contain non-sounding or ambient regions, which leads to the association between the incorrect sounding regions and reference audio signals. These issues hamper the performance of sound localization in unconstrained scenarios where the numbers of sound sources are usually unknown and there may exist objects unseen during training.

In this work, we propose an iterative contrastive representation learning algorithm that does not require any prior assumption or labels for the sound localization task. Starting from conventional contrastive learning (Senocak et al., 2018, 2019), we use the sound localization model obtained in the *previous* epoch to estimate the sounding and non-sounding regions as the pseudo-labels for the current epoch. With such pseudo regions, the model is encouraged to disassociate non-sounding or ambient regions from object sounds and thus explores more negative training samples for contrastive learning. In addition to the relationships between the audio and visual signals within an instance, we correlate audio signals *across* instances. For instance, if the audio clips of two different instances are semantically similar, the image and audio *across* the two instances should be positively correlated and can then serve as a positive pair for contrastive learning, and vice versa. We show an example of two train sounds across instances in *inter-frame relation* of Figure 1. Such a strategy alleviates typical contrastive learning methods from differentiating the representations of the related sounding object and audio signals across instances, and provides more reliable guidance to learn a sound localization model.

We evaluate the proposed method on the Flickr-Sound (Senocak et al., 2018, 2019) and the MUSIC-Synthetic (Hu et al., 2020a) datasets using the consensus intersection over union (cIoU) and area under curve (AUC) as evaluation metrics. Both qualitative and quantitative results demonstrate the effectiveness of the proposed method on the sound localization task. The main contributions of this work are summarized as follows:

- We propose an iterative contrastive learning algorithm to tackle the sound localization task without any data annotations.
- Our method not only leverages regions of interests, but also exploits non-sounding regions as well as the relationship across audio instances to jointly learn the audio and visual representations.
- Qualitative and quantitative experimental results on the benchmark dataset demonstrate that the proposed method performs favorably against the state-of-the-art weakly supervised and unsupervised approaches.

2. Related Work

Some topics related to the development of the proposed approach are discussed in this section.

2.1. Self-Supervised Audio-Visual Representation Learning

Inherent correlation among different modalities of a video provides supervisory signals for learning a deep neural network model. Information sources used in existing self-supervised audio-visual representation learning methods can be broadly categorized as follows. First, audio-visual pairs are extracted from a video clip as positive association. The assumption is that the audio and visual features extracted from the same video clip should be strongly correlated (Arandjelovic and Zisserman, 2017; Arandjelović and Zisserman, 2018; Aytar et al., 2016; Owens et al., 2016; Alayrac et al., 2020; Alwassel et al., 2020; Asano et al., 2020; Ma et al., 2021; Min et al., 2020; Chen et al., 2020). In addition, these schemes differentiate the features extracted from unpaired video clips. Furthermore, some concurrent methods (Morgado et al., 2021b,a) jointly consider the correlations within each modality or across different modalities (i.e., audio and vision). Different from (Morgado et al., 2021b,a) that learn visual information of an entire image, our method leverages pseudo-annotations to provide training guidance from both sounding and non-sounding regions. Second, video temporal information (Owens and Efros, 2018; Korbar et al., 2018; Chung and Zisserman, 2018) is explored to determine strong or weak correlation. Given a video sequence, a few methods sample the audio and visual features from the same time frame as strong correlation and consider those across different frames as weak correlation for the representation. Third, spatial relations among image regions are exploited. Since the binaural recording techniques (spatial audio) preserve the spatial information of the sound origins, some approaches (Gao and Grauman, 2019a; Morgado et al., 2020, 2018; Yang et al., 2020; Zhou et al., 2020; Lin and Wang, 2021; Xu et al., 2021; Lu et al., 2019) jointly model the visual and audio information spatially to construct spatial audio generation systems or learn representations for downstream tasks.

2.2. Sound Source Localization in Visual Scenes

This task aims to find corresponding sounding regions in images from audio signals. We categorize methods addressing this task into three groups. The first group of work (Afouras et al., 2020; Senocak et al., 2018, 2019) leverages the correspondence between audio and visual signals for supervision. These methods assume that the audio and visual features extracted from the same video clip should be more similar than those extracted from different clips. Some sound localization methods (Senocak et al., 2018, 2019) are formulated in a semi-supervised way to deal with limited annotated data. The second line of work uses the class activation map (CAM) (Zhou et al., 2016) to determine discriminative regions for categorical prediction. Owens et al. (Owens and Efros, 2018) learn the audio and visual representations by the audio-visual correspondence and perform sound localization using the CAM model. Similarly, given the event labels, Qian et al. (Qian et al., 2020) use the CAM model to identify sounding regions and corresponding audio clips. As such, the sound and visual object in the same event can be associated. Finally, some models (Hu et al., 2019, 2020b) utilize audio-visual clusters to model audio-visual relationships. These methods cluster different frequencies of an audio signal and visual patches in the images. The centers of the audio and visual clusters extracted from the same video clip are associated during the training stage.

We note that existing sound localization approaches are limited in several aspects. These methods typically require additional information in other modalities (e.g., optical flow (Afouras et al., 2020)), a pre-defined number of sound sources (Hu et al., 2019, 2020b), event labels in both audio and visual modalities (Qian et al., 2020), or single-source videos (Hu et al., 2020a). In this work, we present a sound localization framework that does not rely on any additional annotation or assumption. Furthermore, the correlation between (non-)sounding objects and audio across pairs is jointly considered to further enhance sound localization.

3. Methodology

This section describes the proposed method.

3.1. Sound Localization

Our goal is to localize the source of the detected sound in the image. Specifically, given the input image of size $W \times H \times 3$ and the detected audio, i.e., sound, we aim to estimate the sounding region $\mathbf{S}$. As shown in lower left panel of Figure 2, the proposed sound localization model first extracts the corresponding visual representation $\mathbf{V} \in \mathbb{R}^{w \times h \times d}$ from the input image, and the audio feature representation $\mathbf{a} \in \mathbb{R}^d$ from the short-time Fourier-transformed (Griffin and Lim, 1984) audio. We then use the attention mechanism to compute the response map $\mathbf{R} \in \mathbb{R}^{w \times h \times 1}$ followed by min-max normalization,

$$\begin{aligned} \mathbf{R} &= \mathbf{V} * \mathbf{a}, \\ \mathbf{R} &= \frac{\mathbf{R} - \min(\mathbf{R})}{\max(\mathbf{R}) - \min(\mathbf{R})}, \end{aligned} \tag{1}$$

where the notation $*$ represents the pixel-wise inner-product operation. We then determine the potential sounding region by thresholding the response map $\mathbf{R}$:

$$\mathbf{S} = \mathrm{idx}(\mathbf{R} > \delta_v), \tag{2}$$

where $\delta_v \in [0, 1]$ is a parameter for thresholding. The function $\mathrm{idx}(\cdot)$ returns the spatial indexes of the sampled patches that match the given condition.

In the following, we will illustrate how the proposed method learns to localize sound via audio-visual representation learning. The baseline audio-visual contrastive learning is first introduced. It is used for initializing our model. We then present our iterative training approach and finally discuss how we leverage the relationship given in the audio signals to facilitate the contrastive learning process.

3.2. Audio-Visual Representation Learning

Contrastive Learning. As audio-image pairs extracted from videos provide natural implication of the correlation between the two modalities, we use contrastive learning (Oord et al., 2018) to learn the audio-visual feature representations in an unsupervised manner. The core idea is to maximize the correlation between the audio and visual representations extracted from the same video (i.e., positive pairs) while minimizing the correlation between those from different videos (i.e., negative pairs). Specifically, during the training stage, our model extracts a set of audio features $\{\mathbf{a}_1, \cdots, \mathbf{a}_k\}$ and a set of visual representations $\{\mathbf{V}_1, \cdots, \mathbf{V}_k\}$ from the input batch consisting of k image-audio pairs sampled from the same videos. Then the model is optimized by the following training objective:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{k} \sum_{i=1}^{k} \Big[\log \frac{\exp(\phi(\mathbf{V}_i) \cdot \mathbf{a}_i / \tau)}{\sum_{j=1}^{k} \exp(\phi(\mathbf{V}_i) \cdot \mathbf{a}_j / \tau)} \Big], \tag{3}$$

where the term τ is a hyper-parameter controlling the temperature. The notation ϕ represents the operations of $L2$ normalization on the feature dimension followed by average pooling on the spatial dimensions.

Iterative Contrastive Learning. Since an image typically contains both sounding and non-sounding regions, the training loss in Eq. (3) is less effective as it takes the whole image into consideration at a time, which may associate non-sounding regions with the audio signals extracted from the same video. Moreover, the annotations of the sounding objects are not available under the unsupervised setting.

To this end, we develop an iterative contrastive learning approach. As illustrated in Figure 2, starting from using conventional contrastive learning in Eq. 3 for initialization, we take

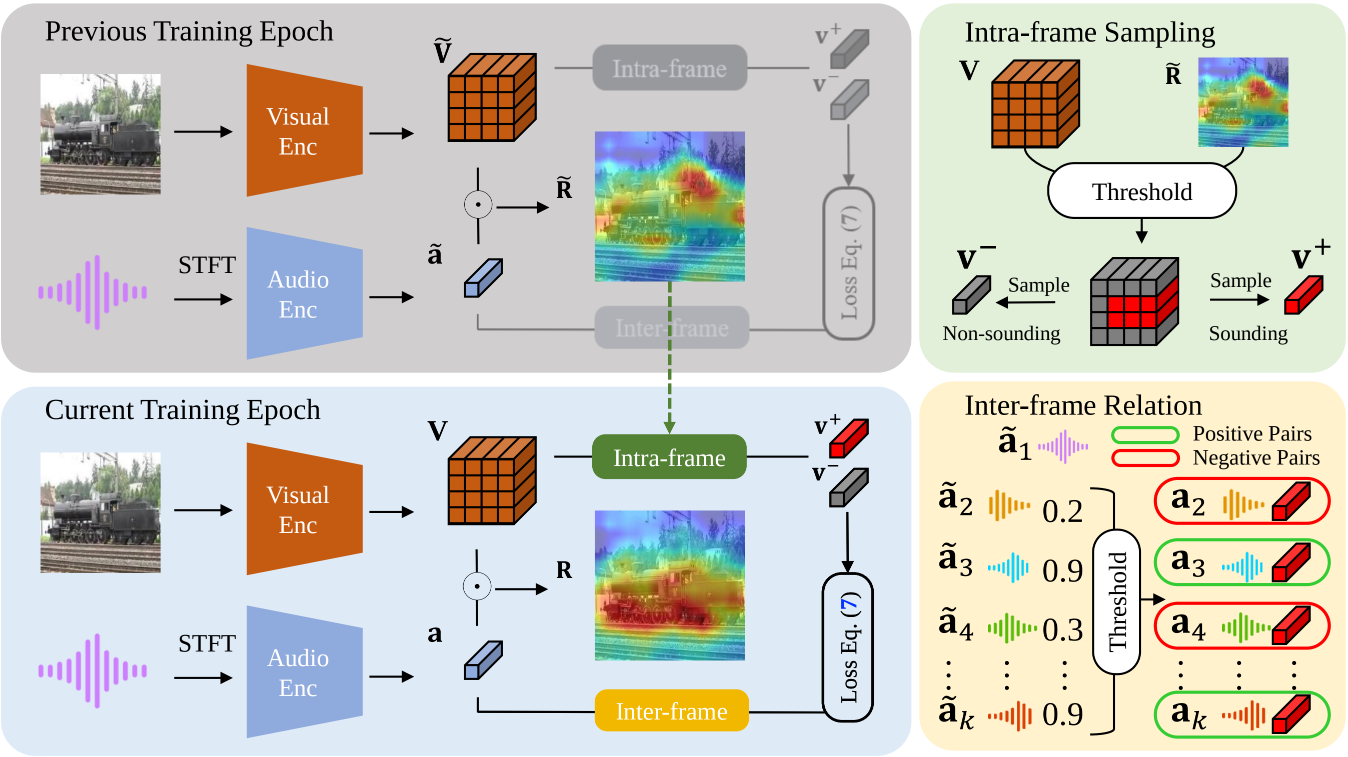


Fig. 2: **Algorithm overview.** Our framework consists of a visual feature extractor, an audio feature extractor, an intra-frame sampling module, and an inter-frame relation module. *(upper-left)* Sound localization $\tilde{R}$ is obtained by computing the correlation between the visual and audio features. *(bottom-left)* Our iterative contrastive learning scheme uses the localization results predicted in the *previous* training epoch as the *pseudo*-labels for the current epoch. *(upper-right)* The intra-frame sampling module uses the pseudo-labels to extract (non-)sounding regions for enhancing the efficacy of the contrastive learning. *(lower-right)* The inter-frame relation module determines the correlation of images and audios sampled across videos by observing the relationship in the audio modality.

the sound localization results predicted in the *previous* training epoch as the *pseudo*-labels for *current* training epoch. Specifically, let $\tilde{\mathbf{R}}_i = \tilde{\mathbf{V}}_i * \tilde{\mathbf{a}}_i$ denote the response map predicted from the model with parameters in the previous training epoch. We randomly sample the visual features from patches, which show high responses on the map $\tilde{\mathbf{R}}_i$ in the previous epoch, as the sounding feature $\mathbf{v}^+$ i.e.,

$$\begin{aligned} \mathbf{x}_i^{\text{pos}} &= \text{idx}(\tilde{\mathbf{R}}_i > \delta_v), \\ \mathbf{v}_i^+ &= \phi(\text{feats}(\mathbf{V}_i, \mathbf{x}_i^{\text{pos}})), \quad i = 1, 2, ..., k, \end{aligned} \tag{4}$$

where function feats$(\cdot)$ returns a set of visual features for the given indexes. We replace the term $\phi(\mathbf{V}_i)$ in Eq. 3 with the sounding feature $\mathbf{v}_i^+$. In this way, the sounding regions are iteratively explored while non-sounding regions are gradually excluded. In practice, we perform min-max normalization for $\tilde{\mathbf{R}}_i$, same as Eq. 1, to prevent the threshold δ_v too high to find confident sounding patches. Furthermore, we adopt attention score normalization and randomly sample positive visual patches for contrastive learning, which make our method less sensitive to the qualities of initial models.

Intra-Frame Sampling. We enhance the efficacy of the proposed contrastive learning by incorporating more negative pairs. However, merely sampling more negative pairs by extracting audio and images from different videos is less effective as the model may easily determine the correlation. Consequently, we propose to use the *pseudo-non-sounding* regions predicted in the previous training epoch to form the negative pairs with the audio clips extracted from the same video. We illustrate the process in Figure 1 (red line and red dotted circle) and Figure 2 (top right). The correlation of these negative pairs is more challenging to determine as they are sampled from the *same* video sequence, thus helping the sound localization model to learn more discriminative audio-visual representations. We call such a strategy intra-frame sampling, which is formulated as follows:

$$\begin{aligned} \mathbf{x}_i^{\text{neg}} &= \text{idx}(\tilde{\mathbf{R}}_i < \delta_v), \\ \mathbf{v}_i^- &= \phi(\text{feats}(\mathbf{V}_i, \mathbf{x}_i^{\text{neg}})), \quad i = 1, 2, ..., k. \end{aligned} \tag{5}$$

Inter-Frame Relation. As the semantically similar contents may appear in different video sequences, contrastive learning can be further improved if it explores the correlation between images and audio signals from different videos. An example is given in Figure 1 (black line and green dotted region). Specifically, we leverage the relationship in the audio modality to determine the correlation of the image and audio clip sampled from different videos. The relationship in the audio modality is estimated by using the audio representations $\tilde{\mathbf{a}}$ computed in the *previous* training epoch. As shown in the bottom-right corner of Figure 2, we determine the correlation $y_{ij} \in \{0, 1\}$ between the i-th image and the j-th audio within the same mini-batch according to the audio representations, i.e.,

$$y_{i,j} = \begin{cases} 1, & \text{if } \left\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \right\rangle \geq \delta_a, \\ 0, & \text{otherwise,} \end{cases} \quad \forall i, j \in \{1, \ldots, k\}, \tag{6}$$

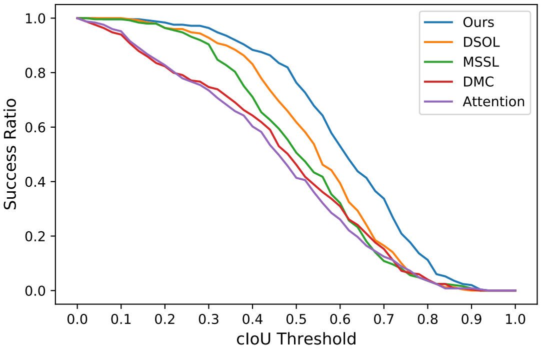


Fig. 3: **Success ratio under different cIoU thresholds.** Success ratio indicates the ratio of all instances whose cIoU scores are higher than thresholds. Note that a larger area under the curve (AUC) indicates better performance.

where the term $\delta_a \in [0, 1]$ is a thresholding parameter. Combining the proposed intra-frame sampling and inter-frame relation strategies, our training objective becomes

$$\mathcal{L}^{\text{iterative}}_{\text{contrast}} = - \frac{1}{k} \sum_{i=1}^{k} \Big[\log \frac{\sum_{j=1}^{k} y_{i,j} \exp(\mathbf{v}_i^{+} \cdot \mathbf{a}_j / \tau)}{\sum_{j=1}^{k} \exp(\mathbf{v}_i^{-} \cdot \mathbf{a}_j / \tau) + \exp(\mathbf{v}_i^{+} \cdot \mathbf{a}_j / \tau)} \Big]. \quad (7)$$

We train our sound localization model using Eq. (3) at the initialization stage, and then iteratively optimize the objective in Eq. (7) until the localization results converge.

4. Experimental Results

4.1. Datasets

We evaluate all methods on the following two datasets:

- **SoundNet-Flickr** (Aytar et al., 2016) dataset consists of more than two million video sequences. We use a 5-second audio clip and the central frame of the 5-seconds corresponding video clip, to form an input pair for the proposed framework. Note that we do not rely on any annotation (e.g., bounding boxes) for model training. In all experiments, we perform the training process with the subsets of the SoundNet-Flickr dataset constructed by Qian et al. (Qian et al., 2020) that contains 10k and 20k audio-visual pairs. Following the protocol in (Qian et al., 2020; Senocak et al., 2018, 2019), we conduct the evaluation using the testing set of the SoundNet-Flickr dataset which consists of 250 audio-visual pairs with bounding box annotations.

- **MUSIC-Synthetic** (Hu et al., 2020a) is a dataset consisting of synthetic audio-visual pairs. Each audio-visual pair is constructed by concatenating four music instrument frames and randomly selecting two out of four corresponding 1-second audios. In other words, for each audio-visual pair, there are two instruments making sound while the other two are silent. We follow the protocol (Hu et al., 2020a) to train the models with all 25k audio-visual pairs in the training set and conduct the evaluation on the testing set consisting of 455 audio-visual pairs with bounding box annotations.

4.2. Implementation Details

Network Architecture. For a fair performance evaluation, we use the official source code of the MSSL (Qian et al., 2020) model. We use the ResNet18 (He et al., 2016) backbone for both the audio and visual feature extractors. Particularly, the stride of 4^{th} residual block in our extractors is set to 1. For the visual representation, we use the feature extracted by the 4^{th} residual block. The dimension of the visual representation is $16 \times 16 \times 512$. As for the audio representation, we also use the feature extracted by the 4^{th} residual block. We then use the average pooling operation to reduce the size of the audio representation to $1 \times 1 \times 512$.

Optimization. We set the temperature terms τ in Eq. 3 and Eq. 7 to 0.25 and 0.05, respectively. We optimize the initial localization model using Eq. 3 the first six epochs. We then use the proposed iterative algorithm to train the sound localization model in the following epochs. Specifically, we randomly sample 12 audio-related visual patches described in Eq. 4. Note that it is the maximum number of sampled patches. We also set the maximum number of sampled non-sounding patches illustrated in Eq. 5 to 12. The threshold parameter δ_v in Eq. 4 and Eq. 5 is set to 0.25, while the threshold term δ_a inEq. 6 is set to 0.85.

We use the SGD optimizer to train our model. The momentum parameter is set to 0.9, and the weight decay parameter is set to $5e - 4$. We set the mini-batch size to 96, the learning rate of visual encoder to $5e - 4$, and the learning rate of audio encoder to $3e - 4$. The training is conducted on a single GTX 1080 Ti GPU with 12GB memory.

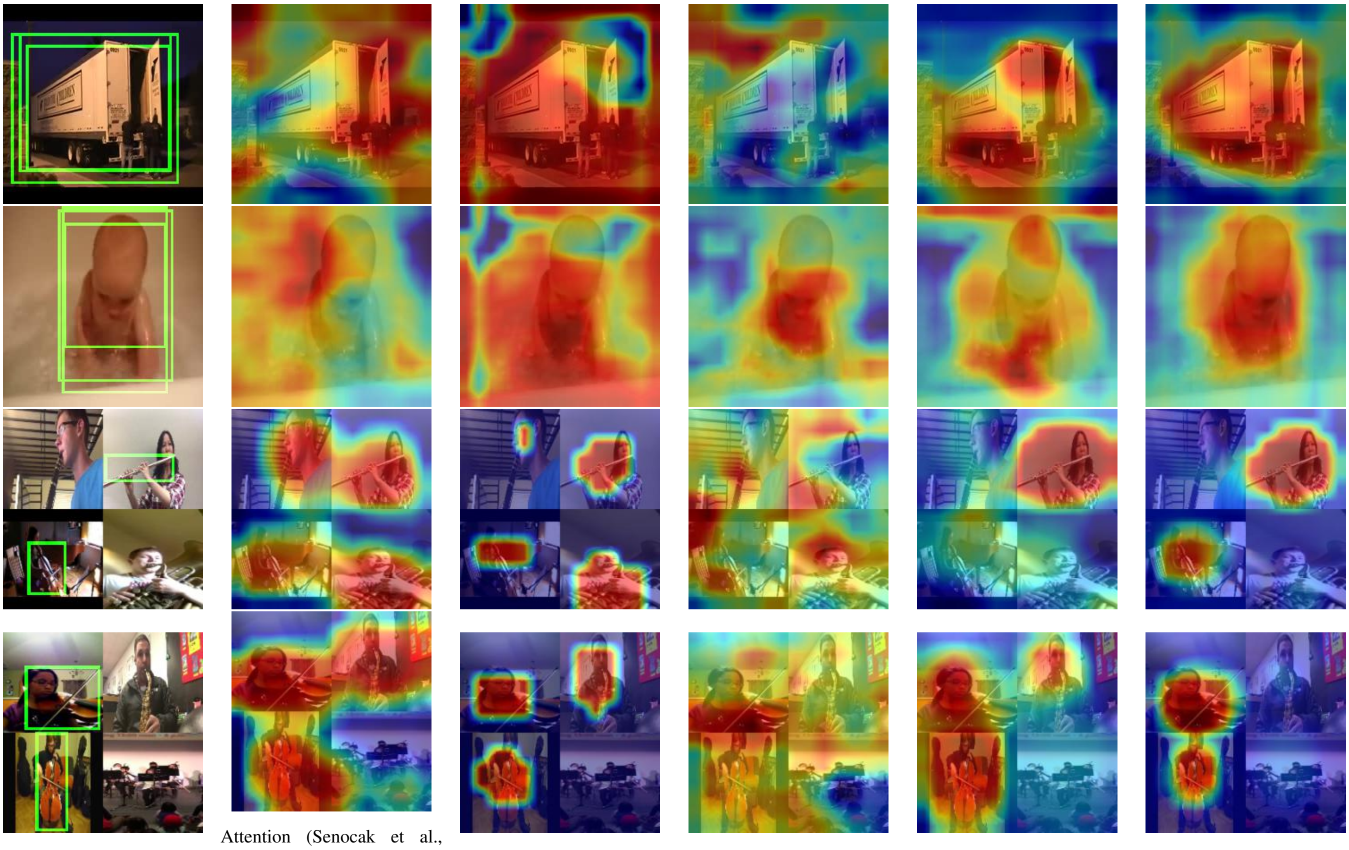
Evaluation Metrics. Following previous work (Hu et al., 2019, 2020b; Qian et al., 2020; Senocak et al., 2018, 2019), we adopt *consensus intersection over union (cIoU)* and *area under curve (AUC)* as the evaluation metrics. Note that the ground-truth sounding region of an image is computed according to the overlapping of the bounding box labels annotated by different people. The response map **R** in Eq. (1) is post-processed to serve as the sound localization results for evaluation. Specifically, we first compute the response map **R** using Eq. (2). Then we recover the resolution of the response map **R** from $w \times h$ to original image resolution $W \times H$ using bilinear up-sampling.

Competing methods. We compare the proposed method to the following weakly- and unsupervised approaches:

- **Attention** (Senocak et al., 2018, 2019) is trained using the audio-visual co-attention mechanism.
- **DMC** (Hu et al., 2019) is an unsupervised approach based on the usage of audio-visual clusters, and requires *a pre-defined number* of sound sources. We set the number of source to one suggested by (Hu et al., 2019, 2020b) for the SoundNet-Flick dataset and set to two for the MUSIC-Synthetic dataset.
- **MSSL** (Qian et al., 2020) reports the state-of-the-art performance on the sound localization task. It requires audio/visual event labels obtained from pre-trained classifiers and the CAM (Zhou et al., 2016) predictions to find the sounding regions.
- **DSOL** (Hu et al., 2020a) is a two-stage approach requiring a large amount of single-source videos for the first stage to

Table 1: **Quantitative results of sound localization.** We evaluate all methods on the SoundNet-Flickr (Senocak et al., 2018, 2019) and MUSIC-Synthetic (Hu et al., 2020a) datasets with cIoU and AUC metrics. Following the evaluation protocol in (Senocak et al., 2018, 2019; Hu et al., 2020a), we evaluate the cIoU@0.5 and cIoU@0.3 for SoundNet-Flickr and MUSIC-Synthetic, respectively.

Method	SoundNet-Flickr 10K		SoundNet-Flickr 20K		MUSIC-Synthetic	
	cIoU@0.5↑	AUC ↑	cIoU@0.5 ↑	AUC ↑	cIoU@0.3 ↑	AUC
Random	7.2	30.7	–	–	0.2	9.6
Attention (Senocak et al., 2018)	42.1	43.8	45.3	46.7	6.9	14.2
DMC (Hu et al., 2019)	41.4	45.0	47.0	47.5	6.6	15.3
MSSL (Qian et al., 2020)	51.2	50.4	53.8	50.6	4.3	12.1
DSOL (Hu et al., 2020a)	56.6	51.5	58.7	52.9	15.4	17.0
Ours	**71.0**	**58.0**	**74.7**	**59.6**	**25.1**	**21.9**



Input Image

Attention (Senocak et al., 2018)

DMC (Hu et al., 2019)

MSSL (Qian et al., 2020)

DSOL (Hu et al., 2020a)

Ours

Fig. 4: **Qualitative comparisons.** We compare with state-of-the-art sound localization methods on the SoundNet-Flicker (Senocak et al., 2018, 2019) (1st and 2nd rows) and MUSIC-Synthetic (Hu et al., 2020a) (3rd and 4th rows) datasets. Sound localization is presented using heat maps, in which red regions indicate the estimated sound source. Note that the bounding boxes are the annotations of sounding regions from multiple annotators for the SoundNet-Flicker dataset. Sound localization is presented by using heat maps, in which redder regions indicate higher probabilities of being the sound source.

build up class-based visual dictionaries and train audio and visual encoders. For a fair comparison, we only train the network in the second stage. We use pre-trained audio and visual encoders and use CAM (Zhou et al., 2016) predictions to replace visual dictionaries.

4.3. Quantitative Results

Table 1 shows the quantitative comparisons on the SoundNet-Flickr and MUSIC-Synthetic datasets. The proposed method performs favorably against the competing approaches on the sound localization task. We note that different from the proposed method, the competing schemes require a pre-defined number of sounding sources (i.e., DMC) or audio/visual event labels (i.e., MSSL). In contrast, the proposed method does not need any prior knowledge about the source number or data annotations. Furthermore, our model trained with 10k audio-visual pairs already outperforms MSSL and DSOL approaches which use more (i.e., 20k) audio-visual pairs during training. In addition to the cIoU metric, the cIoU scores calculated with various thresholds are shown in Figure 3. Our method reports favorable cIoU scores under all thresholds. The consistent performance advantage suggests the effectiveness and efficacy of our iterative contrastive learning algorithm.

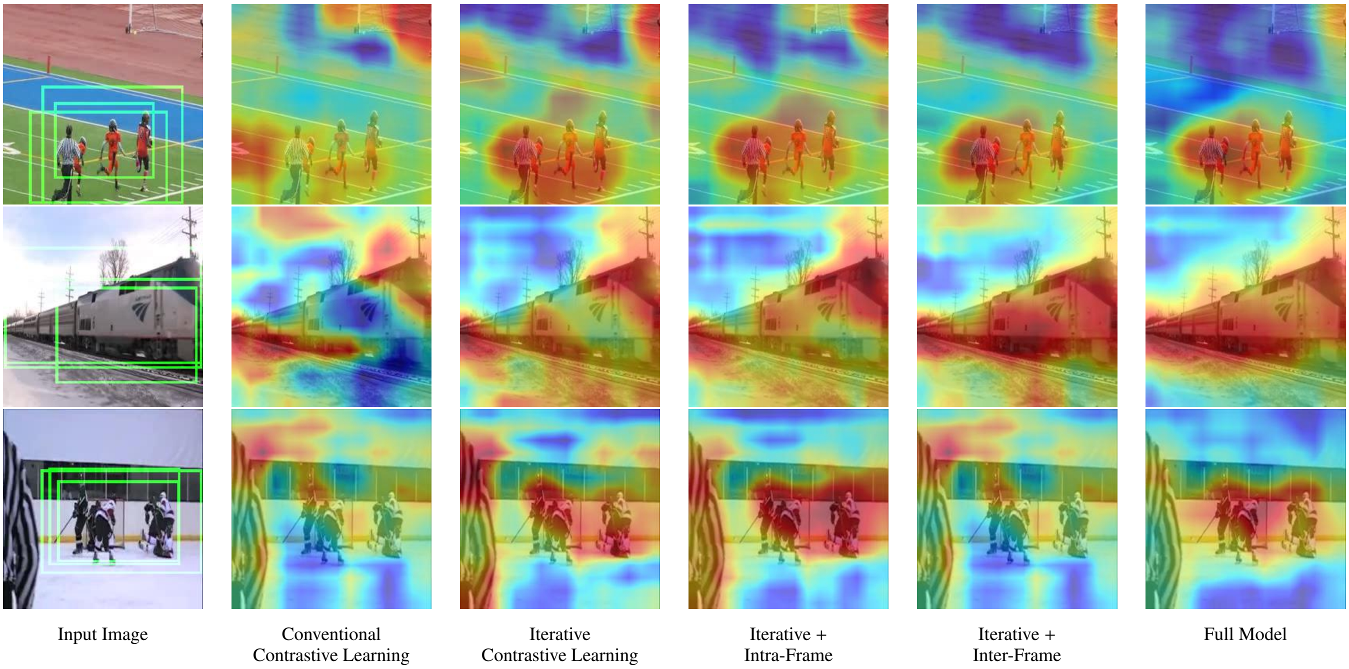


Fig. 5: **Example localization results of using different design components in the proposed method on the SoundNet-Flicker (Senocak et al., 2018, 2019) dataset.** (*from left to right*) We show the qualitative of conventional contrastive learning, iterative contrastive learning, iterative approach w/ intra-frame sampling, iterative approach w/ inter-frame relation, and our full model. Sound localization is presented by using heat maps, in which redder regions indicate higher probabilities of being the sound source.

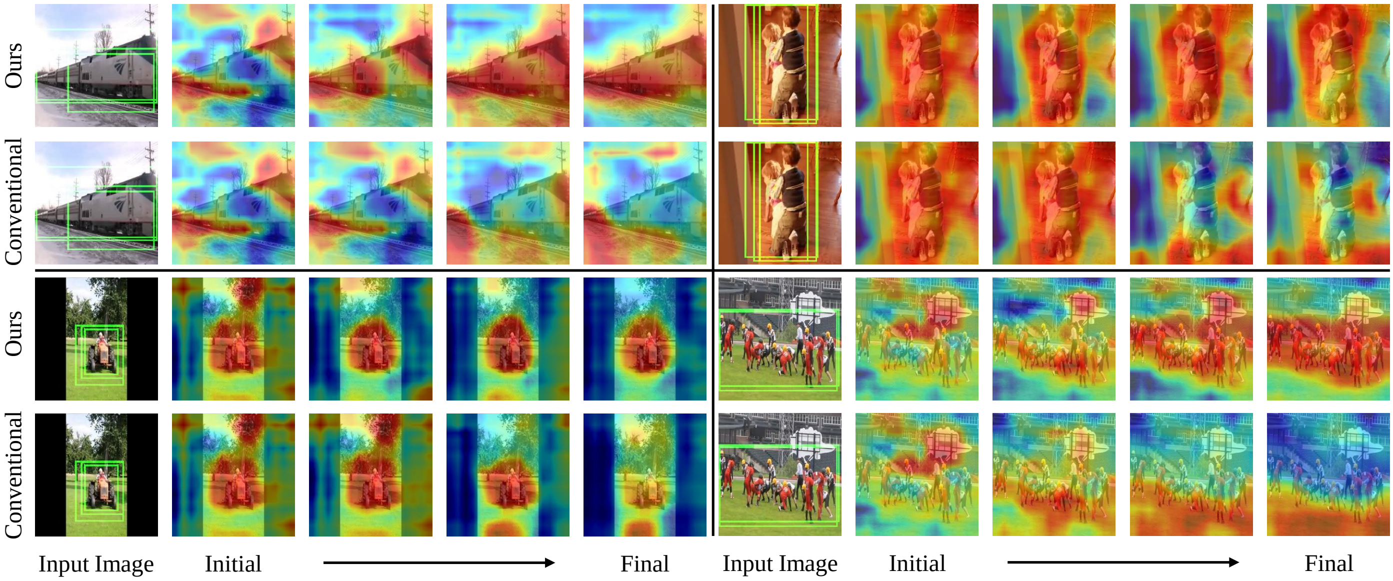


Fig. 6: **Sample localization results at different training epochs on the SoundNet-Flicker dataset (Senocak et al., 2018, 2019).** We present the sound localization results estimated by our method (1st and 3rd rows) and the conventional contrastive learning approach (2nd and 4th rows) at different (initial to final) training epochs. Sound localization is presented by using heat maps, in which redder regions indicate higher probabilities of being the sound source.

4.4. *Qualitative Evaluation*

We demonstrate the qualitative comparisons in Figure 4. The localization results of the proposed method are more accurate compared to those of the competing approaches. The example in the 3rd and 4th row is particularly challenging. Since the multiple-sounding and non-sounding instruments appear in the same scene, it is difficult to localize exact-sounding objects. MSSL and DSOL are both struggling with unrelated background. As for DMC, with the prior defined number of sounding source for the MUSIC-Synthetic dataset, it is more resistant to the unrelated background yet fail to identify the sounding instruments correctly. Compared to these methods, the proposed framework can focus on the sounding objects with better accuracy, while trained without audio-visual event labels or any prior information.

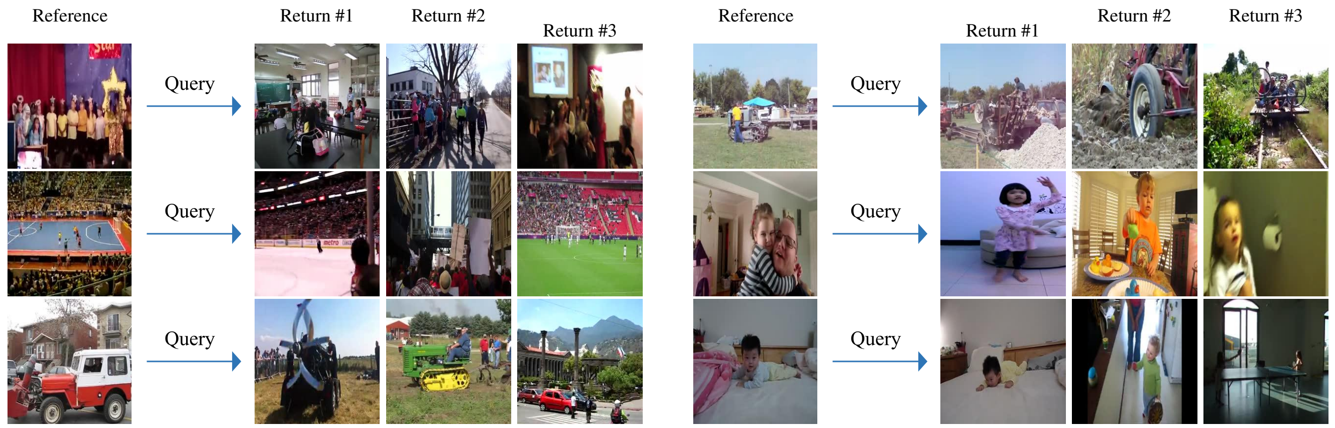


Fig. 7: **Retrieval results from audio signals on the SoundNet-Flicker (Senocak et al., 2018, 2019).** We use the sounds of the reference images as the *queries* to retrieve the top-3 related audio clips and show the corresponding images. The results verify our idea that the relationships in the audio modality can help the association between images and audios extracted across videos.

Table 2: **Ablation study.** (*bottom*) We investigate the effect of using different design components in the proposed method. (*top*) We show how we improve the MSSL approach by modifying the localziation method in Eq. 2 and normalization strategy in Eq. 1.

Method	cIoU@0.5 ↑	AUC ↑
MSSL (Qian et al., 2020)	52.2	49.6
MSSL Stage I	42.2	48.1
MSSL Stage I w/o Labels	10.8	30.2
MSSL Stage I w/ Eq. 2	47.4	48.7
MSSL Stage I w/o Labels w/ Eq. 2	47.0	48.7
MSSL Stage I w/ Eq. 2 Eq. 1	50.2	49.0
MSSL Stage I w/o Labels w/ Eq. 2 Eq. 1	46.6	48.3
Ours Initial	57.8	52.1
Ours Itr(✓) Intra(✗) Inter(✗)	64.2	54.2
Ours Itr(✓) Intra(✓) Inter(✗)	69.4	56.9
Ours Itr(✓) Intra(✗) Inter(✓)	67.1	55.9
Ours 10K	**71.0**	**58.0**

Table 3: **Number of sampled patches.** We show the cIoU and AUC scores of sampling different numbers of patches described in Eq. 4 and Eq. 5 in the paper. The best performance is reported in **bold**.

Number of Patches	cIoU@0.5 ↑	AUC ↑
1	56.6	49.9
2	57.0	49.8
4	57.4	50.4
6	60.6	54.1
8	65.0	55.1
12	**71.0**	**58.0**
16	69.4	57.6
32	66.2	56.3
48	67.4	57.0
64	60.6	53.6
96	57.8	50.3
128	57.0	50.2

4.5. Ablation Study

We conduct the ablation study to analyze the individual effect of each design component in the proposed method. The results are presented in the fourth block of Table 2, where **Itr** indicates the iterative contrastive training that uses the pseudo-sounding regions inferred from the previous epoch, **Intra** represents the usage of the pseudo-non-sounding regions, and **Inter** is the proposed inter-frame relation module. We also demonstrate the qualitative comparisons in Figure 5. Particularly, the iterative strategy (i.e., **Itr**) ensures the localization model focus only on the sounding region compared to the conventional contrastive learning approach (i.e., **Initial**). Both the quantitative and qualitative results confirm the efficacy of individual components designed in our approach.

Comparison with MSSL. The proposed method shares similar backbone with the MSSL (Qian et al., 2020) method. Therefore, we also conduct the ablation study to show the effect of each modification we made, including replacing CAM with thresholding for sounding region localization (Eq. (2)), normalization (Eq. (1)), and conventional contrastive learaning (Eq. (3)). The results are summarized in the first three blocks of Table 2. Since the MSSL method uses a two-stage model trained with audio-visual event labels, we study the case of removing the second stage (**Stage I**) and training without labels (**w/o Labels**). As the results shown in the first block, training with the first stage and without labels both significantly degrade the performance of the MSSL method. We show in the second and third block that using Eq. (2) and Eq. (1) can greatly improve the performance. Finally, we obtain our baseline (**Initial**) by applying Eq. (3) to the MSSL **Stage I w/o Labels** method with Eq. (2) and Eq. (1). To conclude, Table 2 summarizes the effect of the proposed components and the transition from the original MSSL method to the proposed approach.

Localization results in various epochs. Since the proposed iterative method is based on the strategy where the localization results predicted in the previous training epoch serve as the pseudo-label, the iterative localization results are crucial. Therefore, we visualize the localization results at different epochs. As shown in Figure 6, the localization results gradually focus on the sounding regions. The results validate the efficacy of the proposed iterative procedure that takes localiza-

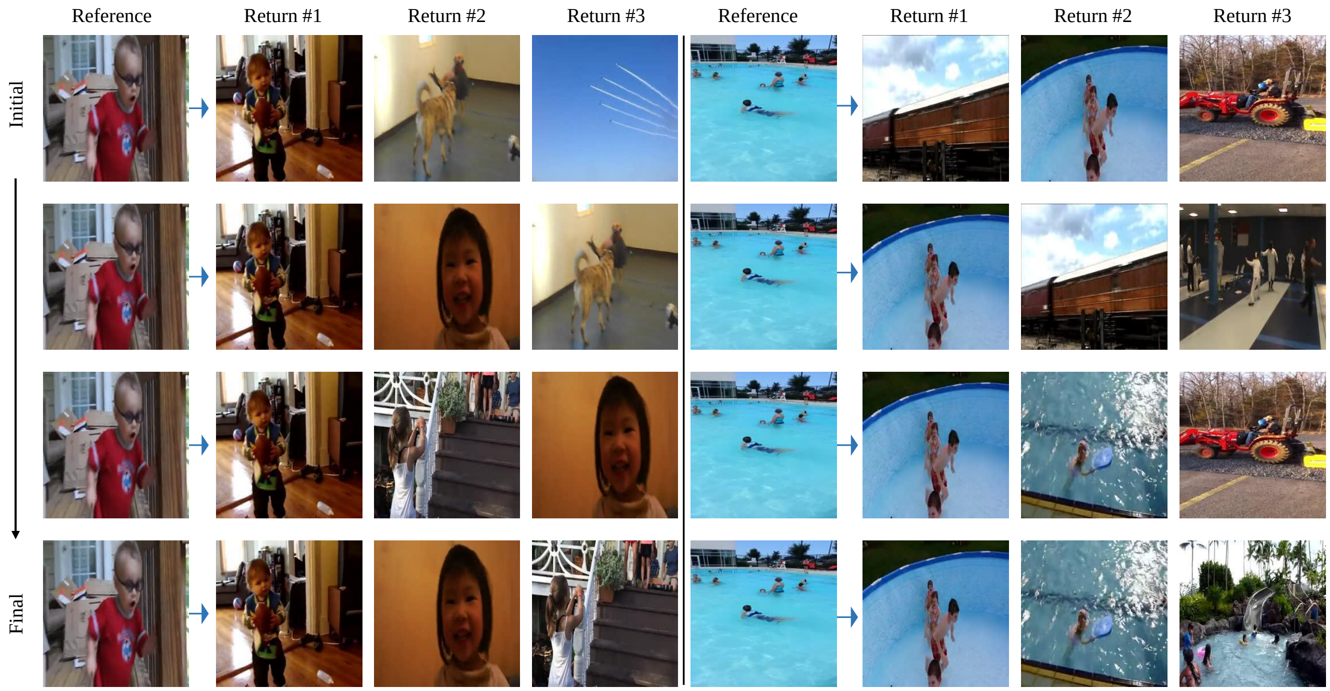


Fig. 8: **Audio retrieval results at different training epochs.** We use the sounds of the reference images as the *queries* to retrieve the top-3 related audio clips and show the corresponding images. Note that rows from top to bottom indicates different training epochs, where the bottom row represents the final epoch. The results verify our idea behind the inter-frame relation module that the correlation between audios can be used to determine the association between the images and audios extracted across videos.

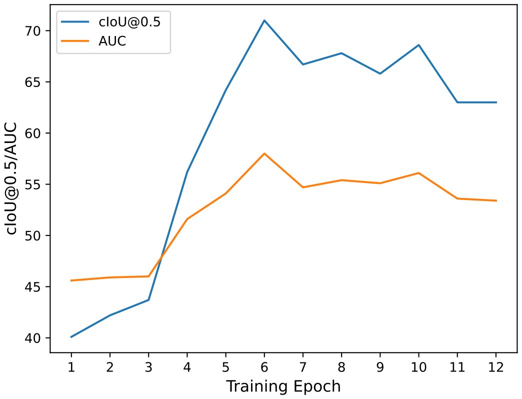


Fig. 9: **Effect of training epochs for initializing the sound localization model.** We show the cIoU and AUC scores of using Eq. 3 to initialize the model with different training epochs.

tion results from the previous epoch as training guidance for the current epoch.

Relationships in audio modality. The proposed inter-frame relation illustrated in Section 3.2 is based on the assumption that the relationships in the audio modality can be the guidance of the contrasting learning. To verify the assumption, we visualize the retrieval results in the audio modality in Figure 7. Specifically, given a reference audio-visual pair, we retrieve the top three audio-visual pairs according to the distances between audio features. We present the images of the reference and retrieved visual-audio pairs in Figure 7. As the reference and retrieved images share semantically similar contents, we validate the intuition behind the proposed inter-frame relation design.

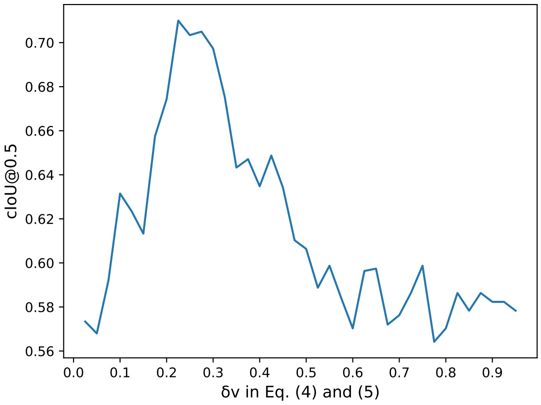


Fig. 10: **Success ratio under different thresholds δ_v.** We present the cIoU@0.5 scores of using the different thresholds δ_v described in Eq. 4 and Eq. 5.

As described in Figure 7, we use the correlation between audios to determine the association between the audio and image sampled from different video sequences. Therefore, we present the audio retrieval results at *different training epochs* in Figure 8. Note that we use the audio signals for the retrieval and present the results using the corresponding images. As demonstrated in Figure 8, compared to the reference audio, the content of the retrieved audio becomes more semantically similar over the training epochs. This verifies the efficacy of the inter-frame relation module in the iterative proposed contrastive learning process.

4.6. Parameter Analysis

Initial Contrastive Learning. To analysis the effect of the initialization by the conventional contrastive learning, we conduct an ablation study on training the initial localization via Eq. 3 with different epochs. The results presented in Figure 9 suggests that simply a few training epochs for the initialization are enough for the following iterative contrastive learning. Moreover, initializing the model with too many training epochs degrades the sound localization performance.

Sampled Patches. As we sample pseudo-(non-)sounding patches in the proposed iterative contrastive learning, we analyze the effect of the number of sampled patches described in Eq. 4 and Eq. 5. The results demonstrated in Table 3 suggest the optimal number of sampled patches to be in the range of $[12, 16]$. Note that we only consider all the patches that match the sounding criterion described in Eq. 4 even though the number of sampled regions is too large.

Thresholding parameters. In the proposed method, we use thresholding parameters δ_v for the intra-frame sampling and δ_a for the inter-frame relation modules to determine the correlation between audios and images, as described in Eq. 4, Eq. 5, and Eq. 6. As shown in Figure 10 and Figure 11, we study the optimal values for the thresholding parameters. The results show that the reasonable ranges for the thresholding parameters δ_v and δ_a are $[0.2, 0.3]$ and $[0.8, 0.9]$, respectively. For δ_v, since we randomly sample positive patches among pseudo-positive regions, a low threshold contributes to selecting diverse patches, thus leading to better performance. Note that the cIoU scores lower than 0.58 (i.e., results of conventional iterative contrastive learning) indicate that the proposed method is not effective and degrades the sound localization performance.

5. Limitation

In the following, we discuss the limitations of the proposed method. Without jointly considering temporal information, our method cannot properly localize sounding objects when multiple objects of similar appearance are present. Our method can only work on a single sounding object (e.g., those on SoundNet-Flicker) and multiple sounding objects with different appearances (e.g., those on MUSIC-Synthetic). Also, our method suffers from the cases where the sounding objects do not appear in the frame.

6. Conclusions

In this paper, we present a novel unsupervised sound localization framework that does not require any prior assumption or data annotation. We propose two modules to provide pseudo positive and negative training pairs based on an iterative contrastive learning pipeline. The *intra-frame sampling* leverages the localization results estimated in the previous epoch as pseudo-labels. The *inter-frame relation* contributes to training pairs across different videos by exploiting the relationships in the audio modality with the audio features learned from the previous epoch. Extensive experimental results show that our approach performs favorably against the state-of-the-art weakly-supervised and unsupervised algorithms.

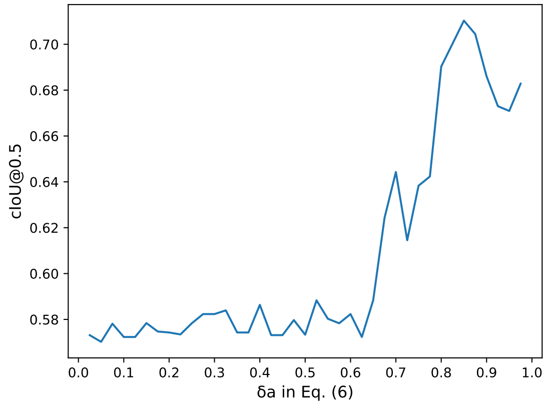


Fig. 11: **Success ratio under different thresholds δ_a.** We present the cIoU@0.5 scores of using different thresholds δ_a described in Eq. 6.

Acknowledgments

This work was supported in part by the Ministry of Science and Technology under grants 109-2221-E-009-113-MY3, 110-2628-E-A49-008, and 110-2634-F007-015. We are also grateful to the National Center for High-performance Computing for providing computational resources and facilities.

References

Afouras, T., Owens, A., Chung, J.S., Zisserman, A., 2020. Self-supervised learning of audio-visual objects from video, in: Proc. Euro. Conf. Comput. Vis., pp. 208–224.

Alayrac, J.B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., Zisserman, A., 2020. Self-supervised multimodal versatile networks, in: Proc. Neural Inf. Process. Syst., pp. 25–37.

Alwassel, H., Mahajan, D., Torresani, L., Ghanem, B., Tran, D., 2020. Self-supervised learning by cross-modal audio-video clustering, in: Proc. Neural Inf. Process. Syst., pp. 9758–9770.

Arandjelovic, R., Zisserman, A., 2017. Look, listen and learn, in: Proc. Int. Conf. Comput. Vis., pp. 609–617.

Arandjelović, R., Zisserman, A., 2018. Objects that sound, in: Proc. Euro. Conf. Comput. Vis., pp. 435–451.

Asano, Y.M., Patrick, M., Rupprecht, C., Vedaldi, A., 2020. Labelling unlabelled videos from scratch with multi-modal self-supervision, in: Proc. Neural Inf. Process. Syst., pp. 4660–4671.

Aytar, Y., Vondrick, C., Torralba, A., 2016. Soundnet: Learning sound representations from unlabeled video, in: Proc. Neural Inf. Process. Syst., pp. 892–900.

Chen, P., Zhang, Y., Tan, M., Xiao, H., Huang, D., Gan, C., 2020. Generating visually aligned sound from videos. IEEE Trans. Image Process. , 8292–8302.

Chung, J.S., Zisserman, A., 2018. Learning to lip read words by watching videos. Comput. Vis. Image Understanding , 76–85.

Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A., 2020. Music gesture for visual sound separation, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 10478–10487.

Gao, R., Feris, R., Grauman, K., 2018. Learning to separate object sounds by watching unlabeled video, in: Proc. Euro. Conf. Comput. Vis., pp. 35–53.

Gao, R., Grauman, K., 2019a. 2.5d-visual-sound, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 324–333.

Gao, R., Grauman, K., 2019b. Co-separating sounds of visual objects, in: Proc. Int. Conf. Comput. Vis., pp. 3879–3888.

Gao, R., Grauman, K., 2021. Visualvoice: Audio-visual speech separation with cross-modal consistency, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 15495–15505.

Gao, R., Oh, T.H., Grauman, K., Torresani, L., 2020. Listen to look: Action recognition by previewing audio, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 10457–10467.

Griffin, D., Lim, J., 1984. Signal estimation from modified short-time fourier transform. IEEE Trans. Acoust. , 236–243.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 770–778.

Hu, D., Nie, F., Li, X., 2019. Deep multimodal clustering for unsupervised audiovisual learning, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 9248–9257.

Hu, D., Qian, R., Jiang, M., Tan, X., Wen, S., Ding, E., Lin, W., Dou, D., 2020a. Discriminative sounding objects localization via self-supervised audiovisual matching, in: Proc. Neural Inf. Process. Syst., pp. 10077–10087.

Hu, D., Wang, Z., Xiong, H., Wang, D., Nie, F., Dou, D., 2020b. Curriculum audiovisual learning. arXiv preprint arXiv:2001.09414 .

Korbar, B., Tran, D., Torresani, L., 2018. Cooperative learning of audio and video models from self-supervised synchronization, in: Proc. Neural Inf. Process. Syst., pp. 7774–7785.

Lee, J.T., Jain, M., Park, H., Yun, S., 2021. Cross-attentional audio-visual fusion for weakly-supervised action localization, in: Proc. Int. Conf. Learn. Represent.

Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G.M., 2016. VideoLSTM convolves, attends and flows for action recognition. Comput. Vis. Image Understanding , 41–50.

Lin, Y.B., Li, Y.J., Wang, Y.C.F., 2019. Dual-modality seq2seq network for audio-visual event localization, in: Proc. Int. Conf. Acoustics, Speech, and Signal Process., pp. 2002–2006.

Lin, Y.B., Wang, Y.C.F., 2020. Audiovisual transformer with instance attention for audio-visual event localization, in: Proc. Asian Conf. Comput. Vis., pp. 274–290.

Lin, Y.B., Wang, Y.C.F., 2021. Exploiting audio-visual consistency with partial supervision for spatial audio generation, in: Proc. AAAI Conf. Artificial Intell., pp. 2056–2063.

Lu, Y.D., Lee, H.Y., Tseng, H.Y., Yang, M.H., 2019. Self-supervised audio spatialization with correspondence classifier, in: Proc. Int. Conf. Image Process., pp. 3347–3351.

Ma, S., Zeng, Z., McDuff, D., Song, Y., 2021. Active contrastive learning of audio-visual video representations, in: Proc. Int. Conf. Learn. Represent.

Mademlis, I., Tefas, A., Nikolaidis, N., Pitas, I., 2016. Multimodal stereoscopic movie summarization conforming to narrative characteristics. IEEE Trans. Image Process. , 5828–5840.

Min, X., Zhai, G., Zhou, J., Zhang, X.P., Yang, X., Guan, X., 2020. A multimodal saliency model for videos with high audio-visual correspondence. IEEE Trans. Image Process. , 3805–3819.

Morgado, P., Li, Y., Vasconcelos, N., 2020. Learning representations from audio-visual spatial alignment, in: Proc. Neural Inf. Process. Syst., pp. 4733–4744.

Morgado, P., Misra, I., Vasconcelos, N., 2021a. Robust audio-visual instance discrimination, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 12934–12945.

Morgado, P., Nvasconcelos, N., Langlois, T., Wang, O., 2018. Self-supervised generation of spatial audio for 360 video, in: Proc. Neural Inf. Process. Syst., pp. 360–370.

Morgado, P., Vasconcelos, N., Misra, I., 2021b. Audio-visual instance discrimination with cross-modal agreement, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 12475–12486.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

Owens, A., Efros, A.A., 2018. Audio-visual scene analysis with self-supervised multisensory features, in: Proc. Euro. Conf. Comput. Vis., pp. 631–648.

Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A., 2016. Ambient sound provides supervision for visual learning, in: Proc. Euro. Conf. Comput. Vis., pp. 801–816.

Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W., 2020. Multiple sound sources localization from coarse to fine, in: Proc. Euro. Conf. Comput. Vis., pp. 292–308.

Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S., 2018. Learning to localize sound source in visual scenes, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 4358–4366.

Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S., 2019. Learning to localize sound sources in visual scenes: Analysis and applications. IEEE Trans. Pattern Analysis and Machine Intelligence , 1605–1619.

Stafylakis, T., Khan, M.H., Tzimiropoulos, G., 2018. Pushing the boundaries of audiovisual word recognition using residual networks and lstms. Comput. Vis. Image Understanding , 22–32.

Tian, Y., Hu, D., Xu, C., 2021. Cyclic co-learning of sounding object visual grounding and sound separation, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 2745–2754.

Tian, Y., Li, D., Xu, C., 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing, in: Proc. Euro. Conf. Comput. Vis., pp. 436–454.

Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C., 2018. Audio-visual event localization in unconstrained videos, in: Proc. Euro. Conf. Comput. Vis., pp. 247–263.

Tzinis, E., Wisdom, S., Jansen, A., Hershey, S., Remez, T., Ellis, D., Hershey, J.R., 2021. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds, in: Proc. Int. Conf. Learn. Represent.

Wu, Y., Yang, Y., 2021. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 1326–1335.

Wu, Y., Zhu, L., Yan, Y., Yang, Y., 2019. Dual attention matching for audio-visual event localization, in: Proc. Int. Conf. Comput. Vis., pp. 6292–6300.

Xu, X., Dai, B., Lin, D., 2019. Recursive visual sound separation using minus-plus net, in: Proc. Int. Conf. Comput. Vis., pp. 882–891.

Xu, X., Zhou, H., Liu, Z., Dai, B., Wang, X., Lin, D., 2021. Visually informed binaural audio generation without binaural audios, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 15485–15494.

Xuan, H., Luo, L., Zhang, Z., Yang, J., Yan, Y., 2021. Discriminative cross-modality attention network for temporal inconsistent audio-visual event localization. IEEE Trans. Image Process. , 7878–7888.

Yang, K., Russell, B., Salamon, J., 2020. Telling left from right: Learning spatial correspondence of sight and sound, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 9932–9941.

Zhao, H., Gan, C., Ma, W.C., Torralba, A., 2019. The sound of motions, in: Proc. Int. Conf. Comput. Vis., pp. 1735–1744.

Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A., 2018. The sound of pixels, in: Proc. Euro. Conf. Comput. Vis., pp. 570–586.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 2921–2929.

Zhou, H., Xu, X., Lin, D., Wang, X., Liu, Z., 2020. Sep-stereo: Visually guided stereophonic audio generation by associating source separation, in: Proc. Euro. Conf. Comput. Vis., pp. 52–69.